TRANSPORTATION RESEARCH

RECORD

No. 1398

Highway Operations, Capacity, and Traffic Control

Traffic Flow and Highway Capacity

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1993

Transportation Research Record 1398
ISSN 0361-1981
ISBN 0-309-05467-2
Price: $28.00

Subscriber Category
IV highway operations, capacity, and traffic control

TRB Publications Staff
Director of Reports and Editorial Services: Nancy A. Ackerman
Senior Editor: Naomi C. Kassabian
Associate Editor: Alison G. Tobias
Assistant Editors: Luanne Crayton, Norman Solomon, Susan E. G. Brown
Publications Coordinator: Sharada Gilkey
Graphics Specialist: Terri Wayne
Office Manager: Phyllis D. Barber
Senior Production Assistant: Betty L. Hawkins

Printed in the United States of America

Transportation Research Record 1398

Contents

Foreword

The 18 papers in this volume are related by their focus on highway capacity, quality of service, traffic flow measurement, and traffic flow theory. The papers cover a wide range of problems reflecting the concerns of both practitioners and theoreticians.

Highway capacity is receiving considerable attention as a result of the research effort toward producing the next edition of the *Highway Capacity Manual*, to be published about 2000. The initial papers in this Record examine the issue of capacity as it relates to two-lane highways, freeway bottlenecks, freeway weaving sections, stop-controlled intersections, traffic circles, left turns, and delay at signalized intersections.

Traffic flow theory, modeling, and control applications are also examined in papers on oversaturated conditions and traffic control, speed estimation and catastrophe theory, headway probability distributions, and simulation modeling of school congestion.

Whether the reader is a traffic engineer trying to determine the capacity and level of service of two-lane highways or a traffic flow theoretician pondering the vagaries of traffic flow equations, the papers in this Record should be both interesting and informative.

Methodology To Assess Level of Service on US-1 in the Florida Keys

RAFAEL E. DE ARAZOZA AND DOUGLAS S. MCLEOD

The methodology developed to assess level of service (LOS) on US-1 in the Florida Keys is presented. Although US-1 is predominantly an uninterrupted-flow two-lane roadway in the Keys, its uniqueness warrants an LOS evaluation different than that found in the 1985 *Highway Capacity Manual* (HCM). US-1 extends from Key West to the Florida mainland with no major roads intersecting it. Furthermore, no other principal arterial serves the Keys or the Keys' resident and tourist population of well over 100,000. Its unique geography, land use patterns, and trip-making characteristics presented a challenge in developing and applying a reasonable and acceptable method of assessing its LOS. A uniform method was developed to assess LOS on US-1 to cover both its overall arterial length from Key West to the Florida mainland and 24 delineated roadway segments. The methodology, which employs average travel speed as the main measure of effectiveness, was developed from basic principles, criteria, and speed relationships contained in Chapters 7 (Rural Multilane Highways), 8 (Rural Two-Lane Highways), and 11 (Urban and Suburban Arterials) of the 1985 HCM. The results of the study correlate well with perceived operating conditions on US-1, and over a 2-year period the methodology appears to have a good level of reliability. The authors recommend that for uninterrupted flow conditions in developed areas (e.g., in communities and along beaches), Chapter 8 of the 1985 HCM incorporate average travel speed as the main measure of effectiveness to determine LOS.

The purpose of this paper is to present the methodology developed by the Monroe County US-1 level of service (LOS) Task Force to assess LOS on US-1 (the Overseas Highway) in the Florida Keys (*1*). The authors are members of the task force.

US-1, which is mostly a two-lane highway, has unique geographic and trip characteristics. It extends through the Florida Keys, covering approximately 180 km (112 mi) from the city of Key West to the Florida mainland. There are 48 bridges that cross water for a total length of 35 km (22 mi), and the longest bridge is approximately 11 km (7 mi). No other road provides vehicular access to the Florida Keys from the rest of Florida or anywhere else. Few local roads are 5 km (3 mi) in length. Consequently, US-1 is not only a regional principal arterial serving intra- as well as interstate travel, but also the local road for most of the trips within the Keys. US-1 annual average daily traffic (AADT) volumes range from 4,700 to 37,200. The road serves a large tourist demand and is one of the most scenic in the United States. The linear geography with the narrow land width of most of the Florida Keys are further characteristics.

Most of the surrounding land use is rural developed and suburban in nature; however, some areas are totally rural and others are urban, such as Key West and its suburbs. With the exception of the few completely rural segments and the bridges, strip commercial stores, motels, and restaurants are common throughout the Keys along US-1. Many driveways and intersecting local roads provide access to the surrounding residential areas.

Part of the growth management process in Florida is to assess roadway LOS to determine if roadway facilities meet standards established by state regulations. From a state transportation perspective, the overall operating condition of US-1 is important, not the condition of any smaller segment. With Key West a major tourist destination at the southern end of the Keys and no alternative routes to it, the logical analysis section of highway extends from Key West to the mainland. From perspectives of local transportation and development approval, shorter segments for analysis are desirable.

For planning purposes the Florida Department of Transportation (FDOT) has adopted LOS standards for all state roads. The applicable peak hour of analysis is the 100th highest hour of the year, representative of a typical peak hour during a 3-month peak season. FDOT's LOS standards vary by road and area type, with LOS C being the applicable standard in the Keys. TRB's *Special Report 209: Highway Capacity Manual* (HCM) (*2*), FDOT's *Level of Service Manual* (*3*), based on the HCM, and accompanying software are used extensively throughout Florida to determine highway capacities, LOS, and compatibility with LOS standards.

The US-1 LOS study encompassed approximately 174 km (108 mi) of US-1 from Key West/Stock Island to the Monroe-Dade county line, broken down as follows:

- 129 km (80 mi), 74 percent two-lane uninterrupted flow;
- 32 km (20 mi), 19 percent four-lane uninterrupted flow; and
- 13 km (8 mi), 7 percent four-lane urban/suburban interrupted flow.

HCM Chapters 7 (Rural Multilane Highways), 8 (Rural Two-Lane Highways), and 11 (Urban and Suburban Arterials) were consulted to determine applicability to the unique conditions and vehicular traffic operations and characteristics of the Florida Keys. Only the 13 km (8 mi) of urban/suburban interrupted flow and the small percentage of two-lane truly rural portions are believed to correlate directly to HCM Chap-

R. E. De Arazoza, Florida Department of Transportation, District 6, 602 South Miami Avenue, Miami, Fla. 33130. D. S. McLeod, Florida Department of Transportation, Mail Station 19, 605 Suwannee Street, Tallahassee, Fla. 32399-0450.

ters 11 and 8. Thus, the challenge was to develop a methodology to assess arterial LOS along US-1 without deviating from the principles of the HCM. Toward that end, a task force was created consisting of representatives from state and local agencies and an engineering consulting firm. An interim methodology was developed during the latter part of 1990, with the final methodology completed in June 1991. The final methodology was applied to the 1992 study data.

NEED FOR SPEED-BASED METHODOLOGY

Chapter 8 of the HCM presents a methodology that applies primarily to the typical rural undeveloped situation. Essentially, these two-lane facilities are long stretches of roads with few side intersecting streets and driveways connecting directly to the roads. Chapter 8 methodology relies mainly on percentage time delay to assess LOS.

Throughout the United States many two-lane uninterrupted-flow highways pass through developed areas such as small communities and pass along beaches. However, the HCM does not directly address ways of handling these two-lane uninterrupted-flow highways. Frequently in these areas, posted speeds are also lower than they are on open highways, and it is believed that motorists expect to be traveling at somewhat lower speeds under these conditions. After much discussion, Florida's LOS Measurement Task Team (as well as the project task team) took the position that most motorists are more concerned about maintaining a decent travel speed under these uninterrupted-flow conditions in developed areas than trying to pass. Similarly, it is believed that the average motorist in the Florida Keys is concerned mostly with operating at an acceptable average travel speed, not with the ability to pass. This assumption is supported by the physical and traffic characteristics of the Keys (e.g., adjacent land development, sightseeing tourists), local knowledge, and discussions with motorists. Furthermore, average speeds comparable to the ability to pass appearing in Table 8-1 of the HCM are appreciably higher than the typical operating speeds of US-1 in the Florida Keys.

With regard to the four-lane uninterrupted-flow portions of US-1, a similar dilemma occurred. HCM Chapter 7 methodology applies to multilane highways with operating characteristics generally unlike those of US-1 through the Florida Keys. For instance, average travel speeds depicted by Table 7-1 of the HCM are also higher than those encountered in the Keys. Furthermore, the methodology inherent in Equations 7-1, 7-2, and 7-3 is closely related to those of freeways with their higher service flow rates, which again neither simulate nor resemble those of US-1 in the Keys. The four-lane portion is found mostly in Key Largo (the northeastern end of the Keys), which has a weighted posted speed limit of 72.5 km/hr (45 mph). Key Largo is developed with strip commercial and residential development. It has many driveway connections and side streets directly accessing US-1.

The remaining 7 percent of the total US-1 mileage is four-lane interrupted flow. These are the portions encompassing Marathon (in the middle of the Keys) and Stock Island (near Key West). The operating characteristics here are truly urban/suburban and interrupted flow in nature, resembling those of HCM Chapter 11. Thus, the methodology of HCM Chapter 11 was used in assessing LOS on these segments.

From the preceding discussion, it was evident that for most of US-1 in the Keys the HCM was not directly applicable and a distinct method to assess LOS on US-1 should be developed. The task team concentrated on keeping consistency with the basic philosophy of the HCM and yet being sensitive to the Keys' uniqueness. Thus, the proposed methodology correlates measured travel speeds along US-1 with LOS speed thresholds developed as part of this study. This is in line with the concept behind the HCM of average travel speed being the main parameter for measuring arterial LOS.

DEVELOPMENT OF LOS MEASURES OF EFFECTIVENESS

Assuming that average travel speed is the most appropriate measure of effectiveness for LOS on uninterrupted-flow facilities in developed areas, the challenge became that of developing appropriate criteria while still generally conforming to the HCM. Or, in other words, can reasonable speed-based criteria be inferred from Table 8-1, and to a lesser extent Table 7-1, of the HCM?

In the United States, posted speed limits for two-lane and multilane highways are no higher than 55 mph even though design or free-flow speeds are generally higher. Although it is accepted practice to base speed limits on the 85th-percentile speed, it is widely recognized that posted speed limits are set below that criterion. As in other parts of the country, posted speed limits in the Keys appear to be influenced more by the number of access points and level of residential and commercial development than by prevailing speeds. From Tignor's research (*4*) it can be inferred that average free-flow speeds in urban areas generally range from 2 to 8 mph—or approximately 5 mph—higher than posted speed limits. Interestingly, the level of service A speed criterion for two-lane (and multilane) highways is approximately the average of typical 55-mph posted speed limits and more typical 60-mph free-flow speeds. Assuming that this relationship is appropriate, analysts may reasonably make LOS A speed criteria for uninterrupted highways on the basis of posted speed limits.

Toward this end, the speed ratios between LOS thresholds from Tables 7-1, 8-1, and 11-1 of the HCM were used in the analysis. These ratios were weighted against actual mileage of US-1 in the Florida Keys to represent the prevailing type of flow: two-lane uninterrupted flow, four-lane uninterrupted flow, and four-lane interrupted flow. For example, from the level terrain portion of HCM Table 8-1, the ratio of LOS B speed to LOS A speed is 55/58, or 0.948. The ratio LOS C/LOS A is 52/58, or 0.897; the ratio LOS D/LOS A is 50/58, or 0.862, and so on. The same process was applied to Tables 7-1 (96.6 km/hr, or 60 mph) and 11-1. Then each ratio was weighted to account for the length of the section of US-1 to which that type of traffic flow applied. Once all the ratios were developed, the weight criteria were applied as in the following example:

Type of Flow	*LOS C/LOS A Ratio*	*Weight*
Two-lane uninterrupted	52/58 = 0.897	74
Four-lane uninterrupted	44/50 = 0.880	19
Four-lane interrupted	22/35 = 0.629	7

Therefore, the overall speed ratio between LOS C and LOS A is

$$\frac{74(0.897) + 19(0.880) + 7(0.629)}{100} = 0.875$$

This process was applied to develop all the required ratios.

Further observations with reference to Tables 8-1, 7-1, and 11-1 yielded the following. From Table 8-1 the difference between LOS A and LOS B speeds is 4.8 km/hr (3 mph), or 4.8 km/hr (3 mph) above an assumed posted speed limit of 88 km/hr (55 mph). From Tables 7-1 and 11-1 the differences are 3.2 and 11.3 km/hr (2 and 7 mph), respectively, with LOS A lower than assumed speed limits. Therefore, from these observations, previous discussion in this paper, and local knowledge, it was determined that the overall US-1 posted speed limit of 79.6 km/hr (49.5 mph) fell reasonably between the LOS A and B thresholds. This assumption is not far from the premise that if a vehicle is able to sustain a travel speed equal to the posted speed limit, then it will correspond typically with the upper ranges of LOS (i.e., LOS A or LOS B).

With these speed differentials and the LOS range premise in mind, the US-1 overall speed thresholds for LOS A and B became 82.1 km/hr (51 mph) [2.4 km/hr (1.5 mph) above 79.6 km/hr (49.5 mph)] and 77.3 km/hr (48 mph), respectively. Applying the ratio developed LOS C/LOS A to the LOS A speed resulted in 72.0 km/hr (45 mph), rounded off [i.e., 0.875 × 82.1 km/hr (51 mph) = 71.8 km/hr (44.6 mph)], which then became the threshold for LOS C. After applying all the ratios, the overall LOS criteria for US-1 were developed, as given in the following:

LOS	*Speed [km/hr (mph)]*
A	≥82 (51)
B	≥77 (48)
C	≥72 (45)
D	≥68 (42)
E	≥58 (36)
F	<58 (36)

Thus, in essence, the state minimum operating speed standard (LOS C) for US-1 became 72 km/hr (45 mph). Or, in other words, drivers can reasonably expect to average at least 72 km/hr (45 mph) at any time of the year from the Monroe-Dade county line to Key West.

The next step was to develop LOS–speed threshold values for the individual segments of US-1. Twenty-four segments were selected, as presented in Table 1. Each segment is fairly homogeneous having a uniform roadway cross section and traffic flow. No further work was needed to cover the 7 percent mileage of the interrupted portions of US-1 found on Marathon and Stock Island, adjacent to Key West. As was discussed earlier, these segments correlate with Chapter 11 of the HCM. Therefore, direct application of Table 11-1 LOS–speed criteria for a Class I arterial was made.

The remaining segments fell within the two-lane and four-lane uninterrupted-flow criteria. It was decided to make the LOS A speed criterion 2.4 km/hr (1.5 mph) higher than the weighted posted speed limit to keep consistency with the overall criteria. LOS C speed was set at 9.7 km/hr (6 mph) below

TABLE 1 Segment Description

Segment No.	Mile Markers	Key(s)
1	4 – 5	Stock Island, Key Haven
2	5 – 9	Boca Chica, Rockland
3	9 – 10.5	Big Coppitt
4	10.5 – 16.5	Shark, Saddlebunch
5	16.5 – 20.5	Lower Sugarloaf, Upper Sugarloaf
6	20.5 – 23	Cudjoe
7	23 – 25	Summerland
8	25 – 27.5	Ramrod
9	27.5 – 29.5	Torch
10	29.5 – 33	Big Pine
11	33 – 40	W. Summerland, Bahia Honda, Ohio
12	40 – 47	7-mile bridge
13	47 – 54	Marathon, Key Colony Beach
14	54 – 60.5	Fat Deer, Crawl, Grassy
15	60.5 – 63	Duck, Conch
16	63 – 73	Long, Fiesta, Craig
17	73 – 77.5	Lower Matecumbe
18	77.5 – 79.5	Fill
19	79.5 – 84	Upper Matecumbe
20	84 – 86	Windley
21	86 – 91.5	Plantation
22	91.5 – 99.5	Tavernier
23	99.5 – 106	Key Largo
24	106 – 112.5	Key Largo, Cross Key

the LOS A speed, consistent with Tables 7-1 and 8-1 of the HCM. LOS B and D speed criteria were set to provide equal increments between LOS A and LOS D [i.e., LOS B 4.8 km/hr (3 mph) below LOS A speed and LOS D 4.8 km/hr (3 mph) below LOS C speed]. LOS E was set 9.7 km/hr (6 mph) below the LOS D speed. This makes the segmental speed differential between LOS thresholds consistent with the differentials in the overall criteria, except for one consideration. On any segment, intersection delay would be deducted from the segment's travel time to account for the influence of that signal on the segment (i.e., signal delay = 1.0 × 15 sec average stopped delay). This corresponds to an LOS C delay due to isolated signals. LOS C delay was chosen because LOS C is the state LOS standard for US-1 in the Florida Keys.

The rationale behind deducting signal delay from the segment analysis was to recognize for the impact of signals in reducing travel time. This provides the required sensitivity in the segment that is to assess the impact not only of regional vehicular trips but also of those that are local in nature. The following tables illustrate the concept and give an example for the US-1 segmental LOS-speed relationship.

The uninterrupted-flow segment criteria are as follows:

LOS	*Speed [km/hr (mph)]*
A	≥2.4 (1.5) above the posted speed limit
B	≥4.8 (3.0) below LOS A
C	≥9.7 (6.0) below LOS A
D	≥14.5 (9.0) below LOS A
E	≥24 (15.0) below LOS A
F	<24 (15.0) below LOS A

A segment having a weighted posted speed limit of 72 km/hr (45 mph) has the following criteria:

LOS	*Speed [km/hr (mph)]*
A	≥74.9 (46.5)
B	≥70.0 (43.5)
C	≥65.2 (40.5)
D	≥60.4 (37.5)
E	≥50.7 (31.5)
F	<50.7 (31.5)

The LOS-speed criteria for interrupted-flow segments (Marathon and Stock Island) are based directly on a Class 1 arterial from Table 11-1 of the HCM.

Speed data from both the overall length of US-1 and the individual segments were compared with the applicable LOS–speed thresholds. This arrangement provided for an assessment of the facility LOS plus an indication of reserve speed, if any.

Under the growth management processes of Florida and Monroe County, if the overall LOS for US-1 fell below the LOS C standard then no additional land development would be allowed to proceed in the Florida Keys, unless the proposed new development traffic impact were mitigated. If the overall LOS for US-1 was C or better, additional development could take place in those segments in which reserve speed was available (i.e., the segment's speed was higher than the standard threshold).

In addition to meeting highway LOS standards there are many other considerations in Florida's growth management process pertaining to the Florida Keys. These are beyond the scope of this paper.

SPEED STUDIES

Considering the types of trips served by US-1, it was decided to conduct travel time and delay runs to cover the entire length of US-1 from Key West to the Monroe-Dade county line (mainland) and each segment of the highway along the way. Travel speeds for the overall length (from Key West to the mainland) provide an indication of the LOS for the regional trips. Travel speeds for each segment also provide an opportunity to assess the impact of local trips.

The floating-car technique, as defined by the *Manual of Traffic Engineering Studies* (*5*) published by ITE, was used. In this technique the test vehicle "floats" with the traffic by passing as many vehicles as pass the test vehicle. A safe operation was maintained and applied in passing maneuvers (where permitted), following other vehicles and changing speeds. The equipment used was an electronic distance measuring instrument with both distance and time features and ability to download data into a personal computer.

The next step in the process was to determine the number of travel time runs and how, when, to where, and from where. Runs were started at both ends of US-1. For example, one run started on Stock Island (Key West city limits) and proceeded to the mainland (Dade County). After reaching this point, the vehicle turned back and proceeded to end the run where it started, on Stock Island. On another day the reverse was true (i.e., the run started in Dade County instead of Stock Island). It was decided to perform 14 two-way runs, or 28 in each direction, covering the 174 km (108 mi) study portion of US-1. Twenty-eight runs provide enough data for statistical significance. Control points were established at each of the 24 segments to record travel time and speed data specific to each one of those segments. Seven runs were started at Stock Island and seven in Dade County covering the hours between 9 a.m. and 7 p.m. Each began at staggered hours to cover the varied trip purposes and time frames within the Keys with the intent to cover peak travel periods of all the segments.

For each run the process provided data, such as running speed and travel speed, in each direction of US-1. Vehicular traffic counts were also collected at three locations covering 7 days.

As stated before, Florida's LOS standards are based on the concept of the typical peak hour during the peak season, which is approximately the 100th highest volume hour of the year in developed areas. Whereas it is normally reasonable to conduct an analysis based on this peak hour for roadways, it is impractical to evaluate a 174-km (108-mi) roadway for a peak hour; full-length trips are approximately 2½ hr in duration. To meet the intent of Florida's LOS standards, it was agreed to conduct the travel time studies during the peak travel hours of the peak month (March) and to use the median speed of the travel time runs. Using this approach, compatibility with LOS standards is developed on the basis of a typical drive during the peak month. Compared with other roads in Florida this approach is somewhat tougher because it uses the highest-volume month instead of the peak season, but it is somewhat more lenient because it uses a typical driving hour instead of one of the highest-volume hours. The median value was also selected, instead of the average, to avoid the influence of extremely high or low speed values at either end of the survey population.

STUDY RESULTS

The 1991 travel time field runs were conducted between March 1 and March 21. Fourteen runs were made to cover both directions along US-1 (total of 28 one-way runs), at the rate of one run per day in each direction. The 1992 field runs were conducted between February 29 and March 20 in the same manner as in 1991. Seven-day traffic counts were made, at three locations, during 1991 and 1992 and converted to AADT. These are given in Table 2.

Tables 3 and 4 present a summary of the results of the field studies for the years 1991 and 1992 (*6,7*). The tables cover both the overall length of US-1 and the individual segments plus the applicable LOS.

The results were not surprising. Overall, the US-1 median travel speeds of 76.7 km/hr (47.7 mph) for 1991 and 75.5 km/hr (46.9 mph) for 1992 reflected an acceptable LOS C. This overall rating is supported by the data collected plus local knowledge and the authors' experience in the area. It is interesting to see that for the overall speed data, the mean (average) speed is either identical or very close to the median speed for both years, suggesting fairly good individual data items with few extremes. Worth noting is the fact that for both years, the survey vehicle was able to maintain an overall median speed relatively close to the weighted posted speed limit [i.e., 2.9 km/hr (1.8 mph) below in 1991 and 4.1 km/hr (2.6 mph) below in 1992]. This fact supports the LOS C assessment of US-1 and the discussions presented earlier in this paper.

The results from the segmental analyses were not surprising either. The resulting LOS accurately reflect traffic operations and perceived levels of congestion. From the 1991 data, only one segment failed to meet the LOS C standard. This is Segment 19 in the Upper Matecumbe area. The two segments (17 and 18) south of it, covering the Lower Matecumbe and Tea Table areas, showed speeds at the lower ends of LOS C with little reserve available. These assessments corresponded well with the authors' knowledge of traffic operations and development conditions of the area. In 1992 these three segments failed the LOS C standard. All three of them were assessed at LOS D.

The authors were pleased with the results of the study. The 1991 and 1992 LOS assessments for US-1 accurately reflect local knowledge, field experience, and perceived conditions in the Florida Keys. Inspection of the survey data also supports this claim. For instance, looking at the 1992 speeds for the segments, with the exception of 3 segments of 24, the difference between the median and mean speeds was less than 1.6 km/hr (1 mph). For the 2 segments the difference between the median and mean was less than 3.2 km/hr (2 mph). In keeping with the 1992 data, 13 of 24 segments (54 percent) have standard deviations less than 4.8 km/hr (3 mph). Of the other 11 segments with standard deviations greater than 4.8 km/hr (3 mph), only 1 had a standard deviation greater than 9.7 km/hr (6 mph). The standard deviations for the segments came out higher for the 1991 data. However, 18 of the 24 had

TABLE 2 Average Annual Daily Traffic

LOCATION	1991	1992	%CHANGE
Big Pine	18,199	17,529	−3.7
Marathon	24,043	25,933	7.9
Upper Matecumbe	17,357	17,564	1.2

TABLE 3 Summary of 1991 Travel Speed Data

Segment No.	Mean (kmh)	Median (kmh)	Std. Dev. (kmh)	Standard (kmh)	Reserve (kmh)	LOS
1	57.7	57.5	8.6	35.4	22.1	A
2	91.1	91.0	2.0	75.3	15.7	A
3	78.5	78.4	6.7	68.4	10.0	B
4	84.0	84.0	5.5	75.3	8.7	C
5	82.4	82.3	5.5	72.6	9.7	C
6	69.8	69.6	5.8	61.6	8.0	C
7	74.5	74.9	5.4	61.6	13.3	A
8	77.8	77.5	5.0	61.6	15.9	A
9	75.7	75.0	5.3	61.6	13.4	A
10	61.9	62.1	6.1	61.6	0.5	C
11	87.0	86.3	2.8	73.4	12.9	B
12	85.1	85.2	5.9	68.4	16.8	A
13	65.4	65.2	3.2	35.4	29.8	A
14	86.2	86.1	4.2	73.9	12.2	B
15	87.4	87.9	4.7	75.3	12.6	B
16	81.2	83.5	15.8	73.9	9.6	C
17	81.6	82.5	6.3	75.3	7.2	C
18	80.0	81.5	6.5	75.3	6.2	C
19	63.0	63.2	5.1	61.6	1.6	D
20	68.7	69.1	7.9	61.6	7.5	C
21	64.2	65.5	8.0	61.6	3.9	C
22	79.6	80.1	3.2	61.6	18.5	A
23	76.2	77.4	4.5	61.6	15.8	A
24	81.2	82.9	6.0	69.7	13.2	B
OVERALL	76.3	76.7	2.5	72.4	4.3	C

TABLE 4 Summary of 1992 Travel Speed Data

Segment No.	Mean (kmh)	Median (kmh)	Std. Dev. (kmh)	Standard (kmh)	Reserve (kmh)	LOS
1	57.3	58.2	7.3	35.4	22.8	A
2	91.3	90.3	3.4	81.3	9.0	B
3	72.9	75.1	11.0	73.2	1.9	C
4	82.9	83.4	4.4	81.3	2.1	C
5	80.2	80.9	4.3	78.0	2.9	C
6	69.5	69.2	6.3	65.2	4.0	C
7	72.0	72.2	4.6	65.2	7.0	B
8	76.9	77.7	4.6	65.2	12.5	A
9	74.6	74.9	4.1	65.2	9.7	A
10	61.3	62.6	6.7	62.1	0.5	C
11	85.8	85.7	3.3	79.2	6.5	B
12	85.2	84.3	5.4	73.2	11.1	A
13	63.9	63.8	2.9	35.4	28.4	A
14	82.2	82.2	6.1	79.6	2.6	C
15	83.9	84.6	5.3	81.3	3.3	C
16	81.8	81.7	4.3	79.6	2.1	C
17	79.9	81.2	5.7	81.3	−0.1	D
18	79.5	79.9	5.9	81.3	−1.4	D
19	64.8	64.6	4.7	65.2	−0.6	D
20	67.1	69.8	7.1	65.2	4.6	C
21	62.0	62.2	4.5	63.4	−1.2	D
22	78.6	77.5	4.5	67.1	10.4	A
23	75.8	75.4	4.3	64.0	11.4	A
24	80.0	82.4	8.3	74.7	7.7	B
OVERALL	75.4	75.5	1.9	72.4	3.1	C

deviations less than 6.4 km/hr (4.0 mph). The results imply that for the most part the assessed LOS will not go beyond one letter change when the standard deviation is added or subtracted to the mean. Finally, the standard deviations for the 1991 and 1992 overall US-1 speeds were 2.5 and 1.9 km/hr (1.6 and 1.2 mph), respectively.

CONCLUDING REMARKS

From the 1992 highway LOS determinations and adopted LOS standards, it was concluded that additional land development may occur in the Florida Keys. However, based on the adopted LOS criteria, unless the traffic impact is adequately mitigated, development should not be approved on the four segments that failed to meet the LOS C standard. Although one other segment, Big Pine, was at LOS C, it had a very low reserve speed. This means that any proposed land development should be closely monitored.

The methodology developed to assess LOS on US-1 in the Florida Keys followed the basic concept of arterial analyses in the HCM with average travel speed as the main measure of effectiveness. The results accurately reflect perceived levels of congestions and local knowledge. For example, motorists can compare their travel speed with the posted speed limit. The methodology has been formally approved by both state agencies principally involved in transportation aspects of Florida's growth management process, the Department of Community Affairs and the Department of Transportation. County commissioners have also incorporated formally the methodology into Monroe County's land use regulations.

As the result of this study, it is recommended that for uninterrupted-flow conditions in developed areas Chapter 8 of the HCM consider average travel speed as the main parameter for determining LOS. Because the results of the study reflect local conditions in the Florida Keys, additional data collection and studies are recommended for nationwide applications.

ACKNOWLEDGMENTS

Thanks and appreciation to the Monroe County US-1 LOS Task Force, which in addition to the authors comprised the following individuals: Rene de Huelbes, Kenneth Metcalf, David Koppel, Mark Rosch, and Jack Schnettler. Gratitude is also expressed to the Florida Department of Transportation for its support.

REFERENCES

1. *Summary of Final Recommendations*. US-1 Level of Service Task Force, Key West, Fla., 1991.
2. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
3. *Level of Service Manual*. Florida Department of Transportation, Tallahassee, 1992.
4. S. C. Tignor. Driver Speed Behavior on U.S. Streets and Highways. *Compendium of Technical Papers*, ITE, Aug. 1990.
5. *Manual of Traffic Engineering Studies*, 4th ed. ITE, Washington, D.C.
6. *1991 Travel Time and Delay Study of US-1 in Monroe County, Key West, FL*. Monroe County Planning Department, Fla., 1991.
7. *1992 Travel Time and Delay Study of US-1 in Monroe County, Key West, FL*. Monroe County Planning Department, Fla., 1992.

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

Comparison of Performance of TWOPAS and TRARR Models When Simulating Traffic on Two-Lane Highways with Low Design Speeds

JAN L. BOTHA, XIAOHONG ZENG, AND EDWARD C. SULLIVAN

A comparison of the TWOPAS and TRARR models when used to simulate traffic operation on two-way rural highways with low design speed and two lanes is reported on in this paper. It was found that the TWOPAS and TRARR models are generally comparable in their ability to simulate traffic operations on two-lane, two-way highways. However, the TWOPAS simulation results compared better with field data for roads with an 80-km/hr (50-mph) design speed. This was the case for both a level terrain site and a rolling terrain site. The comparison was made in travel time and percentage time delay. Both models require further work before they can be applied without reservation to the many types of situations that might arise on two-lane roads. It is recommended, however, that TWOPAS be adopted for analysis related to capacity and level of service.

Computer simulation models are becoming increasingly important in the analysis of traffic on highways and streets. This is particularly true for two-lane highways where frequent changes in alignment and lack of passing opportunities create complex and frequently changing traffic conditions. These conditions are difficult and expensive to analyze with empirical methods because of the large amount of data required. Simulation models are often used to analyze such situations.

The performance of the TWOPAS and TRARR simulation models is compared. The models are used to simulate traffic operation on two-lane, two-way rural highways with low design speeds. TWOPAS is a microscopic stochastic computer model originally developed and documented at the Midwest Research Institute (*1,2*). It was used to generate the basic data presented in the 1985 *Highway Capacity Manual* for the analysis of the capacity and level of service for rural two-lane highways (*3–5*). The TRARR model was developed by the Australian Road Research Board (*6*). TRARR has become noteworthy not only because of its use in Australia but also because it has been applied for important research in Canada and the United States (*7,8*). Both TWOPAS and TRARR were developed specifically for simulation of traffic operations on two-lane rural highways.

The model comparison was carried out as part of a project undertaken for the California Department of Transportation (Caltrans). The aim of the project was to develop a methodology for the analysis of the capacity and level of service for two-lane highways with design speeds of less than 96 km/hr (60 mph).

The comparison is presented in terms of the features of the models as well as the conformance of the model predictions with field data. In addition, noteworthy experience with the models and information on model modifications made during the course of the project are presented. The comparison is discussed, and a summary of major conclusions and recommendations is presented.

J. L. Botha and X. Zeng, Department of Civil Engineering and Applied Mechanics, One Washington Square, San Jose State University, San Jose, Calif. 95192-0083. E. C. Sullivan, Civil and Environmental Engineering Department, California Polytechnic State University, San Luis Obispo, Calif. 93407.

COMPARISON OF MODEL FEATURES

A major difference between the models, before the modifications made during this project, was that TRARR was operational on an IBM-compatible personal computer (PC) whereas TWOPAS was operational only on a mainframe computer. During the course of the project, TWOPAS was modified to operate on an IBM-compatible PC. Recently it was discovered that FHWA also had modified the model to run on a PC.

TRARR used metric units for both input and output. The field data used as input were, however, in imperial units and the output was also required in imperial units. Because a large number of computer runs were made during the course of the project, the input and output were converted to imperial units.

The remaining major model features considered in the comparison are

1. Basic methodology;
2. Model input;
3. Model output;
4. Documentation, support, and computer requirements; and
5. Ease of use.

Basic Methodology

Both TRARR and TWOPAS are microscopic, stochastic models that can simulate uninterrupted traffic on two-lane, two-way highways with or without auxiliary lanes. Both models

operate on a time-scanning basis for updating vehicle movements.

The models are similar in that individual driver behavior and vehicle performance are modeled in detail. Differences in driver behavior with respect to desired speeds are accounted for. Several vehicle types are modeled. Both driver behavior and vehicle performance are restricted by horizontal and vertical alignment as well as other geometric features, such as passing and no-passing zones.

The logic of the models consists essentially of initially placing vehicles on the road and allowing a warm-up time during which the traffic settles into a pattern representative of the prevailing roadway and traffic conditions. Different driver and vehicle types, proportional to the specified flow rates, are generated at each end of the road. A warm-up section precedes the actual test section. The warm-up section allows the traffic to settle into a pattern representative of the field conditions at the entrance to the test section.

Model Input

The models are very similar as far as input is concerned. Both models use data on road geometry, traffic control (passing and no-passing zones), vehicle characteristics (by direction and vehicle type), driver behavior, and entering traffic (by direction and vehicle type). In most cases the data on road geometry, traffic control, and entering traffic are required and default values are provided for the remainder. More details on the possible input items are shown in Figure 1.

Although the input is in principle similar for the two models, there are some notable differences. Vertical and horizontal curves can be directly input into TWOPAS, whereas the vertical alignment has to be directly input into TRARR but the horizontal alignment is input through a road speed index. The road speed index is a function of the curve radius and the 85th percentile of the desired speed. A table of road speed indexes is provided in the manual.

During the course of the project, TWOPAS was modified to increase the number of horizontal curves that it could accommodate from 9 to 50 and the number of grades from 30 to 100 to accommodate the alignment encountered on roads with low design speeds. The alignment specification limitations for TRARR are different. The road is divided into units of which the lengths have to be specified. The limitation on the number of units is 165, but the model was only successfully executed with 150 units. The grade, road speed index, sight distance, and presence of a barrier line as well as an auxiliary lane have to be specified. It is very awkward to make changes to these data because of the constant length of the road unit.

TWOPAS can accommodate 13 vehicle types: 5 types of trucks and buses, 4 types of recreational vehicles (RVs), and 5 types of automobiles.

Characteristics such as acceleration and speed capabilities can be specified or default values used.

Eighteen vehicle types can be specified for TRARR. Existing files containing default values for vehicle characteristics can be used or changed to reflect local conditions. Default values are specified for nine types of large or heavy trucks, one small truck, one car and caravan, and seven types of cars. Some of the vehicles described appear to be vehicles that are found in Australia but not in the United States. No large RVs of the types found in the United States are described by the default values. However, some of the existing vehicle types could be converted to RVs and designated as such in the output.

Model Output

The standard TWOPAS output is much more extensive than the standard TRARR output. The more extensive output of

```
Geometrics
          °Grades
          °Horizontal curves
          °Passing sight distance
          °Passing and climbing lanes
          °Immediate upstream alignment
Traffic Control
          °Passing and no-passing zones
Vehicle Characteristics
          °Vehicle acceleration and speed capabilities
          °Vehicle lengths
Driver Characteristics and Preferences
          °Desired speeds
          °Preferred acceleration levels
          °Limitations on use of vehicle power
          °Passing decisions
          °Behavior in passing and climbing lanes
Entering Traffic
          °Flow rates
          °Vehicle mix
          °Platooning of entering traffic
Simulation Parameters
          °Warm-up time
          °Simulation time
```

FIGURE 1 TWOPAS and TRARR input features.

TWOPAS is an advantage when detailed output is desired, but it is time-consuming and cumbersome when only a few output items are desired, because the detailed output cannot be suppressed. Additional output can be obtained from both models by printing program files.

A summary of the basic output for TWOPAS is given in Figure 2. Additional final output at intermediate times as well as data that can be used to calculate fuel consumption are available on program files. The latter feature was removed from the existing PC version.

A summary of the standard output for TRARR is also presented in Figure 2. A display of vehicles moving along the road, passing, and merging can be generated on the computer screen. TRARR does not automatically provide output for vehicle characteristics, road characteristics, and desired speed distributions (as TWOPAS does), but these data can be printed from input files.

It should be noted that the term "overtaking," as used in Figure 2, has different meanings for TWOPAS and TRARR. In TWOPAS it means catching up, whereas in TRARR it means passing.

The output data for both programs generally are classified by direction and by vehicle type and summarized by subgroup, that is, trucks, RVs, and automobiles.

Documentation, Support, and Computer Requirements

Documentation of each model has about the same level of detail. The TRARR documentation, though, contains terms different from the terminology used in the United States, and the precise meaning is not always clear. In view of the fact that use of these models has increased in the recent past and

TWOPAS

- Reflection of Input Data
- Summary of Specified
  - Simulation times
  - Flow rates
  - Desired speeds
  - Vehicle characteristics
- Detail Road Characteristics
- Desired Speeds for Different Vehicle Types
- Actual Measured Desired Speeds
- Average Speeds
- Operating (85th percentile) Speeds
- Speeds for Zero Traffic
- Speeds on Straight and Level Alignment
- Travel Times for Straight and Level Alignment
- Travel Times for Zero Traffic
- Geometric Delay
- Traffic Delay
- Traffic Snapshots
- Flow Rates at Finish Line
- Time Margins (to oncoming vehicle) in Passes and Pass Aborts
- Passes Started and Aborted
- Platoon Leader Vehicle Types at Finish Lines
- Percent of Time Unimpeded
- Headways at Beginning and Finish Lines
- Platoon Sizes at Finish Lines
- Overtakings Classified According to Speed Distributions
- Overtakings Classified According to Initial Acceleration
- Selected Output for User-Selected Stations
- Selected Output for User-Specified Subsections

TRARR

- Simulation Time
- Specified Percent Following
- Input Flow Rates
- Actual Flow Rates
- At Intermediate Positions
  - Overtakings commenced
  - Spot mean speeds
  - Percent following
- Travel Times
- Journey Mean Speeds
- Percent Time Spent Following
- Number and Rate of Overtakings
- Average Desired Speeds
- Unimpeded Speeds (only accounting for road speed indices)

FIGURE 2 TWOPAS and TRARR output features.

will probably increase in the future, further documentation and increased user-friendly features would be useful.

The TRARR model is supported in the United States by the University of Calgary. It is not clear who is responsible for providing support for the TWOPAS model, but Doug Harwood of the Midwest Research Institute has provided support for this project.

Both models are coded in FORTRAN. Both TRARR and TWOPAS were run on an IBM PS/2 386-55SX computer with 640K of RAM memory. The run time depends on the simulation time, length of road, and volume of vehicles. TWOPAS usually takes longer to execute because it is a larger program.

Ease of Use

Neither model is easy to use if one is unfamiliar with the background and underlying theory. Initially TRARR was easier to use because no programming bugs were discovered, whereas some debugging of TWOPAS had to be carried out. However, as the project progressed, it was found that TWOPAS was easier to use, particularly because changes to the geometric alignment could be made more readily.

COMPARISON WITH FIELD DATA

Study Sites and Data Collection

The two study sites are both in northwestern California in Caltrans District 1. District 1 provided the guidance for this project and also helped with the site selection and data collection.

The following criteria were established for the site selection:

1. Design speed of 80 km/hr (50 mph). Caltrans actually uses the average highway speed (AHS), which is the weighted average of the design speed within a highway section. On nonengineered roads the average highway speed is estimated.
2. Length greater than 1609 m (2 mi). This criterion was established to meet the minimum length specification for general terrain segments as defined in the *Highway Capacity Manual* (HCM) (*5*).
3. Good vantage points should be available for observation.
4. As-built drawings must be available beyond the limits of the test section.
5. Traffic volumes should vary significantly throughout the day, with some high-volume time periods.
6. The section should not include locations of major turning movements.
7. There should be no passing lanes or four-lane sections within 1609 m (2 mi).
8. Horizontal and vertical geometry should be fairly consistent throughout the test section.

The original intent was to have one site that met the requirements for mountainous terrain and a second that met the requirements for rolling terrain, as defined in the HCM. Most of the roads were designed and constructed many years ago, so it was impossible to find as-built drawings for many of the available sites. It was not possible to obtain a mountainous site in this district that met the length requirement, although the region has several mountainous areas. A level terrain site was therefore substituted for the mountainous terrain site. The principal characteristics of the two sites are given in the following:

- MEN101 (on US-101 in Mendocino County, PM 96.42 to PM 98.52)
  1. Rolling terrain
  2. Length = 2851 m (9,355 ft)
  3. 100 percent no-passing zones
  4. Traffic flow data
     a. Direction 1
        (1) Flow rate = 451 vph
        (2) 4.0 percent trucks
        (3) 4.7 percent RVs
     b. Direction 2
        (1) Flow rate = 369 vph
        (2) 5.2 percent trucks
        (3) 5.4 percent RVs
- LAK20 (on State Route 20 in Lake County, PM 19.22 to PM 23.03)
  1. Level Terrain
  2. Length = 6,118 m (20,072 ft)
  3. 94 percent no-passing zones
  4. Traffic Flow Data:
     a. Direction 1
        (1) Flow rate = 210 vph
        (2) 11.9 percent trucks
        (3) 1.9 percent RVs
     b. Direction 2
        (1) Flow rate = 272 vph
        (2) 10.6 percent trucks
        (3) 4.3 percent RVs

The data were collected by placing video cameras at each end of the section and one at approximately the midpoint. The videotapes were subsequently analyzed to obtain the following information:

1. Fifteen-minute flow rates at the beginning of the section, by vehicle type.
2. Average travel time through the section.
3. Percentage of vehicles in platoons at each measurement station. A vehicle was considered to be in a platoon if it was within 5 sec of a leading vehicle. This definition conforms to the definition of the field measurement of percentage time delay, according to the HCM (*5*).

Model Calibration and Validation

For present purposes, the term "model calibration" defines the phase of the comparison with field data in which changes are made to model input, which would not normally be varied, and to the model itself to obtain the best correspondence with field data at a site or sites. During the validation phase, the intent is to determine whether the model is transportable, that is, whether the model can simulate traffic operations at another site or sites without essential modification to the model

or input data that would not normally be varied. No changes to the model itself are made, and only input data directly related to the characteristics of the sites are changed. Major structural modifications to the models were outside the scope of the project.

Both models were calibrated against data obtained at the MEN101 site and validated against data obtained at the LAK20 site. The measures used to test the performance of the models were

1. Travel time over the test section for each direction.
2. Percentage time delay as measured by the model. This is the percentage of time that each vehicle is impeded by another vehicle while traveling over the test section. This is the definition of percentage time delay used in the HCM (*5*).
3. Percentage time delay as measured in the field. This is the percentage of vehicles within 5 sec of a leading vehicle, measured at a point on the highway. This is the surrogate measure recommended in the HCM for representing percentage time delay.

Model Calibration Input

Some of the input data required were not available and had to be estimated. Care was taken to determine either that the models were not sensitive to possible errors in the estimates or that the data were representative of the conditions in the field.

The geometric data available for the MEN101 site extended not far beyond the limits of the test section. Experience with the models, particularly with TWOPAS, indicated that fairly long warm-up sections are required to produce realistic results. Warm-up sections of 1072 and 1622 m (3,520 and 5,322 ft) were used, with a warm-up time of 15 min. The test simulation time was 1 hr.

Passing sight distances were unavailable for the sites. It was assumed that the passing sight distance recommended by AASHTO (*9*) was available. Passing is not allowed at the MEN101 site. Experimentation with the models indicated that the results were not sensitive to the choice of different passing sight distances.

The input requirements for vehicle characteristics had to be treated differently for the two models. As mentioned before, four types of trucks and buses, four types of RVs, and five types of automobiles can be specified for the TWOPAS model. Both vehicle characteristics and proportions of the different vehicle types can be specified. Through communication with Doug Harwood of the Midwest Research Institute, it was established that the subgroups of trucks, RVs, and automobiles represent the performance of these subgroups of vehicles but that the default values for the specific vehicle characteristics do not describe specific vehicle types found on the road. For the truck subgroup, for instance, the collective individual vehicle types in the truck group, as specified through the default characteristics, represent the performance of the total truck population on the road.

The calibration of TWOPAS was carried out using the default values for the vehicle characteristics and also the default values for the distribution of each vehicle type within the subgroup. These default values are identical to those used in the simulation experiment underlying the data used in the HCM to conduct level of service analysis (*3*). Field data were used to determine the proportions of trucks, RVs, and automobiles.

After the calibration was completed, another simulation was carried out using another vehicle type distribution. Vehicle types were identified that most closely resembled the default values for vehicle characteristics. Subsequently the observed traffic was classified according to these vehicle types. No significant deviation from the results obtained previously was observed.

The calibration of TRARR was carried out similarly with vehicle characteristics. In the case of TRARR, 18 vehicle types and their proportions of the traffic stream can be specified. The calibration was first carried out using the TRARR default values for vehicle characteristics and their proportions within the subgroups of trucks and automobiles. RVs were classified as trucks. As in the case of TWOPAS, a simulation was also carried out using the proportions of vehicle types observed in the field.

The observed vehicles were classified according to the default vehicle characteristics. This was difficult to accomplish, because some of the vehicles described in the program documentation are particular to Australia. For instance, the Australian vehicle types do not explicitly provide for the RVs found in the United States. The results obtained using this distribution were, however, better than those based on the default distribution. Consequently, the distribution based on the field data was used for the calibration and validation of TRARR.

It was found that the percentage time delay predicted by the model was very sensitive to the percentage of entering vehicles that are in platoons. In many cases, the predicted percentage time delay was almost the same as the percentage of entering vehicles in platoons. As a consequence, the percentage of vehicles in platoons, as measured in the field at the beginning of the test sections, was specified as entering the warm-up sections.

The mean of the desired speeds of the drivers was determined by simulating the traffic and finding the desired speed that led to the best results for the travel time distribution. In the case of the TWOPAS model, an adjustment was also made to the model itself. The limiting desired speed in horizontal curves was increased by 6.4 m/sec (21 ft/sec). A mean desired speed of 81.6 km/hr (51 mph) combined with a standard deviation of 8 km/hr (5 mph) yielded the best results for TWOPAS. The results discussed in the following section are based on a mean desired speed of 96 km/hr (60 mph) and a standard deviation of 8 km/hr (5 mph) for TRARR.

Default values were assumed for the remainder of the variables.

Model Calibration Results

The results of the model calibration are presented in Figures 3 through 6. From Figures 3 and 4 it can be seen that the TWOPAS output for travel time corresponds well with the field data for both Directions 1 and 2. Figures 3 and 4 also indicate that the TRARR output for Direction 2 corresponded well with the field data but that there is a large difference in

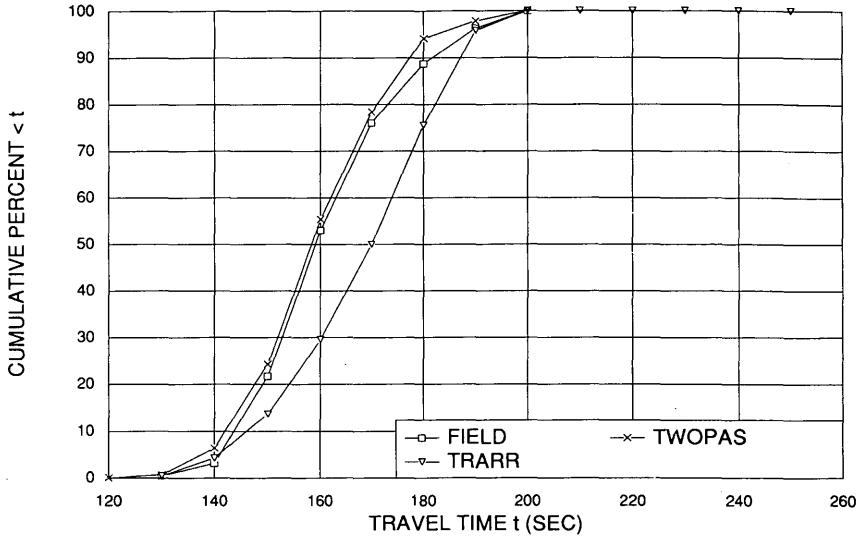


FIGURE 3 Calibration of travel time on MEN101, Direction 1.

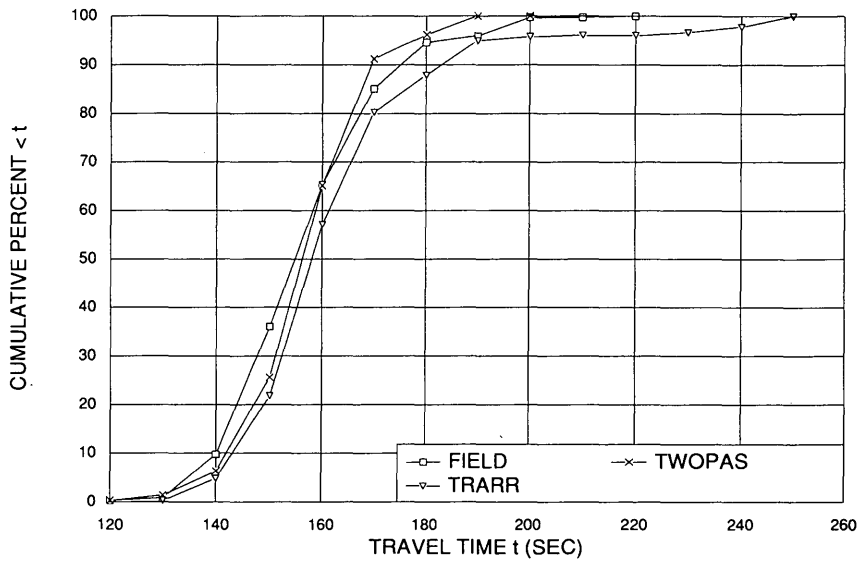


FIGURE 4 Calibration of percentage time delay on MEN101, Direction 1.

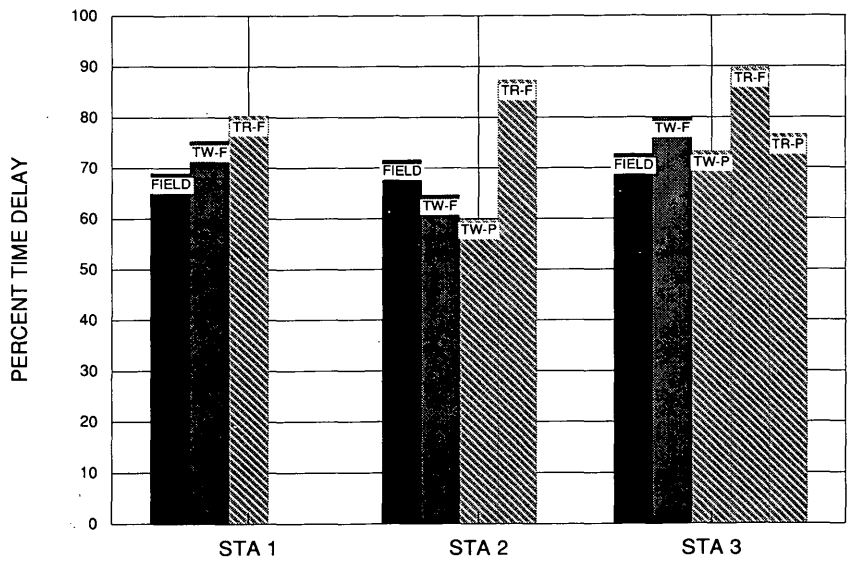


FIGURE 5 Calibration of travel time on MEN101, Direction 2.

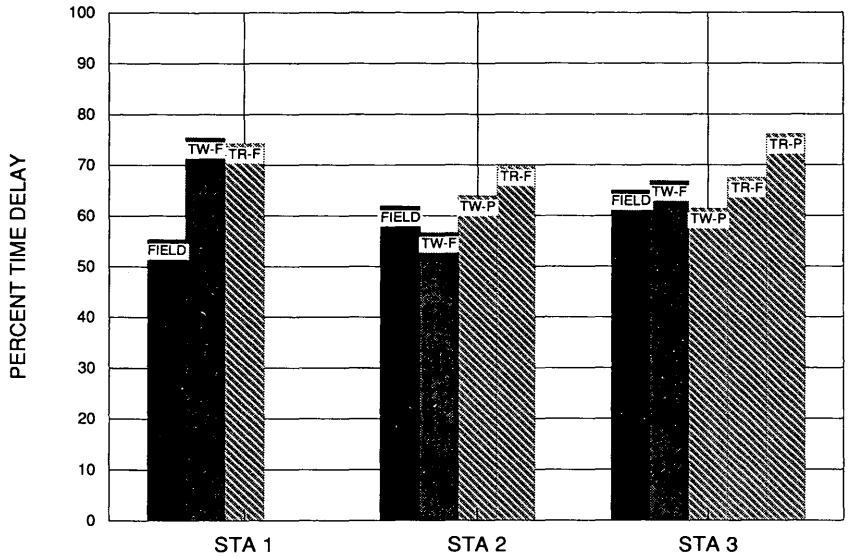


FIGURE 6 Calibration of percentage time delay on MEN101, Direction 2.

Direction 1. These were the best results that could be obtained, given the input data variations discussed in the previous section. Travel time was used as the primary measure for the calibration.

A comparison between the percentage time delay measured in the field (shown as FIELD) and the output from the models is presented in Figures 5 and 6 for the two directions. The model predictions for the percentage time delay, measured over space, are shown as TW-P and as TR-P for TWOPAS and TRARR, respectively. The point measurements are shown as TW-F and TR-F. STA 1 refers to the entrance to the test section, STA 2 to an intermediate point approximately halfway between the beginning and end of the test section, and STA 3 to the end of the test section. TRARR lacks the capability to produce a measurement for the "program definition" at an intermediate point.

Additional output was created to measure the percentage of vehicles in a platoon at a point with the models. The times at which vehicles passed a station were recorded with the aid of the models. Using a separate computer program, a determination was made whether a vehicle was within 5 sec of a leading vehicle.

From the comparison between the field data and the model results for percentage time delay, it can be seen that TRARR generally overestimated the percentage time delay. In some cases TWOPAS underestimated the percentage time delay, whereas in others it produced an overestimation, but generally it produced a closer estimate than TRARR. It is also noteworthy that the percentage time delay measured over space corresponded more closely with the field measurements than the model prediction of percentage time delay measured at a point. This is the reverse of what may be expected.

Model Validation Input

In the case of the LAK20 site, warm-up sections of 1074 and 1465 m (3,522 and 4,806 ft) were used, again with a warm-up time of 15 min. As with the calibration, the simulation time was 1 hr.

It was again assumed that the passing sight distance recommended by AASHTO was available. It was thought that because passing is permitted over 6 percent of the length of the road, the results could have been sensitive to the available passing sight distance. Experimentation with different passing sight distances indicated that the results were not very sensitive to this factor.

Default values were assumed for vehicle characteristics as well as for the distribution of vehicle types within the subgroups of trucks, RVs, and automobiles for TWOPAS. For TRARR, the same procedure was used to specify the vehicle type distribution as was described for the calibration.

Desired speeds were kept at the same values as used in the calibration. Default values were used for the remainder of the variables.

Model Validation Results

The results of the model validation are shown in Figures 7 through 10. TWOPAS estimates of travel time corresponded well with the field data, whereas the TRARR estimates differed substantially in both the mean and the profile of the travel time distribution. These results are presented in Figures 7 and 8.

Experimentation with the mean and standard deviation of the desired speed distribution was undertaken to obtain better results for TRARR. Some improvement was obtained by specifying what could be considered an unreasonably high mean desired speed of 112 km/hr (70 mph). This could indicate that structural changes to the model, recalibration of coefficients internal to the model, or reevaluation of other default values should be considered. Because the TWOPAS model yielded satisfactory results, these possible changes were not considered.

The comparisons of model predictions of the percentage time delay and the field measurements exhibited essentially the same patterns found during the calibration stage. The TWOPAS predictions again were better than the TRARR predictions. The space measurement of TWOPAS compared

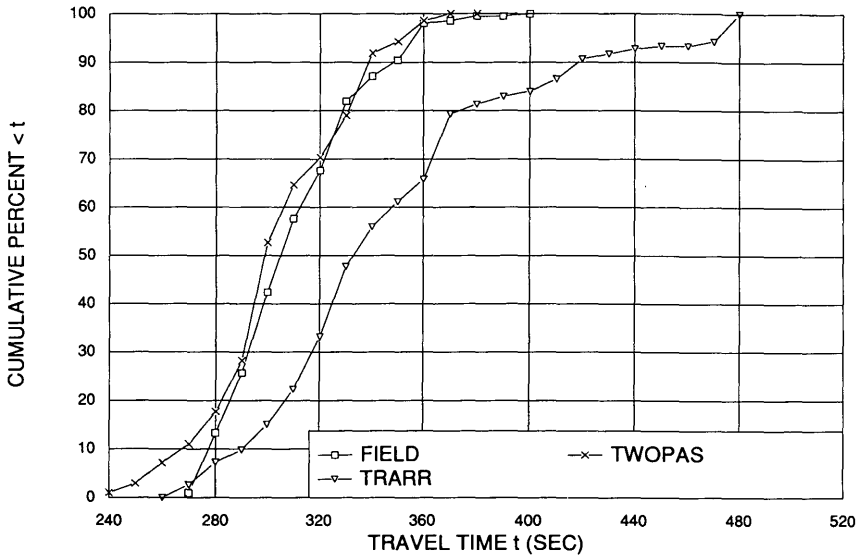


FIGURE 7 Validation of travel time on LAK20, Direction 1.

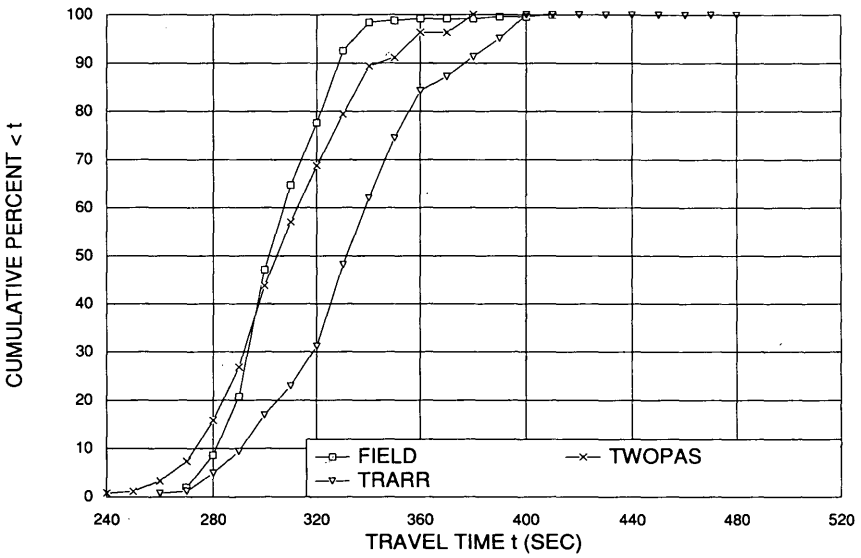


FIGURE 8 Validation of travel time on LAK20, Direction 2.

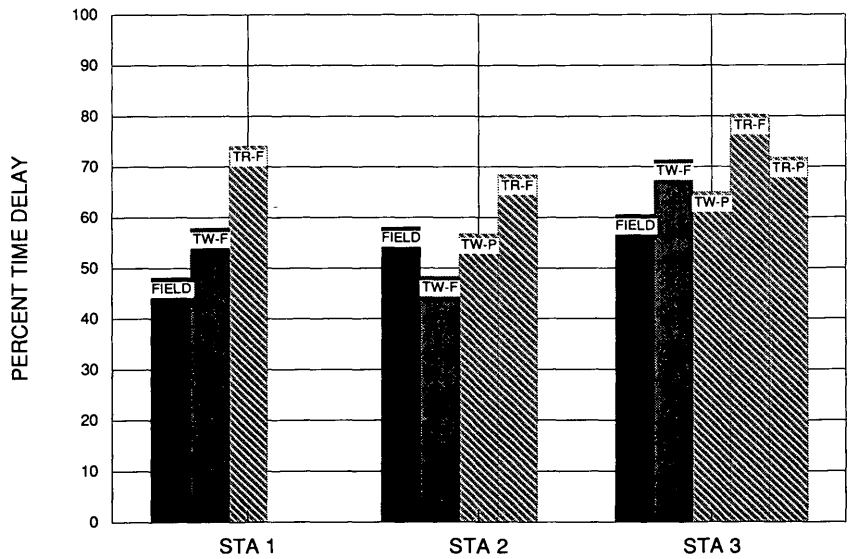


FIGURE 9 Validation of percentage time delay on LAK20, Direction 1.

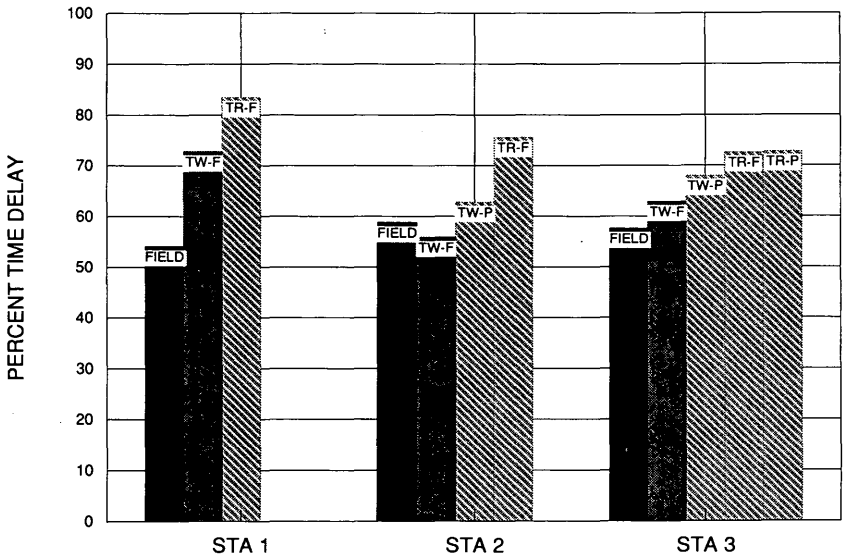


FIGURE 10 Validation of percentage time delay on LAK20, Direction 2.

better with the field point measurement than did the model prediction of the point measurement. It is noteworthy that the TWOPAS space measurement is generally remarkably close to the field point measurement.

DISCUSSION OF RESULTS

As was stated earlier, simulation models appear to be the only affordable way to analyze extensive and complex situations on two-lane roads. Much experience is needed in this area, and it would be useful to continue the development of simulation models and related theories on several fronts. The funding level for work in this area has, however, been very low compared with, for example, research in the area of freeway operations. As a result, there has been some discussion about selecting one two-lane road simulation model to be developed further for application in the United States, to make the best use of available resources.

The results of the comparison of the two models for two-lane highways with low design speeds indicate that TWOPAS performed better. Nevertheless, it may be premature to select TWOPAS for all future work, since clearly both models require further development before they can be applied without reservation to the many different situations that arise in two-lane road operations.

It is useful, however, to provide some perspective on a possible choice of model on the basis of the experience gained in this project. The strong arguments for TRARR revolved around the fact that it alone operated on a PC. That argument is no longer valid. A disadvantage of the TRARR model is that some of the parameters, and particularly the driver and vehicle characteristics, are Australian, whereas the TWOPAS model was created in the United States. It is notable that the TWOPAS vehicle characteristics also may warrant reexamination, as the current default values for vehicle characteristics do not represent specific vehicle types found on the road.

Notwithstanding the fact that it may be premature to commit to one model for use in the United States, there is one very strong argument for making a tentative commitment to the TWOPAS model at this stage. TWOPAS has been used to generate the basic values used in the HCM (*5*) for level of service and capacity analysis, and it is essential that further work in this area be consistent with past work unless the values in the HCM are to be discarded. This appears unlikely in the near future and it does not appear, from the results presented in this paper, that TRARR offers any significant advantage over TWOPAS in this area. It is therefore recommended that TWOPAS be adopted for work related to analysis of capacity and level of service in the areas of two-lane highways. Caution should be exercised in applying the model for this purpose, because its accuracy is questionable in predicting percentage time delay, which is the primary measure used in the HCM (*5*) for capacity and level of service analysis.

SUMMARY OF MAJOR CONCLUSIONS AND RECOMMENDATIONS

1. The TWOPAS and TRARR models are generally comparable in their capability to simulate traffic operations on a two-lane, two-way highway.
2. The application of the TRARR model to highways with low design speeds was carried out with minimal problems, whereas the TWOPAS model had to be debugged on several occasions. It should be noted that the TWOPAS model was converted from a mainframe-based model to run on a PC.
3. TWOPAS simulation results compared better with field data for 80-km/hr (50-mph) design speed roads. This was the case for both a level terrain site and a rolling terrain site. The comparison was made in terms of travel time and percentage time delay.
4. Both models require further work before they can be applied without reservation to the many situations that might arise on two-lane roads and to many possible user require-

ments. Both models should be made more user-friendly. Other major improvements needed for TRARR include improvements in the way it handles horizontal curves, conversion of vehicle types to U.S. types, verification of driver characteristics to ensure conformity with U.S. drivers, and calibration of the factors that influence speeds and percentage time delay. The TWOPAS default values for vehicle characteristics should be modified to represent actual vehicles within the subgroups of trucks, RVs, and automobiles. Although TWOPAS performed better than TRARR in the simulation of vehicle speeds and percentage time delay, further calibration and validation of the model would be beneficial.

5. TWOPAS should be tentatively adopted for analysis in the United States related to capacity and level of service in the short term. After further development of the models, this decision may be reevaluated.

ACKNOWLEDGMENTS

The research was funded by Caltrans and the Department of Civil Engineering and Applied Mechanics of San Jose State University. The authors would like to thank Fred Rooney, Rick Knapp, Pat Secoy, Guy Luther, Ken DeCrescenzo, and Paul Vonada, all of Caltrans, for their help throughout the project. A special word of thanks is also due to Doug Harwood of the Midwest Research Institute for his interest and help with the implementation of the TWOPAS model.

REFERENCES

1. A. D. St. John and D. W. Harwood. *TWOPAS User's Guide.* FHWA, U.S. Department of Transportation, May 1986.
2. A. D. St. John and D. W. Harwood. *TWOPAS Programmer's Guide.* FHWA, U.S. Department of Transportation, July 1986.
3. C. J. Messer. *Two-Lane, Two-Way Rural Highway Capacity.* Final Report. Texas Transportation Institute, College Station, Feb. 1983.
4. C. J. Messer. *Two-Lane, Two-Way Rural Highway Level of Service and Capacity Procedures.* KLD Associates; Texas Transportation Institute, College Station, Feb. 1983.
5. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
6. C. J. Hoban, G. J. Fawcett, and G. K. Robinson. *A Model for Simulating Traffic on Two-Lane Rural Roads: User's Guide and Manual for TRARR Version 3.0.* Technical Manual ATM 10A. Australian Road Research Board, Victoria, May 1985.
7. J. F. Morrall and A. Werner. Measuring Level of Service of Two-Lane Highways by Overtakings. In *Transportation Research Record 1287,* TRB, National Research Council, Washington, D.C., 1990, pp. 62–69.
8. M. Nogueron-Espinosa and A. D. May. Investigating Rural Highway Design Alternatives with Simulation. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1992.
9. *A Policy of Geometric Design of Highways and Streets.* AASHTO, Washington, D.C., 1990.

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

Effects of Location on Congested-Regime Flow-Concentration Relationships for Freeways

PAUL HSU AND JAMES H. BANKS

Two questions related to the effect of location on congested-regime flow-occupancy data and relationships are investigated. A theory accounting for the effect of traffic entering or exiting the freeway between the point of observation and the bottleneck is presented and verified in general terms. According to this theory, maximum flows will decrease in the upstream direction when flow entering exceeds flow exiting and vice versa. The effect of lane drops that occur in the section occupied by the queue is also investigated. It is found that when locations upstream and downstream of the lane drop are compared, the approximate slopes of the congested regime of the flow-occupancy relationship are roughly proportional to the reciprocal of the number of lanes. This implies that average densities (and occupancies) remain approximately unchanged across the lane drop. As this issue was investigated only for the case of drops from five to four lanes, the conclusion must be considered to be tentative.

Recent research has led to a revised understanding of freeway speed-flow and flow-concentration relationships. In a synthesis of this work, Hall et al. (*1*) propose three basic regimes: normal uncongested flow, congested flow, and flow that is accelerating in or just downstream from bottleneck sections (see Figure 1). This interpretation of the data rests on the realization that the entire speed-flow or flow-concentration (flow-density or flow-occupancy) relationship cannot be observed at any one point; instead, the overall relationship must be synthesized from data taken from several different locations, and careful attention must be paid to the way the location affects the interpretation of the data. General explanations of the effect of location on the nature of the data date to Edie and Foote (*2*) and May et al. (*3*), and these efforts have more recently been summarized by May (*4*, p.288).

Two questions related to the effect of location on congested-regime flow-occupancy data and relationships are examined in this paper. Flow-occupancy relationships were investigated primarily because the raw data, which were produced by the San Diego ramp metering system, included flow and occupancy but not speed. The flow-concentration relationship also was chosen because it lends itself readily to theoretical interpretations based on driver behavior. Note that the relationship between density and occupancy is linear except in cases in which vehicle lengths and speeds are correlated; as in the work by Hall et al., the two measures will be considered to be interchangeable. The congested regime was studied because it is perhaps the least understood of the three proposed by Hall and because data interpretation for this regime appeared to involve unresolved (or partially resolved) location-related issues. The questions addressed here are (a) the effect of traffic entering or exiting the freeway between the point of observation and the bottleneck, and (b) the effect of lane drops that occur in the section occupied by the queue.

The first of these questions has been raised in the past, but not clearly. Both Hall et al. (*5*) and Banks (*6*) recognize that apparent discontinuities in the flow-occupancy relationship upstream of bottlenecks may result from the action of the queue in limiting maximum flow at such locations. Also, Banks hypothesized that speed-flow and speed-concentration relations would appear to be truncated in flow upstream of the bottleneck; presumably, this truncation results from traffic entering between the bottleneck and the point of observation, but this is not stated explicitly. More recently, Hall et al. (*1*) propose a three-dimensional diagram of flow and occupancy versus distance in which flows decrease and occupancies increase just upstream of an onramp that adds traffic to a queue; however, they do not elaborate on this point. One objective of this study was to further develop and clarify theory related to the effect of traffic entering or exiting a queue between the bottleneck and the point of observation and to further verify that the truncation does occur.

The second question does not appear to have been raised in the past and was considered here primarily because the study sites happened to contain lane drops in the sections upstream of the bottlenecks. At first glance, there appears to be no reason that the flow-occupancy relationship should be affected by the number of lanes occupied by a queue except that as total flow is limited by the capacity of the bottleneck, flow per lane in the queue upstream should be inversely proportional to the number of lanes. It was found, however, that such lane drops also resulted in shifts in occupancy, at least at the sites studied here.

THEORETICAL BACKGROUND

A basic assumption in the discussion that follows is that there *is* a flow-occupancy relationship in the congested regime—that is, for every level of occupancy there corresponds a particular level of flow. This assumption has not been universally accepted. Congested-flow data tend to display a great deal of scatter, so the relationship can be said to exist only between data averaged over fairly long periods of time. Moreover, not

P. Hsu, California Department of Transportation, 2829 Juan Street, San Diego, Calif. 92186-5406. J. H. Banks, Civil Engineering Department, San Diego State University, San Diego, Calif. 92182-0189.

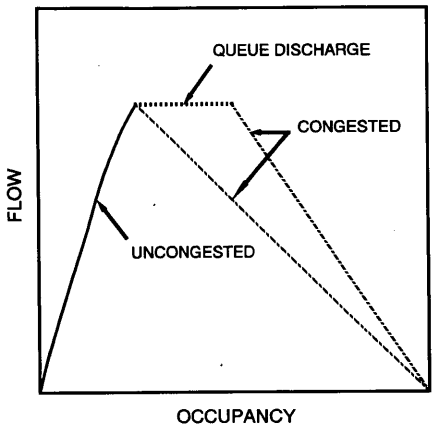


FIGURE 1 Generalized flow-occupancy relationships.

all the variation in the relationship is strictly random, because there appear to be comparatively definite patterns of wave action in congested flow. Congested regime data tend to cluster in certain regions of the flow-occupancy diagram, and some trend can be identified. That being the case, it can at least be said that there should be some sort of distribution of flow for each level of occupancy and vice versa.

There is also no universal agreement on the nature of the flow-concentration trend in congested flow data. Most traditional representations show it as concave to the origin, but Koshi et al. (*7*) show it as convex and Banks (*6*) suggests that it is linear. One difficulty in identifying the shape of the trend line is that data for very high occupancies and very low flows are hard to obtain, since these conditions tend not to occur on the freeways that have been studied recently.

The exact shape of the relationship is not very important to the issues considered here, but the behavioral interpretation of certain features of the relationship is. Banks (*6*) has pointed out that a linear congested-regime flow-concentration relationship, in which the trend line passes through zero flow and 100 percent occupancy (or jam density), implies that the average time gaps between vehicles do not vary with flow. This, in turn, implies that distance gaps vary linearly with speed. This can be extended to state that any straight line on the flow-occupancy diagram passing through zero flow and 100 percent occupancy corresponds to a constant time gap, and that as the line rotates in a clockwise direction, the average time gap decreases. The relationship between the time gap and the slope of any such line is

$$g = -36/b \tag{1}$$

where g is the average time gap in seconds per vehicle per lane and b is the slope of the line in vehicles per hour per lane per percentage occupancy (*8*).

The definition of the time gap discussed elsewhere (*6*) assumes that drivers assign a buffer of fixed length to the vehicle in front of them to provide a margin of safety so that, in behavioral terms, vehicles have effective lengths greater than their physical lengths. Meanwhile, the detectors used to measure occupancy have nonzero length, and the electrical lengths of vehicle are not identical to their physical lengths, so that, for measurement purposes, there is also an effective vehicle length that is greater than the physical length. The assumption that the trend line passes through zero flow and 100 percent occupancy is based on the assumption (*6*) that these two effective lengths are approximately equal. In fact, they are probably different.

If they are different, constant time gaps (in the behavioral relationship) still correspond to a linear congested-regime flow-concentration relationship, but zero flow occurs at some occupancy other than 100 percent.

Let

H_M = measured occupancy;
H_B = "behavioral" occupancy—that is, the fraction of time a point is occupied by vehicles plus any distance buffers assigned to them by drivers of trailing vehicles;
L_M = average effective vehicle length for purposes of measurement;
L_B = average behavioral effective vehicle length—that is, the physical length of the vehicle plus the buffer;
h = average time headway;
q = flow rate = $1/h$; and
u = space mean speed.

Measured occupancy equals the time that the detector is occupied divided by the time headway. If both the time gap and the effective vehicle length are defined in behavioral terms

$$h = L_B/u + g \tag{2}$$

Inverting Equation 2,

$$q = \frac{1}{h} = \frac{u}{L_B + gu} \tag{3}$$

Meanwhile

$$H_M = \frac{L_M/u}{g + L_B/u} = \frac{L_M}{L_B + gu} \tag{4}$$

Solving Equation 3 for u leads to

$$u = \frac{qL_B}{1 - gq} \tag{5}$$

Substituting for u in Equation 4 and simplifying results in

$$H_M = \frac{(1 - gq)L_M}{L_B} \tag{6}$$

Meanwhile, however, the average time gap can be defined as one minus the occupancy (the time not occupied) divided by the flow, or

$$g = \frac{1 - H_B}{q} \tag{7}$$

Solving this for H_B leads to

$$H_B = 1 - gq \tag{8}$$

Substituting for $1 - gq$ in Equation 6 thus leads to

$$H_M = (L_M/L_B)H_B \quad (9)$$

which implies a linear relationship between the measured and behavioral occupancy. If $L_M < L_B$, zero flow (and thus jam density) corresponds to a measured occupancy less than 100 percent; conversely, if $L_M > L_B$, a measured occupancy of 100 percent is reached at a flow greater than zero. In either case, if the behavioral flow-concentration relationship has a linear trend, so will the measured one. This is of some importance, because it appears from the data considered here that most of the congested-regime flow-occupancy trend lines, if linear, pass through zero flow at measured occupancies of less than 100 percent.

The preceding analysis allows all major features of the congested-regime flow-concentration relationship to be interpreted in average time gaps, although it is important to distinguish among behavioral, measured, and physical gaps. The curvature of the trend line expresses the relationship between the average time gap and flow or speed: concave relationships imply that time gaps increase with increasing speed or flow and convex ones imply that they decrease. The scatter in the data corresponds to the variation in time gaps for different groups of vehicles under similar average flow or occupancy conditions. Any rotation or shifting of the trend line between different locations may be related to increases or decreases in the average time gaps and should occur only where there is a reason for this feature of driver behavior to change.

This analysis leads to the expectation that the underlying flow-occupancy relationship will be relatively stable from one location to another as long as driver behavior is stable. There will still be differences in the appearance of the data, however, because the queueing process will act to limit maximum flows at various points upstream of the bottleneck.

Figure 2 depicts a typical freeway section. For the freeways considered here, data are collected by loop detectors that are normally located just upstream of the on-ramps. In this case, the objective is to determine the relationship between q_k and q_{k-1}, the mainline flows at successive detector stations; let $\Delta q = q_k - q_{k-1}$. Ramp flows are q_{on} and q_{off}; let $q_r = q_{on} - q_{off}$. If queues are present and flows are averaged over sufficiently long time intervals, we expect $\Delta q = q_r$, and

$$q_{k-1} = q_k - q_r \quad (10)$$

Equation 10 establishes the relationship that is to be expected between data taken at adjacent detector locations. At both locations, data are expected to be scattered, but the region occupied by data at the upstream station is expected to be offset from that at the downstream station by a flow equal to the average value of q_r. Note that q_r is positive when traffic entering exceeds traffic exiting and vice versa. The data considered here were taken during the morning peak period, when q_r was positive for almost all sections; consequently, it is to be expected that maximum flows will decrease in the upstream direction. This appears to have also been true for the cases considered in previous literature. Note however, that q_r could be negative and, in that case, maximum flows should increase in the upstream direction. This could happen, for instance, for an afternoon-peak bottleneck created by a lane drop.

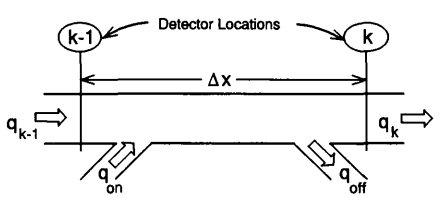


FIGURE 2 Typical freeway section.

DATA

The effects of location on congested-regime flow-occupancy relationships were investigated using morning peak period data from two sections upstream of freeway bottlenecks in San Diego. These sections were Interstate 8 from Jackson Drive to Waring Road and Interstate 805 from Home Avenue to El Cajon Boulevard. Figures 3 and 4 are schematic diagrams of these locations, showing lane configurations and detector locations. Both bottlenecks have been studied extensively in the past (*6,9–11*). The bottleneck on Interstate 8 is just upstream of the College Avenue on-ramp, so this section actually extends to the second set of detectors downstream of the bottleneck; that on Interstate 805 is on an upgrade between the detectors at University Avenue and those at El Cajon Boulevard. When queues are present just upstream, data at the College Avenue detectors on Interstate 8

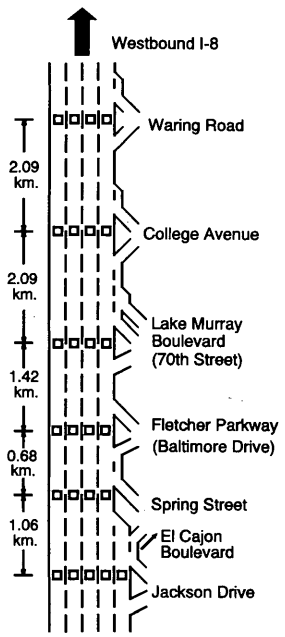


FIGURE 3 Lane configuration and detector location for I-8 (1 mi = 0.62 km).

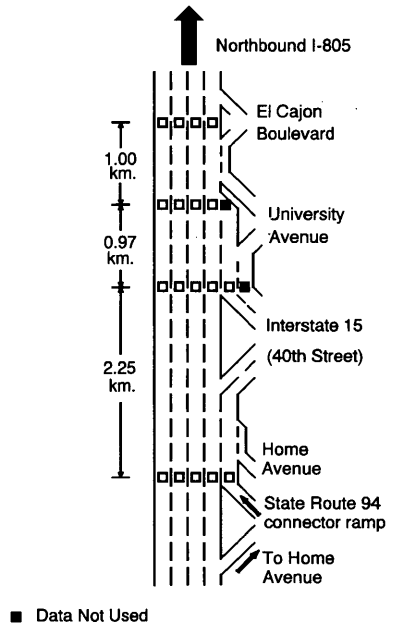


FIGURE 4 Lane configuration and detector location for I-805 (1 mi = 0.62 km).

and the El Cajon Boulevard detectors on Interstate 805 display the characteristics of the bottleneck-acceleration regime identified by Hall (*1*); those further upstream at both sites represent typical congested flow. In addition, some congested flow is encountered at all locations because of incidents downstream.

The data used were produced by the San Diego ramp metering system. Raw data consisted of volume counts and occupancies for each lane, which were recorded every 30 sec. These were aggregated across all lanes and over 6-min intervals, so as to be comparable with those in previous research (*6*). In two cases, there were special problems involved in the aggregation. There are five lanes at the University Avenue detectors on Interstate 805; at the next detector location upstream, just downstream from the branch connector from Interstate 15, there are six. In both cases, the rightmost lane was very lightly used, and data from these lanes were excluded on the grounds that their inclusion would distort volumes and occupancies averaged across all lanes. Unfortunately, use of the rightmost lane at these locations is affected by traffic conditions, with the heaviest use occurring in dense queues. Hence there may be some distortion of the data at these two locations, and they may not be fully comparable with the other locations.

Data were collected for the hours 6:00 to 9:00 a.m. during the months of August 1991 through February 1992. A total of 78 days of usable data were collected for Interstate 805 and 123 days for Interstate 8. Scatter plots of flow versus occupancy were prepared for the data, and data from different detector locations were compared by drawing approximate outer envelopes to the data and by fitting trend lines to the congested-regime data by visual inspection. Outer envelopes were constrained to pass through zero flow and 100 percent occupancy. Two sets of trend lines were prepared: one was constrained to pass through zero flow and 100 percent occupancy and another was unconstrained in its horizontal intercept. A complete description of the data and data analysis process, including scatter plots for all locations, may be found in the literature (*8*).

RESULTS

Figures 5 and 6 show approximate outer data envelopes for each location in the two sections. These tend to verify the theory that maximum flows decrease in the upstream direction in cases in which entering flow exceeds flow exiting. In addition, it can be seen that in some cases the slopes of the data envelopes vary by location.

A more appropriate indicator of the slope of the congested regime is the slope of the trend line. Figures 7 through 10 are sample scatter plots showing both constrained and unconstrained trend lines for four locations upstream of the bottleneck on Interstate 8. As can be seen, the unconstrained trend lines intersect the horizontal axis at occupancies less than 100 percent. Horizontal intercepts for all locations considered (including the ones not covered by Figures 7 through 10) were at occupancies of between 90 and 98 percent, with about 94 percent being most common. This would seem to indicate that

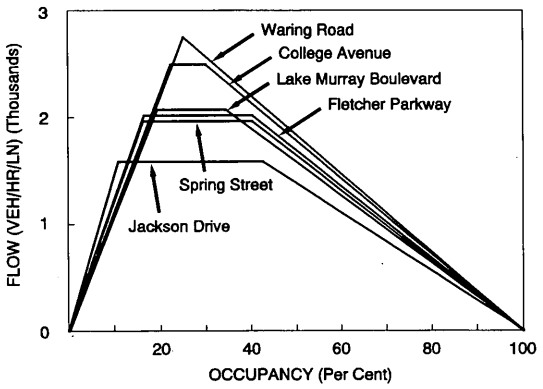


FIGURE 5 Data envelopes for Interstate 8 sites.

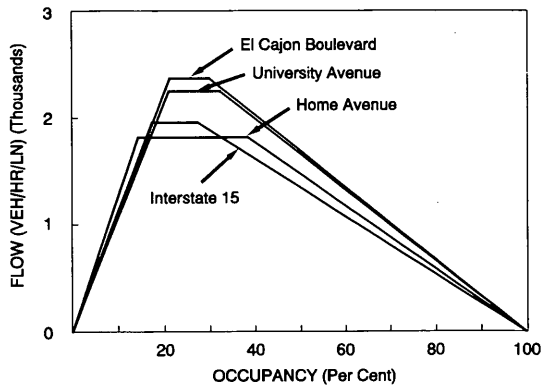


FIGURE 6 Data envelopes for Interstate 805 sites.

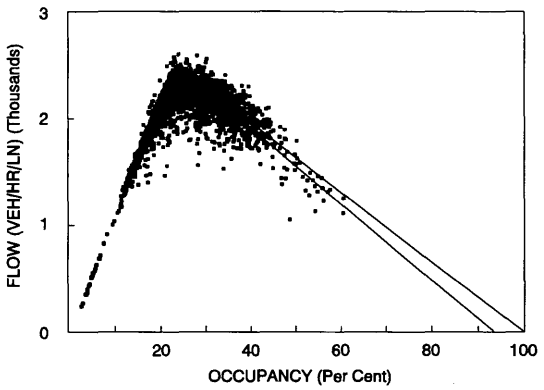


FIGURE 7 Flow-occupancy scatter plot, westbound Interstate 8 at College Avenue.

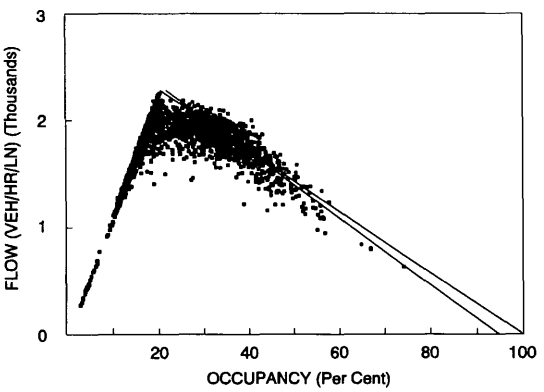


FIGURE 8 Flow-occupancy scatter plot, westbound Interstate 8 at Lake Murray Boulevard.

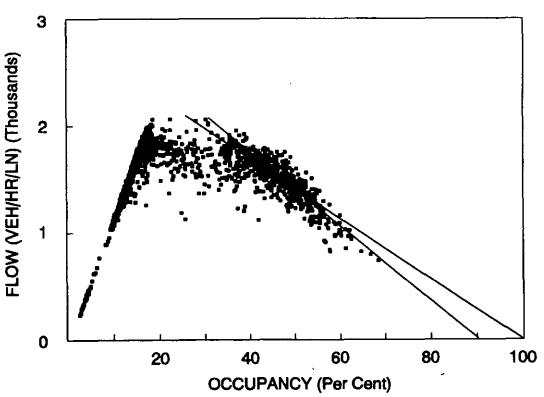


FIGURE 9 Flow-occupancy scatter plot, westbound Interstate 8 at Fletcher Parkway.

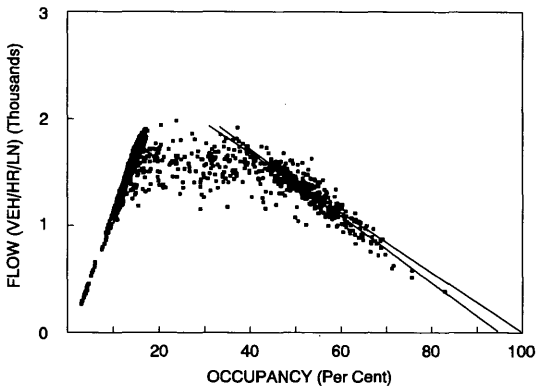


FIGURE 10 Flow-occupancy scatter plot, westbound Interstate 8 at Spring Street.

for the San Diego system, $L_B > L_M$; however, it should be emphasized that this conclusion is based on extrapolations of visually fitted lines, and is at best tentative.

Table 1 gives a summary of the constrained and unconstrained trend-line slopes for the various locations. It can be seen that there are minor variations in slopes among the sites having four lanes in one direction. In particular, constrained slopes at Waring Road and College Avenue on Interstate 8 (which are at or just downstream from the bottleneck) are noticeably steeper than those at 70th Street–Lake Murray Boulevard, Fletcher Parkway, and Spring Street, all of which are definitely upstream of the bottleneck. Constrained slopes for the four-lane sites on Interstate 805 (El Cajon Boulevard and University Avenue)—just upstream and just downstream of the bottleneck, respectively—fall between the two groups on Interstate 8. In all cases, however, slopes for five-lane sites are noticeably flatter than those for four-lane sites.

A rough idea of the relationship between number of lanes and slope of the congested-regime trend line can be gained by comparing the average of the slopes for all four-lane sites in Table 1 with that for all five-lane sites. The average constrained slope for four-lane sites is approximately -29.4, whereas that for five-lane sites is -24.0. For unconstrained slopes, similar averages are -33.5 and -26.7. The ratio of the first two numbers is 0.82 and that of the second, 0.80. In both cases the ratio is approximately 4 to 5—that is, roughly the same as the ratio of the number of lanes downstream to the number of lanes upstream.

This observation provides the basis for an interpretation of the overall effect of the lane drop on flow through the queue. Total flow across all lanes must be unchanged across the lane drop; consequently, as one moves downstream, flow per lane increases in proportion to the number of lanes upstream and downstream.

Let

$$\alpha = N_u/N_d$$

where N_u and N_d are the numbers of lanes upstream and downstream, respectively.

Then

$$q_d = \alpha q_u \qquad (11)$$

TABLE 1 Approximate Slopes for Congested-Regime Flow Concentration Relationships

Freeway	Location	No. of Lanes	Slope	
			Unconstrained	Constrained to (0,100%)
I-8	Waring Rd.	4	—	-31.0
	College Ave.	4	-35.5	-31.9
	70th St./Lk. Murray Blvd.	4	-31.7	-27.9
	Fletcher Parkway	4	-35.2	-27.9
	Spring St.	4	-31.7	-27.9
	Jackson Dr.	5	-24.2	-24.0
I-805	El Cajon Blvd.	4	-33.1	-29.0
	University Ave.	4	-33.8	-30.0
	I-15	5	-27.5	-22.9
	Home Ave.	5	-28.4	-25.0

— Insufficient data.

where q_u and q_d are the flows upstream and downstream of the lane drop. Meanwhile, if the trend line is linear and negatively sloped, the equation for q_u is

$$q_u = a - bk_u \tag{12}$$

where k_u is the concentration (density or occupancy) upstream of the lane drop. The rotation of the slope across the lane drop implies that

$$q_d \cong \alpha a - \alpha b k_d \tag{13}$$

Multiplying Equation 12 by α and substituting into Equation 11 yields

$$q_d = \alpha a - \alpha b k_u \tag{14}$$

which implies that $k_u \cong k_d$; in other words, concentrations (either densities or occupancies) are approximately unchanged across the lane drop. This can be verified by superimposing scatter diagrams of data taken just upstream and downstream of the lane drop on one another. Such diagrams are not presented here because the scatter in the data is great enough that data clusters for upstream and downstream locations overlap, making the diagram difficult to interpret.

Note that because $q = uk$, the relationship of speeds upstream and downstream of the lane drop is

$$u_d \cong \alpha u_u \tag{15}$$

Meanwhile, because average time gaps (however defined) are proportional to the reciprocal of the measured trend line, time gaps decrease across the lane drop, so that

$$g_d \cong g_u/\alpha \tag{16}$$

CONCLUSION

Two questions related to the effect of location on congested-regime flow-occupancy data and relationships have been addressed. The first of these was the effect of traffic entering or exiting the freeway between the point of observation and the bottleneck. It has been shown that the clusters of data representing congested flow at adjacent locations should be offset by a flow equal to the average difference between traffic entering and traffic exiting between the two locations. In the case in which flow entering exceeds flow exiting, this means that maximum flows will decrease in the upstream direction; where flow exiting exceeds that entering, the opposite effect is to be expected. This effect, however, is to be expected only over time intervals long enough so the average change in queue density is approximately zero. Data envelopes for several locations on two freeways in San Diego were presented that tend to verify this theory for the case in which flow entering exceeds flow exiting.

The second question concerned the effect of lane drops that occur in the section occupied by the queue. Both freeway sections studied here contained locations upstream of the bottleneck at which the number of lanes decreased from five to four. It was found (somewhat unexpectedly) that the approximate slopes of the congested-regime flow-occupancy relationship were roughly proportional to the reciprocal of the number of lanes at each detector location. This was shown to imply that average densities (and occupancies) are approximately unchanged across the lane drop, that speeds increase in the downstream direction by a factor proportional to the ratio of the number of lanes upstream and downstream, and that time gaps decrease by a factor proportional to the reciprocal of the numbers of lanes.

This result must be considered tentative, as it is based on only three lane drops in two freeway sections. In addition, in the case of Interstate 805, data from little-used lanes were excluded at two key locations, and this exclusion could have distorted the results, although it seems unlikely. If generally true, this finding raises a number of questions.

- First, is the change in the average time gap (and the corresponding rotation of the congested regime slope in the flow-concentration relationship) purely a local phenomenon caused by the lane drop, or are time gaps proportional to the number of lanes, even over extended sections? Two types of evidence could be brought to bear on this question. If the change in time gap is purely local, it should disappear farther

upstream. In the case of Interstate 805, one of the sites (Home Avenue) is a considerable distance upstream of the ultimate lane drop, at University Avenue. Note from Figure 3, however, that there is a four-lane section between the Interstate 15 branch connector and Home Avenue, so there are in fact two separate lane drops. Consequently, this type of evidence cannot be brought to bear without extending the study sections. As queues only rarely extend upstream of the locations included in this study, this appears to be impractical. In addition, this question could be addressed by comparing flow-occupancy diagrams for separate freeway sections with different numbers of lanes. Based on the limited amount of comparable data available, it appears unlikely that average time gaps in congested flow are proportional to the number of lanes as a general rule. For instance, the average time gap implied by the data in Figure 2 of Hall and Agyemang-Duah (*12*), for a site with three lanes in one direction, is at least as great as those for the five-lane sites reported here.

- Second, is the change in time gap really proportional to the number of lanes upstream and downstream, or does this apply only to drops from five lanes to four? This question could be addressed by finding lane drops involving different numbers of lanes in sections occupied by queues.
- Third, if the change in the average time gap is purely a result of the lane drop, would the reverse effect occur where lanes are added in a section occupied by a queue?
- Finally, at the microscopic level, what is the behavioral explanation for the change in time gap at the lane drop, and how should it be modeled? Past models of car-following behavior have assumed (or implied) that distance separations are continuous functions of the relative speeds of successive vehicles. Yet here is an example of a situation in which there is apparently a change in the relative speeds of successive vehicles but no change in the distance separation. The only plausible explanation seems to be that drivers increase their distance separations in advance of the lane drop, in anticipation of merging. A microscopic investigation of how and why they do so would be interesting.

ACKNOWLEDGMENTS

This paper is based on data provided by California Department of Transportation. Special thanks are due to Don Day of the Traffic Systems Division of the San Diego District of Caltrans for his assistance in making the data available.

REFERENCES

1. F. L. Hall, V. F. Hurdle, and J. H. Banks. Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992.
2. L. C. Edie and R. S. Foote. Traffic Flow in Tunnels. *HRB Proc.*, Vol. 37, 1958, pp. 334–344.
3. A. D. May, P. Athol, W. Parker, and J. B. Rudden. Development and Evaluation of Congress Street Expressway Pilot Detection System. In *Highway Research Record 21*, HRB, National Research Council, Washington, D.C., 1963, pp. 48–70.
4. A. D. May. *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, N.J., 1990.
5. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research*, Vol. 20A, Elmsford, N.Y., 1986, pp. 197–210.
6. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 53–60.
7. M. Koshi, M. Iwasaki, and I. Ohkura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory, 1981*, University of Toronto Press, Ontario, Canada, 1983, pp. 403–426.
8. P. Hsu. *Effect of Location on Freeway Traffic Flow-Concentration*. M.S. thesis. Department of Civil Engineering, San Diego State University, Calif., 1992.
9. J. H. Banks. Flow Processes at a Freeway Bottleneck. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 20–28.
10. J. H. Banks. Two-Capacity Phenomenon at Freeway Bottlenecks: A Basis for Ramp Metering? In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 83–90.
11. J. H. Banks. *Evaluation of the Two-Capacity Phenomenon as a Basis for Ramp Metering, Final Report*. Civil Engineering Report Series 9002, San Diego State University, Calif., 1990.
12. F. L. Hall and K. Agyemang-Duah. Freeway Capacity Drop and the Definition of Capacity. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 91–98.

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

Some Observations on Speed-Flow and Flow-Occupancy Relationships Under Congested Conditions

FRED L. HALL, ANNA PUSHKAR, AND YONG SHI

There is a need for a better understanding of freeway operations under congested conditions. Not only are major freeways in many cities congested several hours a day, but intelligent vehicle-highway system programs to be able to function effectively under all conditions, aggregate traffic behavior under all conditions needs to be accurately modeled. Three models for congested flow-occupancy data have been put forward in the past. Using data from the Highway 401 freeway management system in Toronto, evidence in support of any of those is sought, as it is for implied models of speed-flow relationships. The data suggest that all three models can be supported. Each may be appropriate under different conditions.

This paper attempts to begin to provide insight into the nature of congested operations on freeways, with particular reference to the flow-occupancy curve and the speed-flow curve. This may seem like a pointless task. Transportation agencies would prefer to avoid operating their facilities under such conditions. Why then should the 1985 *Highway Capacity Manual* (HCM) (*1*) or theoretical discussions pay attention to those conditions?

It is important for three reasons to know how traffic behaves within congestion. The first is the fact that in many jurisdictions, the major urban freeways are congested several hours a day, a situation that appears likely to get worse before it gets better. Consequently, transportation professionals need a better understanding of those conditions. The second is that the Committee on Highway Capacity and Quality of Service is currently attempting to gain that understanding and is in the process of revising the chapters of the HCM that describe these fundamental relationships. This is therefore a good time to contribute to that discussion. The third reason looks to the future, and to intelligent vehicle-highway systems (IVHS). For the promise provided by IVHS to come to fruition, there must be a solid understanding of aggregate traffic behavior under congested conditions as well as under free-flow conditions. For IVHS to provide realistic advice to motorists (or to make the vehicle route decisions for them, as has been suggested in more comprehensive IVHS schemes), the system will need to be able to predict travel conditions in the near future. An understanding of behavior within congestion is critical to that task.

The first section of this paper provides the background for the investigation that follows, by looking at what is known about congested operations and what has previously been suggested about how best to study them. The second section describes the freeway system from which the data for the paper have been drawn and the methods used to select and reduce the data. The third and fourth sections contain the substance of the investigation, in the form of discussion of a number of figures displaying relevant data and different models derived from them, first for flow-occupancy curves, then for speed-flow curves. The final section provides the conclusions that can be drawn from the study.

Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada.

BACKGROUND

To judge by the 1985 HCM, which is the most recent representation of the consensus view of North American highway professionals, there is not currently a sound understanding of freeway operations within congestion. The figures in the 1985 HCM showing speed-flow relationships provide only a dotted representation for the congested part of the curve, suggesting some uncertainty about its location (e.g., HCM Figures 3-4 or 3-6). The density-flow graphs provide an even stronger indication of uncertainty, showing only the first quarter or so of the congested portion of the curve beyond capacity (HCM Figures 3-3 and 3-6). This hesitation about defining the congested portion of the curve is continued in the revised Chapter 7 of the HCM, recently published (*2*), which shows only the uncongested portion of the speed-flow relationship and makes no effort to describe the congested portion (for multilane roadways).

With regard to the freeway material, the reluctance to provide a representation for congested operations is entirely understandable in light of the paucity of information available in the literature. Going back as far as Greenshields' (*3*) seminal work (which admittedly was not of freeway data), there is only one data point within the congested part of the curve. The best data set for dealing with the congested part of the curve is probably that used by Drake et al. (*4*) to compare a number of models of traffic relationships. For their analysis, they had between 50 and 60 observations within the congested area (depending on how that is defined), in which each observation represented 1 min of data for the middle lane of a 3-lane expressway. The other study with a considerable amount of congested traffic data was that conducted by Ceder (*5,6*). In both of these analyses, however, parameters related to the fitting of relationshps to the congested data were determined in part by the uncongested data. The most recent analyses,

by Hall et al. (*7*), suggest that the uncongested data are misrepresented by those models; hence it is very likely that Ceder's models distort the congested part of the curves. The 1992 paper by Hall et al., although making a start at defining the shape of the two curves, could define closely only the uncongested and queue discharge portions of the curves. Because of a lack of data, the authors had to rely on abstract logic rather than data to specify the congested part of the speed-flow curve; they could not decide between two placements of the congested part of the flow-occupancy curve on either logical or empirical grounds.

An additional reason for the lack of willingness to represent the congested portion of the curves appears in one of the earlier studies. Greenberg, 1959, noted (*8*, p.84) that in the queue upstream of a bottleneck section, flow is controlled by the bottleneck. [May (*9*, p.288) illustrates this effect.] If flow in the queue is controlled by somewhere else, one might decide that there is little sense in studying congested operations with reference only to what happens within the queue. Certainly the effect of the downstream bottleneck needs to be considered in any effort to understand congested operations. Because it may not be clear where that bottleneck is, a recursive approach may help, in which operations at one section are identified as a function of conditions there and at the next downstream section (which in turn may require looking farther downstream). Nevertheless, the presence of this effect should only serve as a warning for analyses of congested flow; it should not preclude such study entirely.

Wattleworth (*10*) discussed the nature of the flow-density curve that can arise in a bottleneck when its flow is governed by an upstream queue and ramp that together do not supply enough traffic to the bottleneck. He concluded that "it may not be possible to obtain empirically a true volume-density curve for a bottleneck since part of the observed curve may merely be reflecting the influence of conditions upstream of the bottleneck." (*10*,p.20) This observation complements that just noted by May, in the sense that it is likely that no single location can provide the full range of conditions necessary to define the curves and that any location's data can be affected by upstream or downstream conditions. Hence it is particularly important to be aware of surrounding conditions during the time data are being acquired.

To assist the discussion of the congested section of the curves, it is helpful to have a model for the other parts of them. The curves identified by Hall et al. (*7*) will be accepted for subsequent discussion: a linear uncongested section that falls off slightly at higher flows, and a queue discharge segment at constant volume. In the following discussion, the three parts of the curves will be referred to as uncongested, queue discharge, and congested operations. Although queue discharge may be regarded as a form of congestion, in that drivers are traveling at less than their desired speed, for convenience only operations within a queue will be called congested.

For the flow-occupancy curve, Banks has suggested that within-queue flows can be represented by a straight line (*11*). If this is accepted, the question that remains is how this section relates to the other two—in other words, how and where the within-queue section of the curve is to be placed. Three suggestions appear in the previous literature, two of which were included in the flow-occupancy figure by Hall et al. All three have been included in Figure 1. For convenience in later

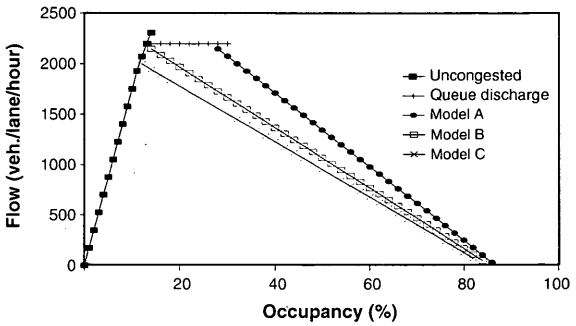


FIGURE 1 Proposed models for congested flow-occupancy.

discussion, they have been termed Models A through C. Model A arises from work in Chicago and joins the congested segment to the right-hand end of the queue-discharge flow section (*12,13*). Model Bel arises from both Ontario and California data and has the congested segment at least aiming toward; if not joining, the uncongested section at the peak of prequeue flows (*11,14*). Model C arises from Japanese data and has been called the reverse lambda model (*15*).

There is a necessary connection between the decision made as to the placement of the uncongested segment on a flow-occupancy graph and that on the speed-flow graph. Not only must they reflect identical flow conditions within the queue, but the speeds and occupancies at any given flow are inversely related: the lower the speed, the higher the occupancy. If occupancy were a constant multiple of density, as Athol (*12*) suggested, then one could specify the nature of the inverse relationship precisely, starting from the fundamental identity,

$$\text{flow} = \text{speed} * \text{density} \quad (1)$$

to calculate that

$$\text{flow} = \text{speed} * \text{occupancy} * k \quad (2)$$

where k is a constant of proportionality. However, Hall and Persaud (*16*) showed that the linear relationship between occupancy and density holds only under restrictive assumptions, namely, either uniform vehicle lengths or uniform vehicle speeds. Under congested conditions, vehicle speeds are not uniform across vehicles. Vehicle lengths are not uniform even within the lane closest to the median of the roadway, although use of data from only that lane would minimize the variation. However, for data from the full roadway, vehicle lengths vary widely. Hence the preceding equation cannot be used to determine one of the speed-flow and flow-occupancy relationships from the other. Nevertheless, as a starting point for this analysis, Figure 2 shows the three possibilities for the congested portion of the speed-flow curve that correspond with the flow-occupancy curves in Figure 1 based on Equation 2.

Given this background to the problem, the task can now be more clearly specified: what, if any, are the conditions under which each model is correct? That they might all be correct appears likely when one considers the fact that the queue discharge portion of the curve is due to upstream conditions and the congested portion is due to downstream conditions. Each station on the freeway is related differently to

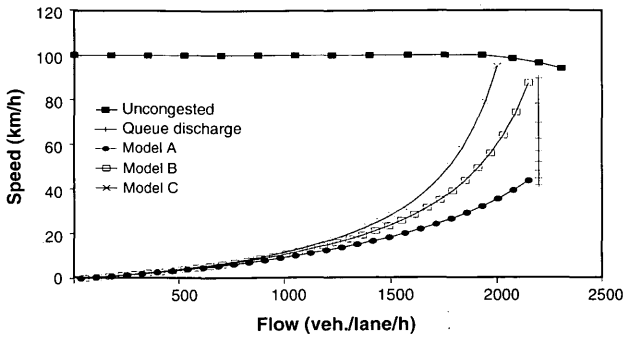


FIGURE 2 Implied models for congested speed-flow relationships.

the upstream and downstream bottlenecks that create these traffic patterns. Hence the queue discharge and congested portions of the curve may not be related in exactly the same way for different stations. The data patterns may then depend on how congestion is arrived at, in the sense that the congested portion of a curve will look different for a location that has been experiencing queue discharge flow, as compared with one that moves into congestion from normal uncongested operations. The investigation needs to take into account two considerations. The first, and potentially more important, is the governing downstream bottleneck. The second is the fact that it is unlikely that the full range of the congested parts of the curves can be specified with data from a single location. Hence it may be helpful to combine data from several locations.

DESCRIPTION OF AVAILABLE DATA

Although it will be necessary to investigate this problem with data from a number of locations, such an investigation must start somewhere. Relevant data were available from a freeway traffic management system (FTMS) that was put into operation January 30, 1991, on Highway 401 in Toronto. Although there is more to the system than inductive loop detectors, those are the only aspects of the system used for this analysis, so they are the only parts described.

Spacing of detector stations on the 401 system averages roughly 500 m, with some as close as 380 m and others as much as 770 m apart. There are double-loop detectors at every third station, where speeds are calculated for each vehicle over the 4.5 m distance between the loops, and average speeds are reported every 20 sec, along with the volume counts and occupancies. For single-loop stations, speeds are estimated on the basis of vehicle lengths calculated at a nearby station, based on the speeds there. These speed estimates are not accurate enough for our purposes. Hence there are potentially three times as many stations available for the flow-occupancy discussion as for the speed-flow graphs.

The part of the 401 highway on which the FTMS has been installed is an express-collector system. Within the 15.7 km of the section on which the FTMS is operating, there are five transfers from the express lanes to the collectors, and three transfer opportunities from collector to express, in each direction. Major bottlenecks occur at some of these, and at a few particular entrance ramps to the collector system. Because the focus of this study was congested traffic, locations upstream of these bottlenecks were of primary interest. Despite the size of the 401 system, and the perception that there is considerable congestion in the system, in fact only a limited number of stations could be used for the analysis. The congestion is concentrated at a relatively small number of stations, and not all of those provided speed data. This study focuses on the section of the 401 east of Weston Road and west of Yonge Street (Figure 3), as it contains several potential bottleneck areas, and most of the detectors were operating properly in the area.

On the 401, the data are available for 20-sec intervals. Although this level of detail is useful for identifying traffic incidents quickly, plots of such data contain considerable random variation that sometimes obscures the underlying relationship. Figure 4 shows an example of the 20-sec data obtained from one of the westbound stations. Although a pattern is apparent, considerable variation is clear within the congested data. To see which presented the clearest picture of the relationship, intervals of 1, 2, 5, 10, and 15 min were examined. [The treatment of missing data within the averaging program was to assume it had the same value as the rest of the interval if up to one-third of the intervals (e.g., five of fifteen 20-sec periods in a 5-min interval) were missing or invalid, and to treat the whole interval as missing if more than one-third of the data was missing.] The 5-min interval was chosen as the best one for the present purpose as it minimized the random variation yet retained some indication of the data patterns. Figure 5 shows the 5-min data for the same station and time period as Figure 4. The transitions to and from congestion occur at approximately the same flow rates in each figure, and the congested data center on 40

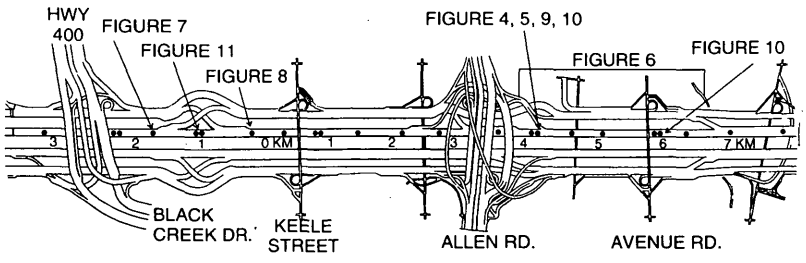


FIGURE 3 Schematic diagram of portion of Highway 401 FTMS (COMPASS) showing westbound detector locations, and locations for data in subsequent figures (modified from figure produced by Ministry of Transportation, Ontario).

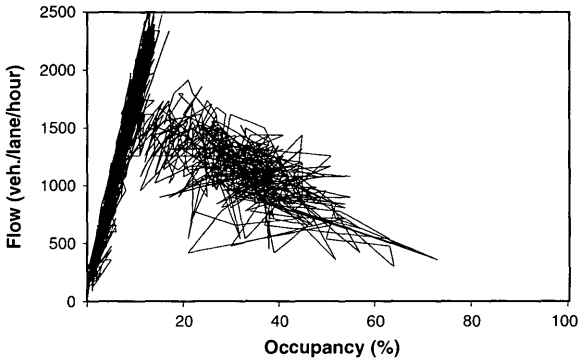


FIGURE 4 Twenty-second data for two peak periods (0500–0900 and 1600–1900).

percent occupancy and a flow of 1,000 vehicles per hour (vph) in both figures. However, the range and scatter of the congested data have been greatly reduced by the averaging.

For the 401 system, the station data are stored in separate files from the loops data, which would provide information on the individual lanes. Although it might be of interest to study the behavior in each separate lane across a station, the analyses here have focused on the station data. Within congestion it appears that all lanes operate similarly, and the primary question for applications is how the roadway as a whole operates.

FLOW-OCCUPANCY RELATIONSHIPS

Before turning to consideration of the shape and position of the speed-flow curve, it is useful to examine the flow-occupancy curves, primarily because more flow-occupancy data are available from the Highway 401 system. Any conclusions from the study of these graphs will help to determine the speed-flow congested curve, on the basis of the approximate relationship specified in Equation 2 and reflected in the deviation of Figure 2 from Figure 1.

The 401 system has some lengthy congested sections in which queues can back up for several miles. The first approach was to take data from a number of stations within one queue, to learn if they provided a similar picture for the congested part of the curve. This was done for five consecutive stations on the westbound express lanes, with data for March 6, 1991. The five are the seventh through the eleventh east of the control center. One of the five is the station shown in Figures 4 and 5. At this station, the data went from normal uncongested operations to congestion at flows of about 1,500 vph, over perhaps 15 min, then stayed within congestion for more than an hour before returning, at lower flows, to uncongested operations. This pattern is fairly typical of the five stations, in both the flows at which congestion set in and the duration of congestion.

The combined data for the five stations appear in Figure 6. The difference between movement into and out of congestion is lost in the general scatter of the combined data. An attempt was made to fit a third-order polynomial on occupancy to the congested data, using a stepwise regression analysis, but only one of the three variables could be made to enter. The first-order correlation coefficients are very similar for all three variables in the polynomial: the square of occupancy produces an R^2 of .748, occupancy alone an R^2 of .737, and the cube of occupancy an R^2 of .722. However, on the basis of the jam occupancy that results from putting each of these into a separate equation, the linear function is most sensible, as it is the only one to produce an x-intercept of more than 70 percent. It might be expected that the jam occupancy should be 100 percent, but data such as shown in Figure 4 suggest that the x-intercept lies between 80 and 90 percent. Of course, few if any of these data come from completely stopped traffic, but even when traffic is at a standstill, space remains between vehicles. The equation selected to represent these data is

$$\text{flow} = 1919 - 21.8 * \text{occupancy} \quad (3)$$

This equation, as is obvious from Figure 6, is not consistent with either Models A or B of Figure 1. The line would join the uncongested line in the vicinity of a flow of 1,700, as in Model C, the reverse lambda suggested by Koshi (*15*).

However, these data are not the only pattern observed. Figure 7 shows data from a station that was experiencing what appears to be queue discharge flow, at a flow rate in the vicinity of 1,750 vph and occupancies of 30 percent, when a queue from a downstream incident reached the location. The data seem to follow Model A, the Chicago model, quite well.

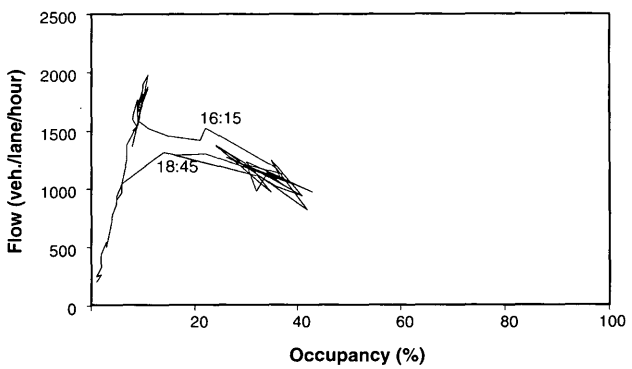


FIGURE 5 Five-minute aggregate data for two peak periods.

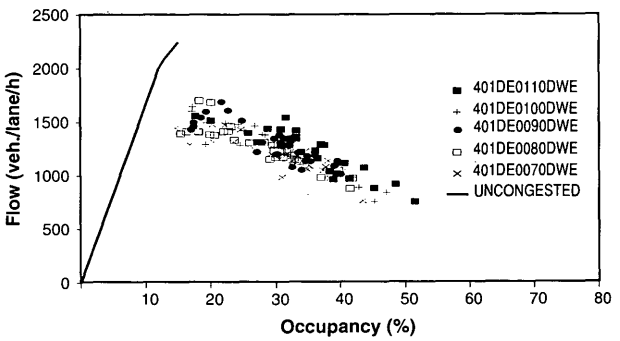


FIGURE 6 Combined congested data for five westbound express stations.

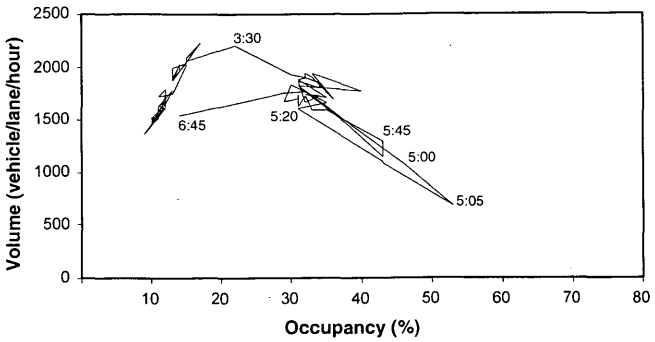


FIGURE 7 Flow-occupancy pattern where queue discharge precedes operations within a queue.

In addition, if the congested segment of the curve is extrapolated back to the uncongested segment, it would appear to be consistent with Model B. Regardless of whether Model A or B (or both) is supported by Figure 7, it is clear that these data occur at higher flow rates than the data in Figure 6, for the same occupancy. Hence at least two different patterns of congested flow-occupancy behavior can occur, depending on the circumstances.

The picture of flow-occupancy behavior can be even more complicated. Figure 8 shows the data pattern for a location that was experiencing normal queue behavior (as represented by the large cluster of points at flows of 600 to 800, and occupancies of 40 percent or so) at the time that a jackknifed tractor-trailer downstream of this location caused a severe capacity reduction. The series of points from 12:10 to 13:15 do not fall on any of the curves that have been identified. The relationship between flow and occupancy within congestion may be much more complex than one might hope.

SPEED-FLOW RELATIONSHIPS

Because of the absence of speed traps at two-thirds of the stations on the 401, only two of the five stations used in Figure 6 provide good speed data. Consequently, two days from each station have been used to provide a comparable estimate of the shape of the congested portion of the speed-flow curve. As an example of the full context for such data, Figure 9

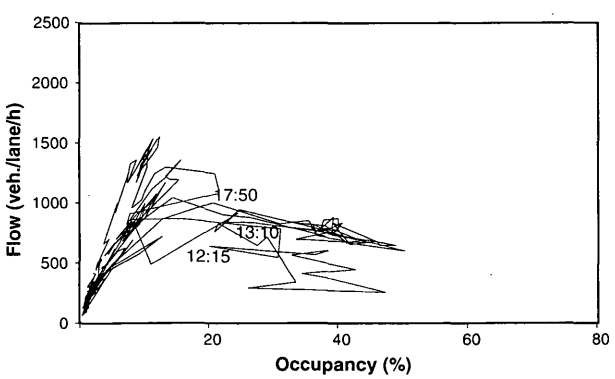


FIGURE 8 Flow-occupancy pattern with queueing resulting from a severe capacity restriction.

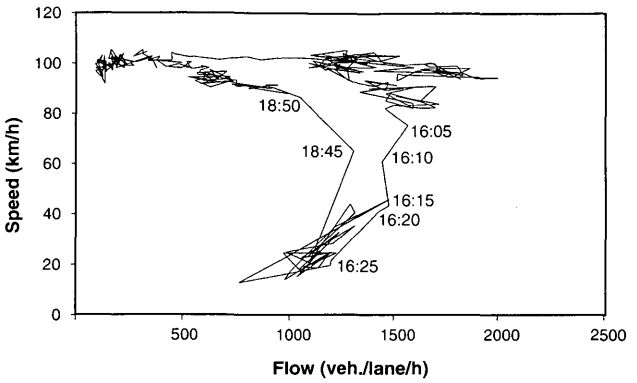


FIGURE 9 Speed-flow data corresponding to flow-occupancy data of Figure 5.

shows the 5-min averages of the speed-flow data corresponding to the flow-occupancy data of Figure 5.

The four sets of congested data are combined in Figure 10. As with the flow-occupancy data, an attempt was made to fit a third-order polynomial, in this case forced through the origin, to the data. For the speed-flow data, two terms were needed, the linear term and either the quadratic or cubic term. With the former, the R^2 was .936; with the latter, .937. The resulting equations give, respectively, speeds of 72 and 81 km/hr at a flow of 2,000 vph, and speeds of 87 and 103 km/hr at flows of 2,200 vph. Since the 103 km/hr speed is certainly too high, the quadratic equation will be accepted for the moment. The equation is therefore

$$\text{speed} = 0.003 * \text{flow} + 0.0000166 * \text{flow}^2 \quad (4)$$

This model would seem to be most nearly consistent with Model B of Figure 2.

It is possible that the top four points in the graph are really transitional ones and not truly congested data. Both equations changed when these points were omitted from the analysis, with the coefficient on the linear term becoming larger and that on the quadratic or cubic term being reduced by about a third. The R^2 for both equations became .939. Speeds at 2,000 vph were only 62 and 68 km/hr respectively. Such a model might seem to be consistent with Model A of Figure

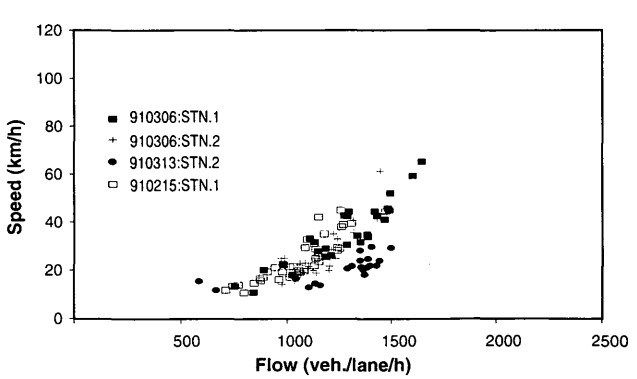


FIGURE 10 Combined congested data from two stations and 3 days.

2, but other data suggest that those speeds are not low enough to reflect Model A.

Figure 11 shows the pattern for data immediately upstream of the major transfer from collector lanes to express lanes. When congestion sets in at this station, speeds drop abruptly, to the vicinity of 30 km/hr, for flows ranging from 1,500 to 2,000 vph/lane. These data are clearly not consistent with the data in Figure 10, yet the pattern in Figure 11 is reproduced regularly. These data are consistent with Model A, suggesting that exclusion of the top four points in Figure 10 would not be appropriate, and that the best representation of the combined speed-flow data in Figure 10 is as in Equation 4.

CONCLUSIONS

Each of the models identified in Figures 1 and 2 seems to have data to support it, based on the examination of some data from Highway 401 in Toronto. It seems most likely that each could apply at different locations along the freeway or under different conditions. For example, Model A in the flow-occupancy case seemed to apply when a queue extended upstream to a station that had previously been operating in queue discharge mode. The speed-flow version of Model A was found in the first station upstream of a major transfer lane. Flow-occupancy Model C was found within an extended queue; the speed-flow data from within that queue seemed most likely to match Model B.

One conclusion of the analysis, then, is that the simple approximation provided by Equation 2 (relying on occupancy being a constant multiple of density) is not a close enough approximation to be of help in specifying the nature of relationships within congested operations. If it were, then the same model would apply to both the flow-occupancy and the speed-flow data from a particular location. A theoretical derivation explaining the problems with that approximation appeared in Hall and Persaud (*16*), as mentioned earlier. These results emphasize the problems of continuing to use the fundamental equation within congested operations.

Some data were also found, in the queue upstream of a major accident, that did not match any of the models. Contrary to the simple smooth curves that might have been expected, such as those that provide reasonable representations of uncongested operations, the situation within congested operations seems quite complex. Further investigation of congested operations is warranted. It may also help to unravel the complexity if some analyses are conducted lane by lane.

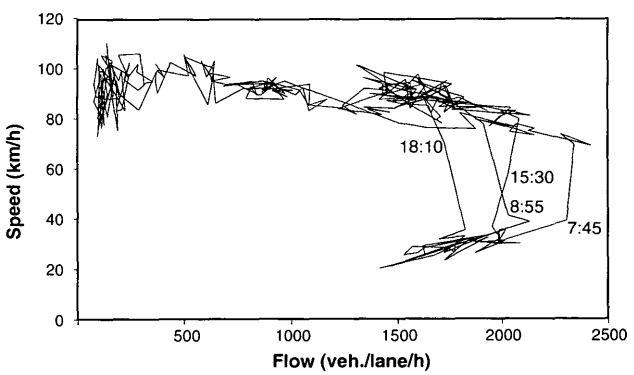


FIGURE 11 Speed-flow data for a station just upstream of a major ramp.

One important question that should be addressed for further analyses is what the appropriate averaging interval should be. This paper has relied primarily on 5-min data, in part for reasons explained above, but also in part for consistency with the types of analyses that underlie the current HCM. Certainly much information is lost by moving from the detail available at shorter intervals (such as in Figure 4) to a 5-min average. If regression analyses are to be used to try to develop equations from the data, then it may be better to retain the data for the shortest available interval. If the results of analyses such as these are to be used in IVHS installations, shorter intervals would also make sense. On the other hand, few facilities collect data at the 20-sec intervals used on the 401. There is an argument for some standardization of analyses.

ACKNOWLEDGMENTS

The work described here has been supported by the Natural Sciences and Engineering Research Council of Canada, through both a research grant to Fred Hall and a summer research assistantship to Anna Pushkar. The assistance of the Ministry of Transportation of Ontario, through the Freeway Traffic Management Section, in making available the data is greatly appreciated.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. *Special Report 209S: Supplement to Highway Capacity Manual: Revised Chapter 3, Chapter 7, Table of Contents, and Index.* TRB, National Research Council, Washington, D.C., 1992.
3. B. D. Greenshields. A Study of Traffic Capacity. *HRB Proc.*, Vol. 14, 1935, pp. 448–477.
4. J. S. Drake, J. L. Schofer, and A. D. May Jr. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, 53–87.
5. A. Ceder. *Investigation of Two-Regime Traffic Flow Models at the Micro- and Macroscopic Levels.* Ph.D. thesis. University of California, Berkeley, 1975.
6. A. Ceder and A. D. May. Further Evaluation of Single- and Two-Regime Traffic Flow Models. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 1–15.
7. F. L. Hall, V. F. Hurdle, and J. M. Banks. Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992.
8. H. Greenberg. An Analysis of Traffic Flow. *Operations Research*, Vol. 7, 1959, pp. 79–85.
9. A. D. May. *Traffic Flow Fundamentals.* Prentice-Hall, Englewood Cliffs, N.J., 1990.
10. J. A. Wattleworth. Some Aspects of Macroscopic Freeway Traffic Flow Theory. *Traffic Engineering*, Nov. 1963, pp. 315–320.
11. J. H. Banks. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 53–60.
12. P. Athol. Interdependence of Certain Operational Characteris-

tics Within a Moving Traffic Stream. In *Highway Research Record 72*, HRB, National Research Council, Washington, D.C., 1965, pp. 58–87.
13. A. D. May, Jr., P. Athol, W. Parker, and J. B. Rudden. Development and Evaluation of Congress Street Expressway Pilot Detection System. In *Highway Research Record 21*, HRB, National Research Council, Washington, D.C., 1963, pp. 48–70.
14. F. L. Hall, B. L. Allen, and M. A. Gunter. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research A*, Vol. 20A, 1986, pp. 197–210.
15. M. Koshi, M. Iwasaki, and I. Ohkura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory*, 1983, pp. 403–426.
16. F. L. Hall and B. N. Persaud. An Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.

The ideas expressed here represent the opinions of the authors only.

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

Study of Freeway Bottlenecks in Texas

JOHN RINGERT AND THOMAS URBANIK II

Observations of flow rates much higher than 2,000 passenger cars per hour per lane and the recent revision of the multilane highway chapter in the *Highway Capacity Manual* have led to questioning the current value of freeway capacity and the speed-flow relationship. An analysis of free-flow and queue discharge flow rates at three freeway bottlenecks in Texas found less variability in queue discharge flow rates than in free-flow flow rates. Average free-flow flow rates ranged from 2,096 to 2,210 vehicles per hour per lane (vphpl) across all lanes, whereas queue discharge flow rates averaged approximately 2,175 vphpl for the study sites. In addition, higher flows did not occur in free-flow conditions in all cases. As a result of lane interaction, some lanes are prematurely transitioned into queue discharge without reaching high flow rates in free-flow conditions.

During the past decade, much attention has been given to freeway capacity and the relationship of speed, flow, and density for freeway bottlenecks. Freeway capacity plays a critical role in the planning, design, and operation of freeways in general and urban freeways in particular. As a result of steadily increasing congestion in urban areas throughout the United States, traffic engineers and transportation planners have identified problems associated with the freeway capacity numbers given in the 1985 *Highway Capacity Manual* (HCM) and the corresponding speed-flow relationship (*1*). Many studies have concluded that flows measured on freeways are frequently greater than the values given as capacity. A study by Hurdle and Datta (*2*) in 1983 concluded that the value of 2,000 passenger cars per hour per lane (pcphpl) was still a good estimate of capacity. A more recent study by Agyemang-Duah and Hall (*3*) concluded that the capacity flow rate was approximately 2,300 pcphpl. A study by Urbanik et al. (*4*) found peak 15-min flow rates between 2,100 and 2,300 vehicles per hour per lane (vphpl) at four sites in Texas.

In addition to the measurement of high flow rates, many studies have concluded that a reduction in capacity occurs once a queue forms. Hall and Agyemang-Duah concluded that the flow rate dropped from 2,300 to 2,200 pcphpl in queue discharge conditions (*3,5*). If two capacities exist, a possible problem exists with the use of the peak 15-min flow because it would give a flow that may not be sustainable at higher demand levels where queue discharge occurs.

Although many studies on freeway capacity have been done, a good understanding of the characteristics of flow during free-flow and queue discharge has not been achieved. The characteristics of flow at freeway bottlenecks, whether a reduction in flow exists once a queue forms, the shape of the speed-flow relationship, and recommendations regarding capacity are examined in this paper. Because the definition of capacity plays a significant role in the determination of the value to use as capacity, the maximum sustainable flow will be the focus of this paper. Maximum sustainable flow is the maximum flow rate that can be maintained for an indefinite period with sufficient demand. For capacity to be useful, it is important that it be sustainable and repeatedly achievable.

J. Ringert, Kittelson & Associates, Inc., 1455 Response Road, Suite 120, Sacramento, Calif. 95815. T. Urbanik II, Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843.

STUDY SITES

Data were collected at three study sites. A schematic of each study site is shown in Figures 1, 2, and 3. The primary study site that was used to evaluate the operation at freeway bottlenecks was on US-290 in Houston, Texas. This study site was chosen because of the frequent occurrence of queues during the p.m. peak period.

The other two study sites were on I-410 in San Antonio, Texas, and at the merge of I-35 and US-67 in Dallas, Texas. These two study sites were chosen to validate the flow characteristics found at the US-290 site. These locations were chosen because of the occurrence of congestion during a.m. peak periods.

RESULTS OF ANALYSIS OF US-290

Data Collection

Data were collected for 15 days at the US-290 study site. The data were collected for 2 to 3 hr on each day using pairs of inductive loops in each traffic lane.

Analysis of Flows at US-290 Study Site

To be consistent with existing convention, 15-min peak flows were measured for each of the sample days. The average peak 15-min flow rates are given in Table 1. The flow rates in Table 1 represent the peak 15-min flow rates during the study period across all lanes for the 15 samples. Individual lanes can have higher peak 15-min flow rates occurring at different times.

Although truck traffic was present, no attempt was made to adjust for truck passenger car equivalents in the preliminary analysis. Truck percentages averaged 3.0 percent for all lanes combined. The highest percentages were in the outside lane (Lane 3), which averaged 5.3 percent trucks, whereas the inside lane (Lane 1) had the lowest average truck percentage, with 0.7 percent. These truck percentages reflect the truck traffic for the study periods.

Although peak 15-min flow rates are the existing convention detailed by the 1985 HCM, it is likely that they are not

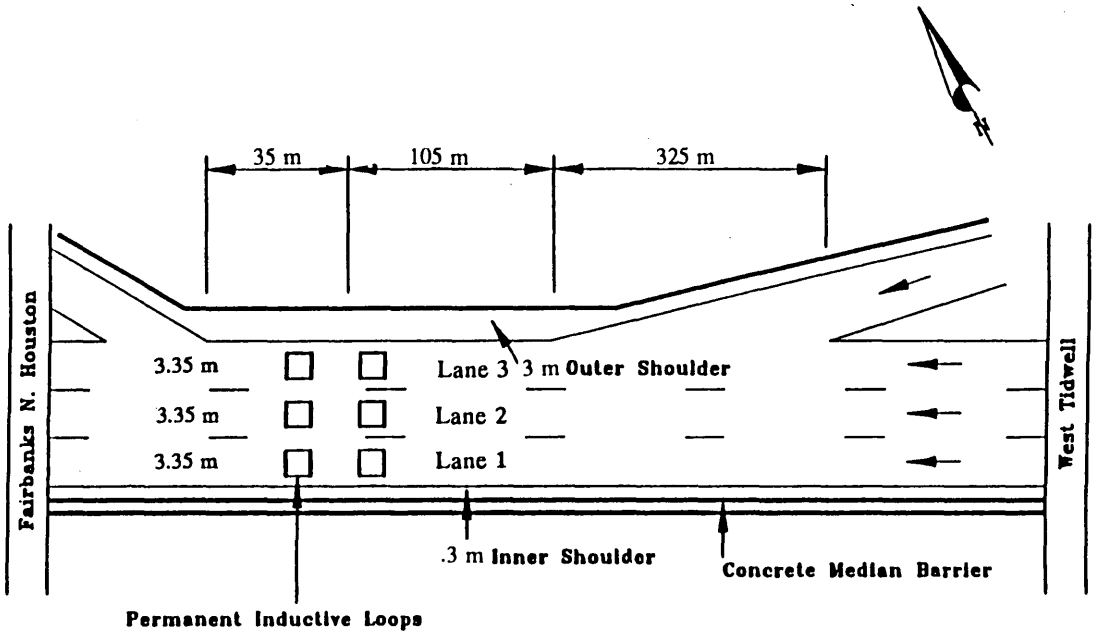


FIGURE 1 US-290 at Tidwell study site.

sustainable. The two-capacity hypothesis asserts that two separate capacities exist: one during free-flow conditions and one when demand exceeds capacity, creating queue discharge conditions. Therefore, distinguishing between regions is necessary to ensure that a measurement does not contain data from both regions at the same time.

To compare the characteristics of flow before and during the period when demand exceeds capacity, the point at which demand exceeds the service rate of the bottleneck was identified. Demand is considered to exceed the service rate when vehicle speed is controlled by the service rate of the bottleneck, thereby making vehicles wait to resume their desired speed. The excess demand generates queues and produces queue discharge flow. The point at which a rapid drop in 30-sec average speed occurs, which is shown conceptually in Figure 4, was determined to be the beginning of queue dis-

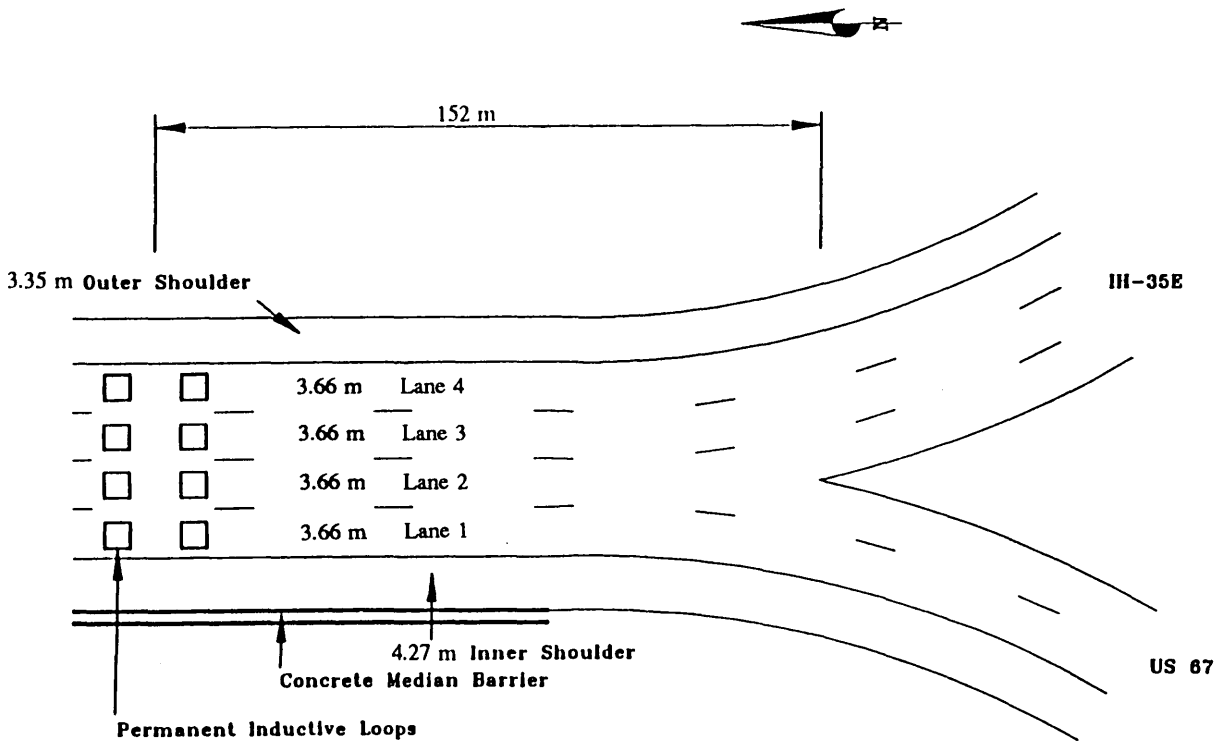


FIGURE 2 I-35/US-67 study site.

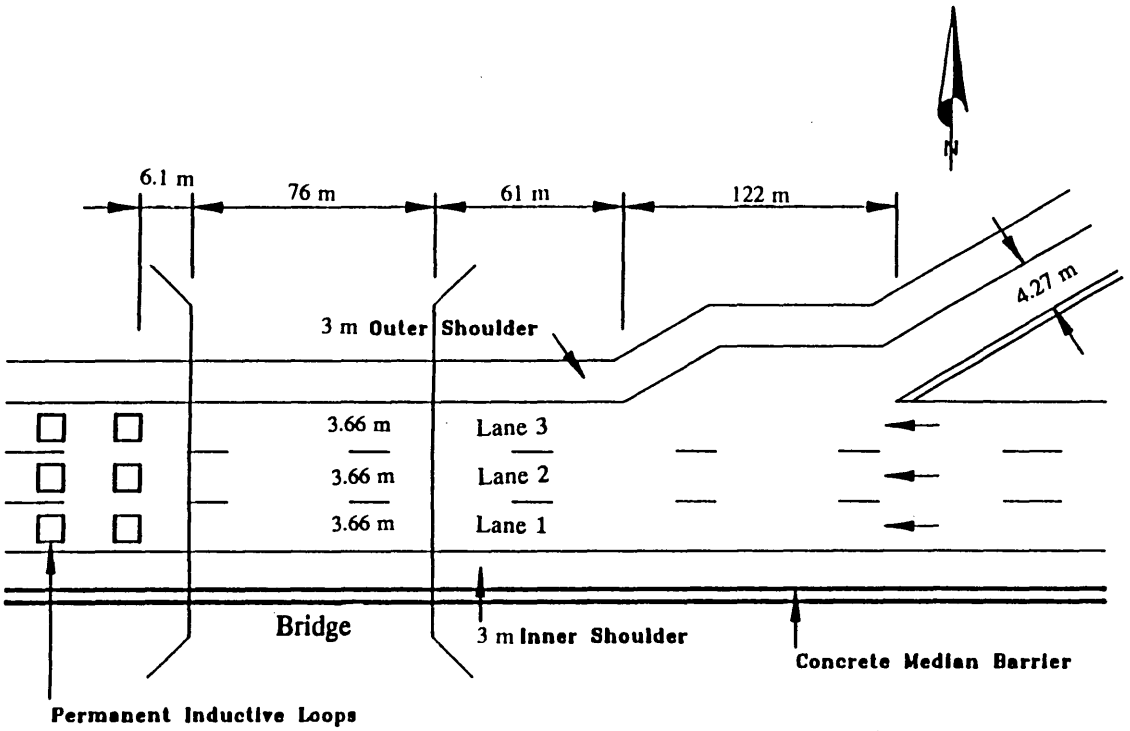


FIGURE 3 I-410 study site.

charge. Vehicles are forced to a lower rate of speed than previously desired under prevailing conditions. The following three time periods were used in the analysis:

- The 5 min immediately before the speed drop (free flow);
- The 5 min immediately after the speed drop (queue discharge); and
- The entire period after speed drop until end of queue discharge, formation of a downstream queue, or the end of data collection (queue discharge).

These time periods in relation to a typical speed profile are illustrated in Figure 4. The flow rates were calculated from the average headways during each interval. In most cases the

TABLE 1 Peak 15-min Flow Rates Across All Lanes at US-290 at Tidwell

Highway	Observation	Peak 15-Minute Flow Rate (vphpl)			
		Lane 1	Lane 2	Lane 3	Average
U.S. 290	1	2336	2256	2348	2313
	2	2320	2200	2140	2220
	3	2300	2244	2456	2333
	4	2380	2240	2060	2227
	5	2288	2172	2180	2213
	6	2368	2204	2312	2295
	7	2260	2224	2120	2201
	8	2228	2196	2228	2217
	9	2220	2268	2132	2207
	10	2384	2256	2188	2276
	11	2496	2244	2172	2304
	12	2252	2220	2348	2273
	13	2408	2344	2156	2303
	14	2248	2356	2028	2211
	15	2312	2240	2068	2207
	Average	2320	2244	2196	2253

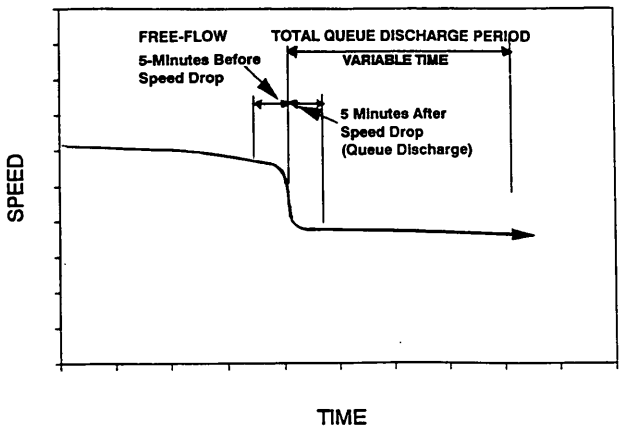


FIGURE 4 Time intervals used for analysis, typical speed profile.

speed drop occurred at the same time in all lanes, but for consistency between lanes the location of the speed drop for each sample was determined on the basis of highest flow lane (Lane 1 at the US-290 site).

Statistical Analysis of Speed Drop

A statistical analysis of the speed drop was performed in two steps. To determine whether a significant change in flow characteristics occurred after the speed drop, an *F*-ratio was calculated to compare the variances of headways during the 5-min intervals before and after the speed drop. A *t*-test was then performed to determine whether significant changes in mean headways occur after the speed drop. In using both the *F*-ratio and the *t*-test, it was assumed that the sample mean headways were independent and the headways within them have a normal distribution. These assumptions are considered valid because of the large sample of headways and the central limit theorem, which states that as sample size increases, the distribution of sample means approaches normality.

The *F*-ratio indicated that, for the most part, the two time periods have unequal variances. The variance before the speed drop was significantly higher than the variance after the speed drop for all 15 samples for Lane 1, for 9 samples for Lane 2, and for 10 samples for Lane 3. Because the variances were significantly different in several cases, the mean headways before and during queue discharge were compared using a *t*-test approximation for significance of difference between two means with different variances. The results of this *t*-test indicated that only three samples in Lane 1, two samples in Lane 2, and six samples in Lane 3 had statistically different mean headways before the speed drop compared with after the speed drop. Therefore, most samples did not experience a statistically significant change in mean flow after the speed drop. Because of the high variance in headways and corresponding flows, a large change is required to be statistically significant.

Average Flows for all Samples Combined

To evaluate the characteristics of flows between days, statistics were calculated for the mean flow rates for the 15 sample days. Individual flow rates for each sample were calculated from mean headways. The statistics for the sample flows are given in Table 2.

Table 2 gives the means, standard deviations, and a 95 percent confidence interval on the means for each time interval for the 15 days on which data were collected. The confidence interval assumes that the distribution of sample means follows a normal distribution, which should be applicable because a relatively large sample size was used.

The average flows show some interesting trends. Whereas the mean average flow rate across all lanes does not change much for the 5-min period before the speed drop compared with the 5-min period after the speed drop during queue discharge, the distribution of traffic across the lanes does change. In Lane 1 the mean traffic volume substantially increased in queue discharge. The opposite occurred in Lane 3, where the

TABLE 2 Statistics for Daily Averages for US-290 at Tidwell

Time Period	Statistic	Lane 1	Lane 2	Lane 3	Average
5-min Before Speed Drop (Free-Flow)	Average Flow (vphpl)	2076	2002	2210	2096
	Std Dev.	136	159	187	132
	95% CI (+/-)	75	88	104	65
5-min After Speed Drop	Average Flow (vphpl)	2266	2035	1989	2097
	Std. Dev.	145	134	193	115
	95% CI (+/-)	80	74	107	64
Total Queue Discharge	Average Flow (vphpl)	2246	2161	2090	2166
	Std. Dev.	81	68	101	67
	95% CI (+/-)	45	38	56	37

average volume decreased after the speed drop in queue discharge. This could be explained by the fact that in free-flow conditions Lane 1 has not reached its maximum flow and operates at very high speeds while Lane 3 is already becoming congested. When Lane 3 becomes congested and slows down even further, traffic merges into Lanes 2 and 1 and subsequently drops the speed and transitions the flow into queue discharge conditions. Therefore, if the maximum flow occurs in free-flow conditions, it would not be seen in Lane 1 because Lane 1 transitioned from free flow with a flow rate below maximum flow directly to queue discharge.

The analysis of means points out another problem with using free-flow conditions to obtain a maximum sustainable flow rate. Although a maximum flow rate in free-flow conditions exists, in certain lanes it may never be reached, and therefore it would be impossible to achieve. For example, suppose Lane 1 had a maximum flow of 2,400 vehicles per hour (vph), but because of the turbulence created by Lanes 2 and 3, the flow rate in Lane 1 transitioned directly from 2,076 to 2,266 vph in queue discharge as the data suggest. The flow rate of 2,400 vph would never be reached and therefore could not be considered the maximum flow rate.

Although it appears that for individual samples the variance decreased in queue discharge, the variance among samples (standard deviation squared) remained high in the 5-min period after the speed drop. One explanation for this occurrence is the way in which the boundary of the free-flow and queue discharge conditions was chosen. As shown in Figure 4, the boundary was determined on the basis of the speed drop. As can be seen in Figure 5 for Lane 1, a rapid speed drop into what appears to be queue discharge conditions does not occur in all samples. To investigate the effects of samples that did not appear to drop directly into queue discharge, the samples that appeared to drop directly into queue discharge were evaluated. The results of the evaluation of samples that had rapid speed drops into queue discharge indicate that the variance (and standard deviation) between samples decreases in the 5-min interval after the speed drop.

Speed-Flow Relationship

Figure 6 shows the speed-flow plots for all samples combined for each lane. Five-min averages were used for the plots. The lines represent the estimated shape of the speed-flow relationship on the basis of the data. The ends of the lines are equal to the average total queue discharge flow. Note that some data points include downstream congestion, also shown in Figure 6.

The hypothesized speed-flow relationship resulting from the analysis is plotted in Figure 7, showing that the operational characteristics of each individual site determine the capacity. Because of lane interaction, when Lane 3 transitions into queue discharge the other lanes quickly follow without reaching their theoretical maximum flow. As mentioned previously, this transition may be nearly instantaneous in some cases. The dashed lines show the hypothesized shape of the end section of the speed-flow relationships in Lanes 1 and 2, assuming Lane 3 does not break down and prematurely transition Lanes 1 and 2 into queue discharge.

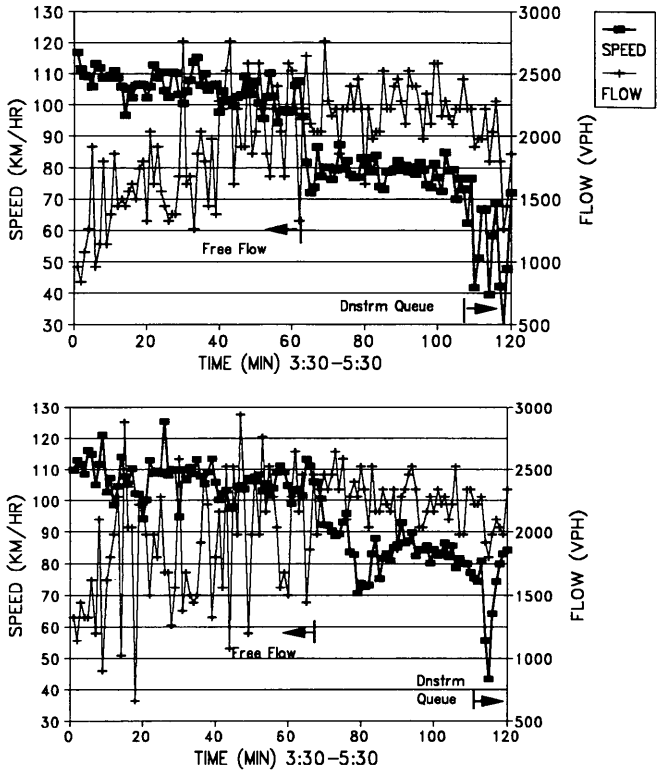


FIGURE 5 Types of speed drops during transition to queue discharge, 1 min, Lane 1: *top*, rapid transition (sample 2); *bottom*, unstable transition (sample 6).

Results of US-290 Analysis

From the analysis of the US-290 site the following observations can be made:

- The variance in flow decreases after the speed drop.
- The flow rate in Lane 3 decreased after the speed drop while flow rates in Lanes 1 and 2 increased, indicating that when Lane 3 transitions into queue discharge, Lanes 1 and 2 also transition to queue discharge.
- The transition to queue discharge can occur nearly instantaneously or over a longer period of time.
- Because of the variability and instability of flow during free-flow conditions and in the direct vicinity of the speed drop, use of these regions as maximum sustainable flow and capacity might not be practical.

COMPARISON WITH OTHER BOTTLENECK SITES

Analysis of Flows at Validation Sites

To verify the results of the analysis for the US-290 site, data from the I-410 and I-35/US-67 sites were collected. Data were collected for 4 days at the I-410 site and 3 days at the I-35/US-67 site. Because one sample at the I-410 site was taken during a rainy morning, only three samples were used in the analysis. The peak 15-min flow rates for each of these sites are given in Table 3.

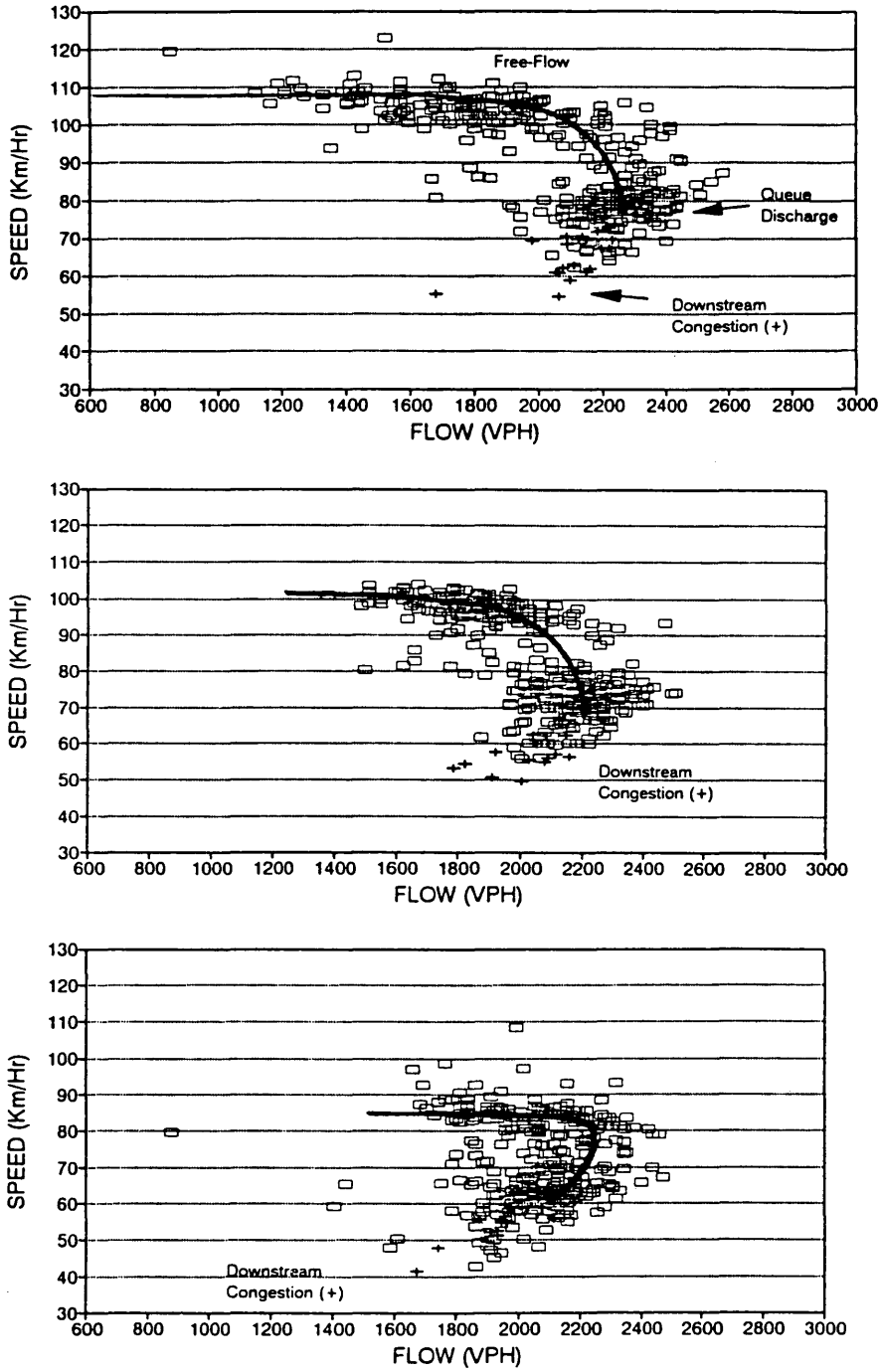


FIGURE 6 Speed-flow plots for US-290 study site, 5-min averages: *top*, Lane 1; *middle*, Lane 2; *bottom*, Lane 3.

The peak 15-min flow rates show some apparent differences between the two study sites as well as the US-290 study site. A comparison of flow rates for individual lanes indicated that the median lane (Lane 1) is the highest flow lane for all the sites and is predominant for the I-410 and the I-35/US-67 sites. These two sites have very large peak 15-min flows of 2,496 and 2,492 vph. The lowest flows tend to be in the outside lanes, which are merge lanes for the US-290 and I-410 sites but not for the I-35/US-67 site.

The average calculated flow rates for the 5 min before the speed drop, 5 min after the speed drop, and entire time during congested conditions are given in Table 4. Truck percentages averaged 1.7 percent at the I-35/US-67 sites and 2.1 percent at the I-410 site.

The I-410 site, which is similar to the US-290 site, had much different flow rates and distribution of traffic across lanes. At the US-290 site the outside lane had the highest volume before the speed drop, indicating that it reached its maximum flow

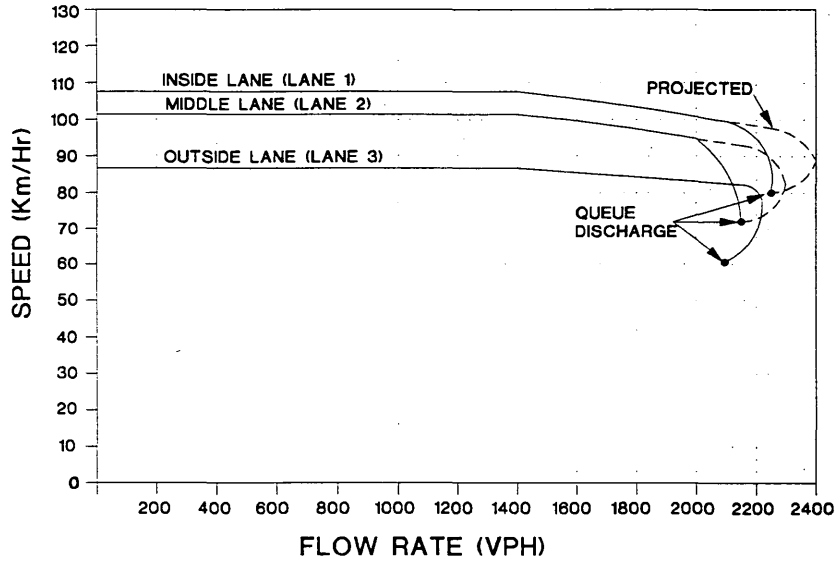


FIGURE 7 Speed-flow relationship for US-290 study site.

TABLE 3 Peak 15-min Flow Rates Across All Lanes

Highway	Observation	Peak 15 Minute Flow Rate (vphpl)				
		Lane 1	Lane 2	Lane 3	Lane 4	Average
U.S. 290	Average	2320	2244	2196	-	2253
I-410	1	2376	2212	1856	-	2148
	2	2476	2196	1856	-	2176
	3	2636	2096	1688	-	2140
	Average	2496	2168	1800	-	2155
I-35/US 67	1	2480	2180	2144	2124	2232
	2	2588	2320	2224	1856	2247
	3	2408	2240	2324	2172	2286
	Average	2492	2247	2231	2051	2255

TABLE 4 Comparison of Flow Rates Before and After Speed Drop at All Sites

Sample	5 Minutes Before Speed Drop (vphpl)					5 Minutes After Speed Drop (vphpl)					Entire Period After Speed Drop (vphpl)				
	Lane 1	Lane 2	Lane 3	Lane 4	Avg	Lane 1	Lane 2	Lane 3	Lane 4	Avg	Lane 1	Lane 2	Lane 3	Lane 4	Avg
U.S. 290															
Average	2076	2002	2210	-	2096	2266	2035	1989	-	2097	2246	2161	2090	-	2166
Std Dev	136	159	187	-	132	145	134	193	-	115	81	68	101	-	67
95% CI	75	88	104	-	65	80	74	107	-	64	45	38	56	-	37
I-410															
Average	2463	2166	1856	-	2162	1839	1790	1585	-	1738	1954	1864	1667	-	1828
Std Dev	285	89	188	-	153	164	14	49	-	46	180	95	46	-	68
95% CI	709	222	466	-	381	400	34	121	-	113	448	236	115	-	169
I-35/US 67															
Average	2679	2341	2134	1687	2210	2542	2238	2231	1856	2217	2396	2154	2172	2017	2185
Std Dev	128	79	254	182	137	90	75	56	86	70	47	56	41	80	47
95% CI	318	197	631	452	341	222	185	139	213	174	117	140	103	199	118

before the other lanes and broke down first. After the speed drop the inside lane increased in flow whereas the outside lane decreased in volume. A much different situation occurred at the I-410 site. All of the lanes decreased in flow after the speed drop. Further study found that the speed drop at the I-410 site was caused by downstream congestion that resulted in the flow reduction. Because of a downstream slowdown, the speed drop at the study site did not occur independently and the queue discharge flows were controlled by the downstream bottleneck. Although the site was affected by downstream congestion, the inside lane had a very high flow before congestion set in. This shows that it is possible to reach much higher flow rates under free-flow conditions than measured at the US-290 site in the inside lanes. This agrees with the hypothesis that the inside lane at the US-290 site prematurely dropped into queue discharge before reaching the maximum flow under free-flow conditions.

The most interesting results came from the I-35/US-67 site. At this site, traffic volumes were substantially higher than those measured at the US-290 site. The average flows before the speed drop were 2,679, 2,341, 2,134, and 1,687 vph for Lanes 1, 2, 3, and 4, respectively. After the speed drop the flow decreased in Lanes 1 and 2 and increased in Lanes 3 and 4, indicating that Lanes 1 and 2 broke down and transitioned Lanes 3 and 4 into queue discharge. This corresponds to the results and the model for US-290, which determined that once a lane or lanes break down, the other lanes are also subsequently broken down. Figure 8 shows the comparison of the flow rates between the I-35/US-67 site and the US-290 site for the 5 min before the speed drop and the total queue discharge period.

The flow rates for the I-35/US-67 site show the same statistical trends as the US-290 data. The standard deviations in mean flows between days tended to be lower after the speed

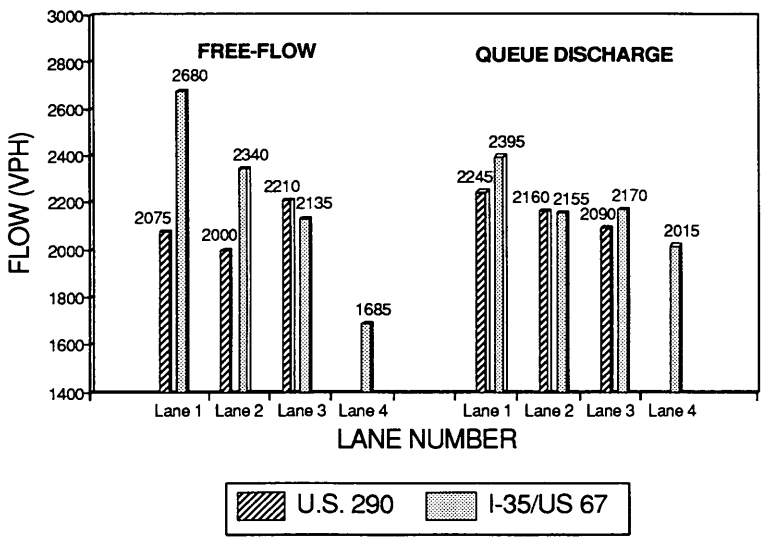


FIGURE 8 Flow-rate comparison between I-35/US-67 and US-290 at Tidwell.

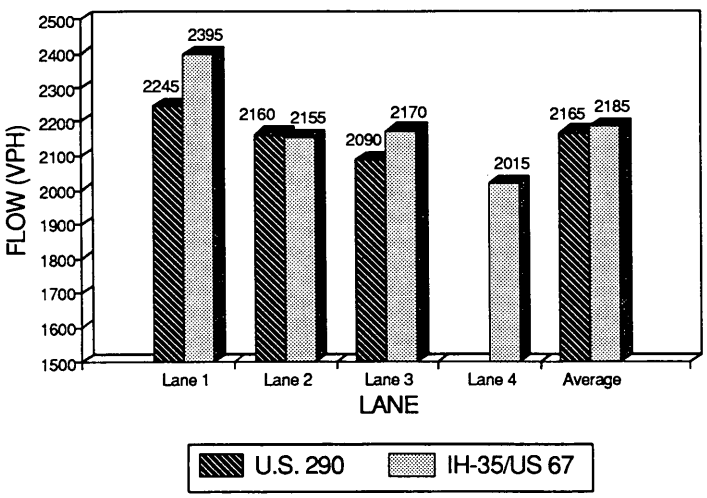


FIGURE 9 Average queue discharge flow rates.

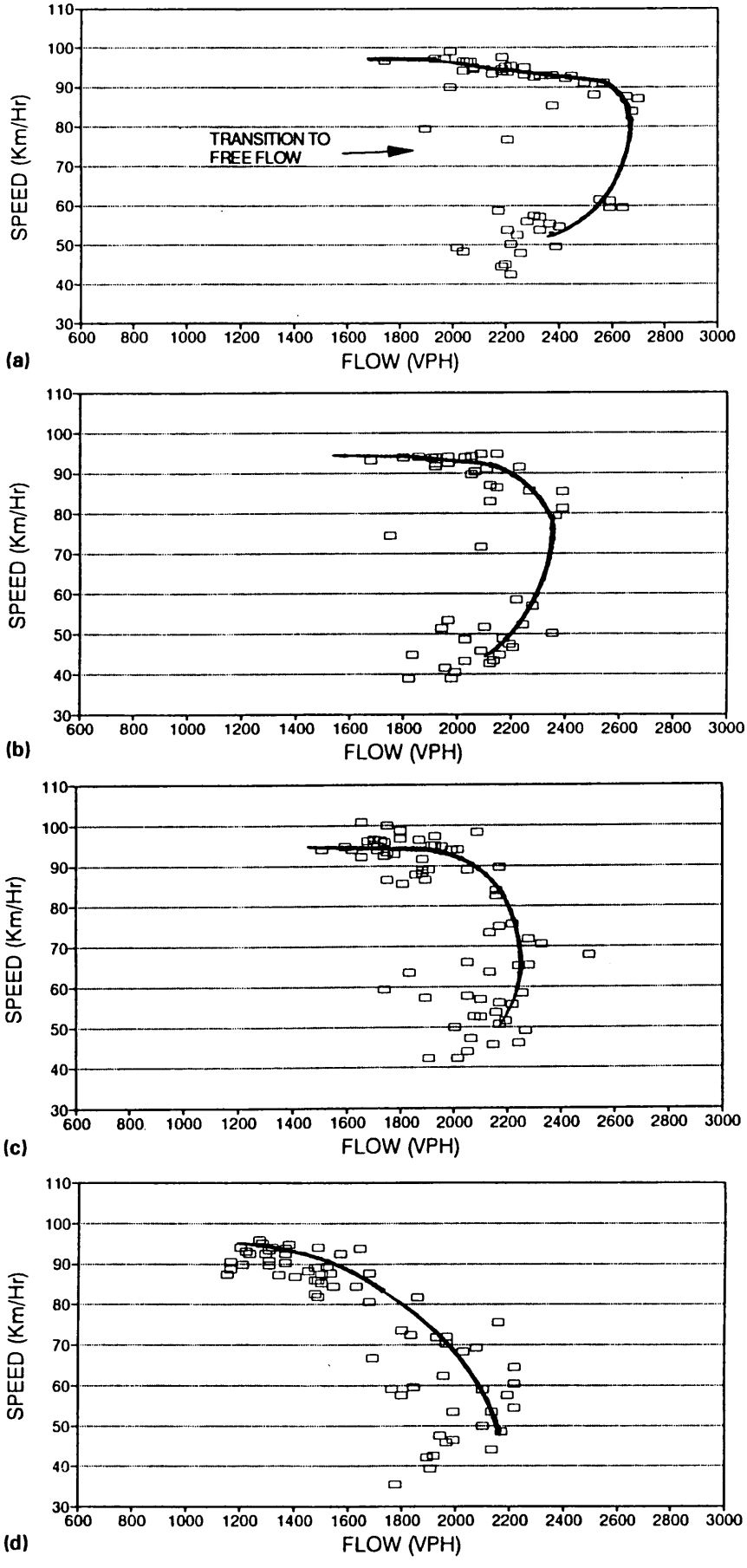


FIGURE 10 Speed-flow plots for I-35/US-67 study site, 5-min averages; *a*, Lane 1; *b*, Lane 2; *c*, Lane 3; *d*, Lane 4.

drop than before the speed drop. This further supports the conclusion that flows in the free-flow regions contain a large amount of variability, whereas the flows during queue discharge are not nearly as variable. Even with only three samples the confidence intervals for the average queue discharge flow rates are low, with 2,396 ± 117, 2,154 ± 140, 2,172 ± 103, and 2,017 ± 199 for Lanes 1, 2, 3, and 4, respectively. The average across all lanes was 2,185 ± 118 for the three samples. Although the individual lanes differ significantly, the average across all lanes during queue discharge was not statistically different than at US-290, which had an average flow of 2,166 ± 37. Figure 9 shows the average queue discharge flow rates adjusted for heavy vehicles (assuming an E_t of 2.0) for the US-290 and I-35/US-67 sites. The flows ranged from 2,200 to 2,400 pcphpl for individual lanes except in Lane 4 at the I-35/US-67 site, which was observed to not be a preferred lane and had lower flows. The averages for both sites across all lanes were nearly identical, with 2,230 and 2,220 pcphpl. For these reasons, queue discharge appears to be the most consistent flow for estimating the maximum sustainable flow of a facility. Although much higher flows are obviously possible, they typically do not occur across all lanes and would be difficult to maintain.

Even in queue discharge, Lane 1 at the I-35/US-67 site continually had very high flows, averaging 2,396 vph. It is difficult to determine the reason for these very high queue discharge flow rates. One explanation for the high flow rates during queue discharge in Lane 1 could be the difference in the type of bottleneck at the I-35/US-67 site.

Speed-Flow Relationship at Validation Sites

Because the I-410 site was affected by downstream congestion, the relationship between speed and flow was not evaluated at this site. Figure 10 shows the speed-flow plots for each lane at the I-35/US-67 site. Also shown in Figure 10 are the projected speed-flow relationships suggested by the data. The scattered points in the center of the relationship (low flow and higher speed) represent recovery to uncongested conditions. This phenomenon was not present in the US-290 data.

Figure 11 shows the speed-flow model for the I-35/US-67 site. As can be seen in this figure, Lanes 1 and 2 reach their peak flow rates during free-flow conditions while Lanes 3 and 4 are prematurely transitioned into queue discharge. One interesting aspect of this site is the relation of Lanes 2 and 3. Both have approximately the same free-flow speeds and transition into nearly identical queue discharge flow rates, yet Lane 2 reaches it peak in free-flow conditions and Lane 3 peaks after the speed drop in queue discharge. These two lanes illustrate the effects of lane interaction. Because Lane 2 broke down, the turbulence transitioned Lane 3 into queue discharge before it reached its maximum flow rate. The extremely high flows in Lane 1 make it difficult to determine if the turbulence created by Lane 1 restricted Lane 2 from reaching its maximum free-flow flow rate.

These results support the hypothesis that when one or more lanes break down the other lanes are prematurely transitioned into queue discharge conditions. The data from the I-35/US-67 site show that extremely high flows are possible in free-flow conditions and that relatively high flows also occur in queue discharge. Therefore, the measured flows are a function of the interactions between the lanes.

Although lane interaction is not a new idea, it seems to have been forgotten in many studies. One of the most significant problems is when lanes are combined to form an overall speed-flow relationship. In free-flow conditions all lanes operate similarly, and all lanes can be combined to produce average flows. Once the facility begins to break down, average flow rates and speeds are significantly different from the individual lanes, and combining all lanes to produce a single relationship is misleading. The interaction between lanes may help explain the variety of results obtained in earlier studies.

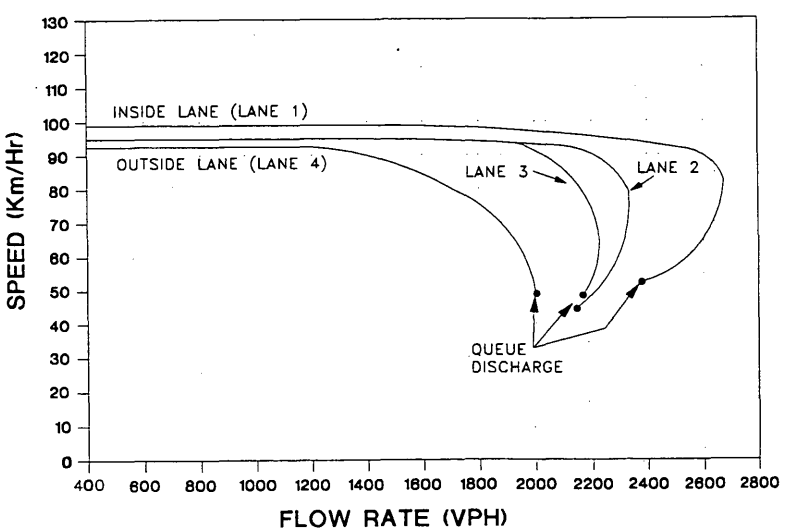


FIGURE 11 Speed-flow relationship for I-35/US-67 study site.

RESULTS AND CONCLUSIONS

On the basis of the analysis of the US-290, I-410, and I-35/US-67 study sites the following conclusions can be made:

- Variance in flow rate decreases after the speed drop to queue discharge.
- Peak flows for individual lanes occur in free-flow conditions before breakdown.
- Peak flows during free-flow conditions do not generally occur in all lanes on a facility because of an imbalance of flow rates between individual lanes. This prematurely transitions the flow from free-flow into queue discharge conditions.
- Bottleneck configuration may influence the maximum possible flow obtainable during free-flow and possibly queue discharge conditions.
- Queue discharge appears to be the best estimate for maximum sustainable flow and capacity.

ACKNOWLEDGMENTS

This study was sponsored by the Texas Department of Transportation and FHWA, U.S. Department of Transportation.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. V. F. Hurdle and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.
3. K. Agyemang-Duah and F. L. Hall. Some Issues Regarding the Numerical Value of Freeway Capacity. *Proc., International Symposium on Highway Capacity*, July 1991, pp. 1–15.
4. T. Urbanik, W. Hinshaw, and K. Barnes. Evaluation of High-Volume Urban Texas Freeways. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 110–118.
5. F. L. Hall and K. Agyemang-Duah. Freeway Capacity Drop and the Definition of Capacity. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 91–98.

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

Suggested Procedures for Analyzing Freeway Weaving Sections

BARBARA OSTROM, LANNON LEIMAN, AND ADOLF D. MAY

Most speed-oriented methodologies have proved unsatisfactory in describing the behavior of freeway weaving sections. Research done at the Institute of Transportation Studies at the Berkeley campus of the University of California, with the support of the California Department of Transportation, has investigated several types of simple weaves using point flow estimation. For major weaves, a type of simple weave, an analysis method was developed using point flow estimates based on movement percentages. This methodology estimates point flows within 10 percent of the actual observations for 70 percent of the cases within empirical limits. For ramp weaves a regression-based set of equations to estimate total point flow directly has proven more effective. The equations predict total point flow within 10 percent of the empirical values for 90 percent of the data. The methodology developed for major weaves has been implemented in an interactive menu-driven computer program, FREWEV. This is the first step towards an integrated freeway model, FRELANE, which currently includes major weaves and ramp weaves.

Weaving sections on freeways are one of the greatest areas of conflict in normal freeway operations. They are also the most difficult sections of the freeway to analyze satisfactorily. Various approaches and measures of effectiveness have been used during the last half century. Studies to validate the weaving methodology presented in the 1985 *Highway Capacity Manual* (HCM) have shown speed to be a very poor predictor of weaving operations (*1–3*). Ongoing research in California is exploring the use of point flow estimation and using density as a measure of effectiveness.

HISTORICAL REVIEW

The 1950 HCM presented the first method for predicting the capacity and operating speeds of freeway weaving sections (*4*). The 1965 HCM contained a revised version of this 1950 HCM graphical method with added emphasis on quality of flow (*5*). *Interim Materials on Highway Capacity*, published in 1981 by Polytechnic Institute of New York (PINY) (*6*), contained a new method for estimating weaving and nonweaving speeds for simple weaving sections (*7*) and a modification of the earlier 1965 HCM method (*8*). The 1985 HCM chapter on weaving analysis was based on the JHK algorithm, which predicted weaving and nonweaving speeds (*9*).

In 1987, on the basis of recognized needs for additional research on freeway weaving in California and encouraged by national recognition for similar research, the Institute of Transportation Studies at the Berkeley campus of the University of California (ITS-UCB), with support from the California Department of Transportation (Caltrans), began a 5-year research program.

Institute of Transportation Studies, 109 McLaughlin Hall, University of California at Berkeley, Berkeley, Calif. 94720.

ONGOING RESEARCH

The first phase of the research at ITS-UCB was to evaluate major weaves using existing methods. The methods evaluated included those of HCM, 1965 (*5*); Leisch (*10*); PINY (*7*); JHK (*11*); HCM, 1985 (*9*); and Fazio (*12*). These speed-based predictive methodologies produced estimates with errors of more than 10 percent for weaving and nonweaving speeds. In modeling the section as an overall unit, the speed-based methods were perceived to overlook the importance of the interactions within and between lanes. The speed predictions were also poor when a similar analysis was done with ramp weave sections (*1,2*). The Caltrans Traffic Bulletin 4 method (Level D), an alternative lane flow method for ramp weaves, was also evaluated (*13*). When applied to the same data set as the speed prediction methods, Level D predicted volumes that were within 10 percent of ramp movement volumes along the weaving section. Based on this information, a lane flow approach such as the Level D method appeared to be more promising than the speed-based methodologies for analyzing major weaves (*3*).

The second phase of the research was directed toward improving methods to analyze and design major weaving sections. Comprehensive field data was collected for 10 major weaving sections in California. Using these data sets and applying the INTRAS simulation model (*14*), a new analytical approach, point flow by movement, was developed. The proposed procedure predicts vehicle lane flow rates at frequent intervals within the weaving section as a function of prevailing traffic flow and geometric conditions. This work has been discussed in previous publications (*3,15*).

The third phase of the research program, just being completed, is concerned with developing point flow methodologies for other types of freeway weaving and ramp configurations. The models suggested predict total point flows at frequent intervals along the freeway (*16*).

Data sets for ramp weaving sites were provided by Caltrans from its own extensive study of ramp weaving sections (*1,2*). Using this data, a total point flow method was developed for analyzing ramp weaving sections (*17*).

DEFINITION OF WEAVING SECTIONS

A weaving section exists when at least one movement must make at least one lane change to enter or exit the freeway. Only simple weaving sections are considered in this discussion.

A simple weave has only one on-ramp and one off-ramp connected by one or more auxiliary lanes. It is isolated from the influence of other ramps. Simple weaves include ramp weaves and major weaves. A ramp weave is a one-lane on-ramp connected to a one-lane off-ramp by an auxiliary lane. A major weave has an on-ramp connected to an off-ramp by one or more auxiliary lanes; at least one of the ramps must have two lanes and the other ramp may have one or two lanes. Examples of the weaving sections that are included in this discussion are shown in Figure 1.

ALTERNATIVE METHODOLOGIES

All of the methodologies described in the following are used to predict the flows at a set of points within the critical area of the weaving section. The critical area includes all lanes beginning, ending, or beginning and ending at ramps and the right-most through freeway lane. The critical areas for selected weaving sections are shaded in Figure 1. The points at which the analysis is done include the merge, the diverge, 76.2 m (250 ft) downstream of the merge, and at 152.4-m (500-ft) increments from the merge to the end of the section. The use of multiple analysis points comes from the Level D methodology. The inclusion of the 76.2-m (250-ft) point comes from empirical analysis of data collected during the research. The analysis at this point is important because it can be shown that the majority of the lane changing takes place within 152.4 m (500 ft) of the ramp of interest (*3*).

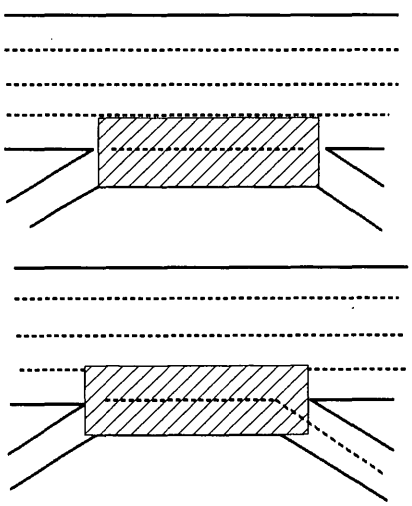


FIGURE 1 Simple weaving sections: *top*, ramp weave; *bottom*, major weave.

Level D Method

The Level D method was developed by Caltrans in the early 1960s (*13*). It is intended to evaluate ramps and ramp weave sections. The Level D method is only appropriate for a ramp weaving section operating under high or near capacity traffic flow conditions. Given the section length and volumes in the weaving section, Level D predicts the distribution of traffic in the two right-most lanes of the freeway weaving section. The distribution of each ramp movement is solely a function of section length. The amount of through traffic in the right-most through freeway lane is a function of the total freeway flow. The method allows the analyst to calculate estimates of the point flows for individual movements throughout a section. The results are highly sensitive to the estimate of non-weaving traffic in the right-most through freeway lane. Errors in estimation of total volumes at points in this lane can be attributed to failure to correctly predict freeway to freeway volumes (*17*).

Point Flow by Movement Method

Point flow by movement has its origins in the Level D method. The point flow by movement method models the distribution of movements and the amount of lane changing within the analysis area. It predicts the distribution of each movement throughout the section and estimates the total volume at a point as a sum of the individual movements. Unlike the Level D methodology, the point flow by movement method uses ramp movement percentages, which are dependent on volumes, or on weaving section length or on both volumes and weaving section length. Freeway-to-freeway percentages are not necessarily constant along the section and may be functions of volume or site geometry. The point flow by movement method for major weaves uses average percentages obtained from field data and simulation of longer sections under observed volume conditions. With a fairly wide range of values for most observed percentages, the use of an average represents the best fit.

The many equations required to produce point flows with this method and an example can be found in the initial research report (*3*).

Total Point Flow Method

Total point flow is a regression-based methodology that directly predicts the total flow at an analysis point within a weaving section (*16*). The equations, estimated separately for each analysis location, may generally be expressed as

flow in Lane N at XX ft

$$= \Theta_0 + \Theta_1 FF + \Theta_2 FR + \Theta_3 RF + \Theta_4 RR$$

where

Θ_i = coefficients,
FF = freeway-to-freeway movement,
FR = freeway–to–off-ramp movement,
RF = on-ramp–to–freeway movement, and
RR = on-ramp–to–off-ramp movement.

The coefficients in the equations can provide information on the influence of the component flows within a conflict area, but the coefficients do not represent the percentages of each movement at a specific analysis point.

Movement flows can be estimated using some very strong assumptions about the behavior of individual movements and net lane changing. Net lane changing is calculated using the difference in total volume between two points across one or more lanes.

PROPOSED ALTERNATIVE METHODOLOGY FOR SELECTED WEAVES

Major Weaves

Before the research conducted at ITS-UCB, there was no analysis methodology that addressed major weaves. A major weave with at least one two-lane ramp has not been reliably analyzed by methods based on speed or the Caltrans Traffic Bulletin 4. ITS-UCB proposed a point flow by movement methodology for analyzing this type of section. The validity of the model was investigated in the initial research (*3*). The predictive ability of the method was investigated with a simple random sample of empirical data. Approximately 70 percent of the point totals are within 10 percent of observed values (*18*), as shown in Figure 2.

The total point flow methodology was also applied to major weaving sections. It was no better at predicting the total point flows than the point flow by movement method (*18*). The point flow by movement methodology allows for separate predictions of each movement. It permits the analyst to investigate individual movements without assumptions about their behavior. It also provides a secondary check on the existence of under capacity conditions. With these advantages, the point flow by movement methodology was selected for analyzing major weaves.

Ramp Weaves

The three alternative methodologies were investigated for analyzing a ramp weave section: Level D, point flow by movement, and total point flow (*17*). A subset of the Caltrans data (*1,2*), which contained information on all of the movements throughout the entire length of the weaving section, was used for the evaluation of the three methodologies when applied to ramp weaves.

Using a set of sites for which all three methods could be applied, it was determined that Level D produced estimates of total point flows within 10 percent for 40 percent of the analysis points. Seventy percent of the errors occurred in the right-most through lane because of incorrect estimates of freeway-to-freeway volumes. The point totals calculated using point flow movement estimates were better than those calculated by Level D, primarily through improved freeway-to-freeway volume estimates (*17*).

A set of total point flow equations was developed on the basis of data from three sections of similar length. The resulting total point flow estimates were an improvement over the point flow by movement method. The regression equations predicted total volumes within 10 percent of the observed volumes for 90 percent of the analysis points (*17*). A comparison of the predictions made with the total point flow equations and the predictions from the Level D method is shown in Figure 3. The Level D method and the total point flow method both continue to be evaluated for analyzing ramp weaves.

IMPLEMENTATION OF ALTERNATIVE METHODOLOGY FOR MAJOR WEAVES

An interactive menu driven computer program, FREWEV, has been developed for designing and analyzing major freeway weaving sections (*18,19*). The analysis method imple-

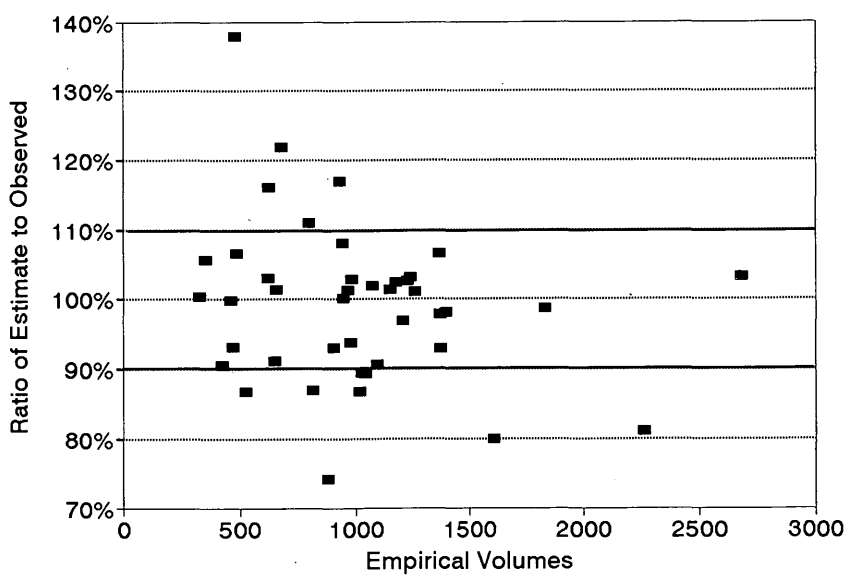


FIGURE 2 Accuracy of major weave estimates.

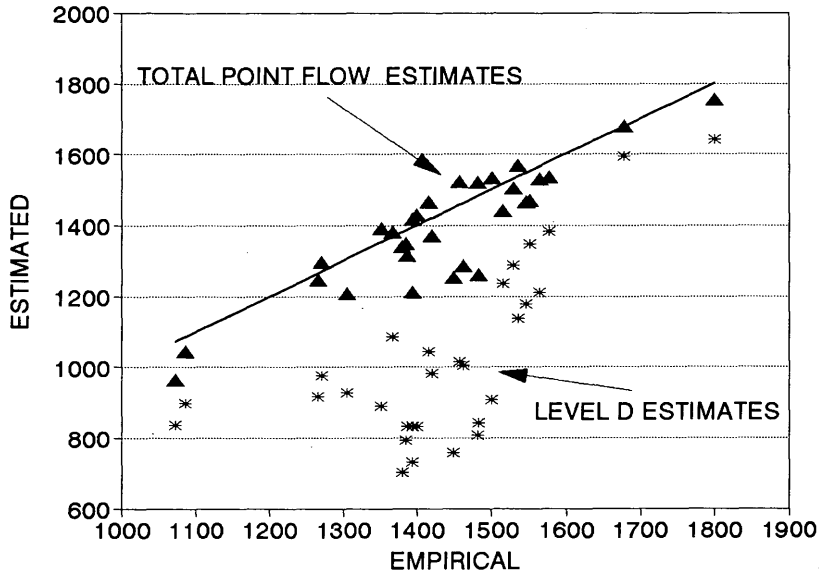


FIGURE 3 Comparison of total point flow and Level D estimates.

mented within FREWEV is the point flow by movement method described in the Alternative Methodologies section of this paper and discussed in detail in the initial research report (*3*).

Overview of FREWEV Model

The FREWEV model can analyze the five types of major freeway weaving sections that are shown in Figure 4. These weaving sections can be classified as Type B or C according to the HCM (*9*). The model evaluates only those lanes that are actually influenced by weaving movements. Those lanes define the critical area and are shaded in Figure 4. Note that the model on which the FREWEV program is based assumes that the freeway is not congested and that the weaving section is not influenced by ramps located upstream or downstream of the waving section.

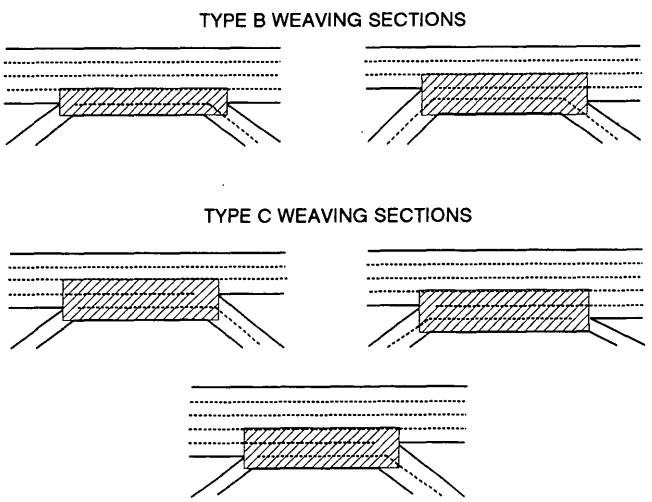


FIGURE 4 Types of weaving sections analyzed by FREWEV.

The input for FREWEV consists of design and demand data for the weaving section and for the sections just upstream and downstream of the weaving section. The design features include the subsection lengths, subsection capacities, position and capacities of on- and off-ramps, number of lanes, and subsection grades. The demands need to be provided for the mainline origin and each on-ramp and off-ramp as well as the ramp-to-ramp demand for the major weaving section. Demand data also include percentage trucks and peak hour factors. The user may supply truck conversion factors and freeway-to-freeway percentages for special circumstances. The data set can be saved for later retrieval and additional analysis.

The FREWEV model calculates the amount of traffic by movement at points along a weaving section for each of the lanes in the conflict area. The four movements involved are those traditionally associated with weaving—through freeway (FF), freeway to off-ramp (FR), on-ramp to freeway (RF), and on-ramp to off-ramp (RR). By summing the flows of the four individual movements at each point, the total flow at each point is determined. The density at each point is then calculated as a function of volume and number of lanes. Density is the criterion to determine the level of service (LOS) at each point based on the LOS range as defined in the initial phase of the research (*3*). The model also calculates the amount of traffic crossing the lane boundaries for each segment between the points that are used for point flows.

FREWEV produces a variety of screen displays and hard copy outputs. Sample output from the FREWEV analysis of a Type C, five-lane major weaving section with a one-lane on-ramp and a two-lane off-ramp is shown in Figures 5 and 6. Figure 5 displays the point flows by movement superimposed on the schematic of the three-lane conflict area. Figure 6 shows the three subsection freeway segment with the density-based LOS printed at the defined points for each lane in the conflict area.

The FREWEV model can be run in two different modes. The empirical mode allows the program to analyze only weaving sections that adhere to the design and demand values for

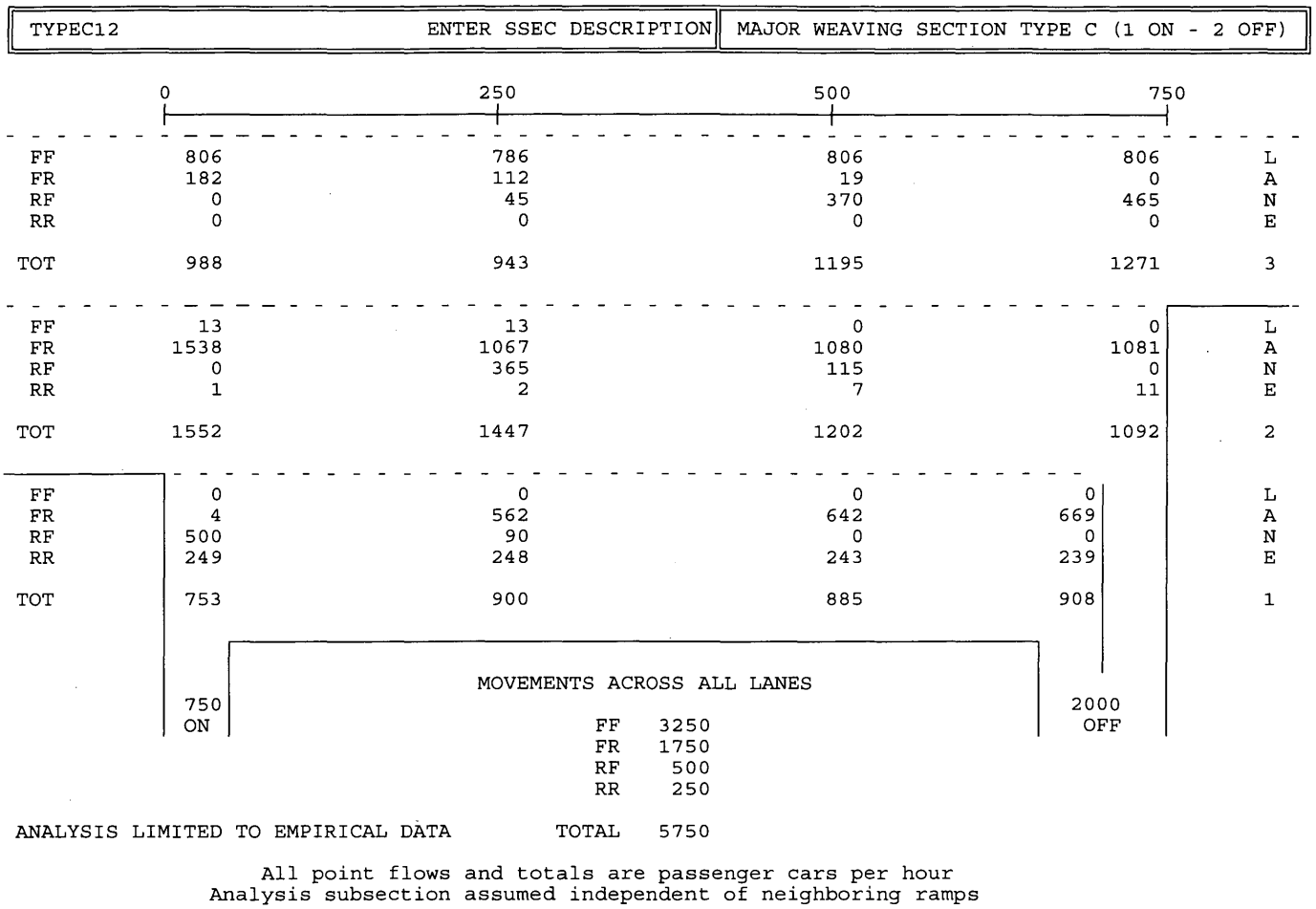


FIGURE 5 FREWEV output: point flows superimposed on schematic of major weaving section.

which empirical data were observed. The simulation mode allows the program to analyze weaving sections even if they have design and demand values outside the range for which empirical data were collected. The extension of the data ranges has been developed through simulation. The tables of ranges that are valid for each of the two modes and a detailed description of the method used for generating the simulation ranges can be found in the *FREWEV User's Guide* (*19*) and a separate technical document (*18*).

FREWEV is written for the IBM personal computer and requires a math coprocessor. Printouts of the freeway geometry require a printer with the IBM character set.

CONCLUSIONS

Point flow estimation techniques developed in this research produced more reliable estimates of freeway weaving section behavior than other identified tested methods. Sets of movement percentage tables have been derived on the basis of extensive field data sets and extended by simulation for freeway major weaving sections. Regression equations for empirical conditions have been produced for ramp weaving analysis. The analysis for both major weaves and ramp weaves is provided for the right-most lanes at frequent intervals within the weaving sections. Two interactive menu-driven computer programs have been developed to design and analyze freeway weaving sections. The FREWEV computer model analyzes major weaving sections and permits comparison between alternative designs. The FRELANE model, an extension of the FREWEV model, currently permits the analysis and comparisons of major weaving and ramp weaving sections. Further research is under way to extend the FRELANE model to include multiple weaving sections and a variety of ramp merge and diverge situations.

ACKNOWLEDGMENTS

Many individuals from the Berkeley campus of the University of California have participated in this multiyear research project. The authors wish to recognize Michael Cassidy, Patrick Chan, Bruce Robinson, Mark Vandehey, George Mazur, James Holmes, Daniel Preslar, and Daniel Thompson.

The authors thankfully acknowledge that this research was sponsored by the California Department of Transportation and FHWA. The authors also wish to thank the following individuals, who have provided advice throughout the life of this project: Fred Rooney, Howard Fong, and Greg Tom from Caltrans; and Stephen Cohen and Jeffrey Lindley from FHWA.

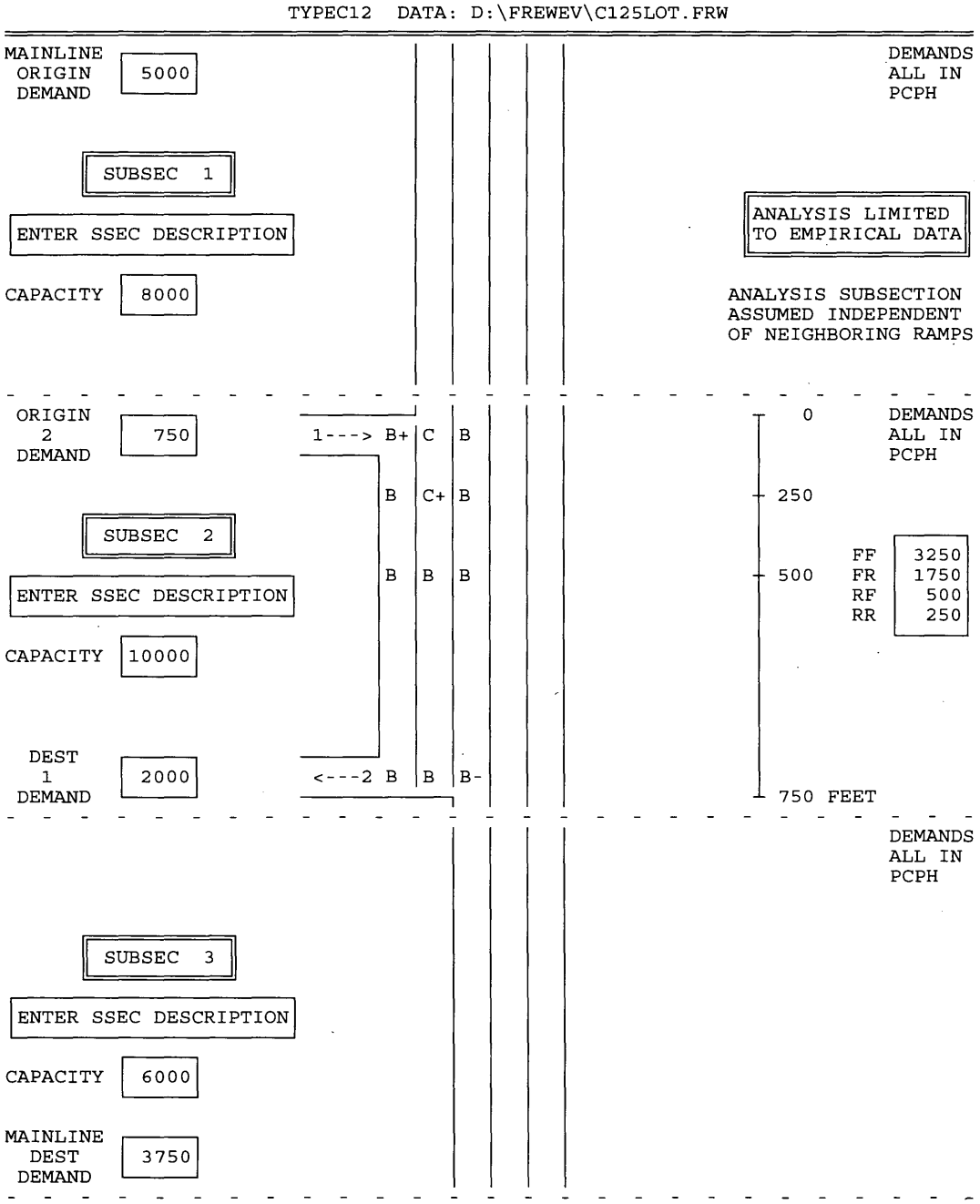


FIGURE 6 FREWEV output: freeway geometry with LOS in critical area.

REFERENCES

1. H. F. Fong and F. D. Rooney. *Weaving Data Near One-Lane Ramps, Peak Periods.* California Department of Transportation, Sacramento, 1990.
2. H. K. Fong and F. D. Rooney. *Weaving Data Near One-Lane Ramps, Non-Peak Periods.* California Department of Transportation, Sacramento, 1990.
3. M. J. Cassidy, P. Chan, B. Robinson, and A. D. May. *A Proposed Analytical Technique for the Design and Analysis of Major Freeway Weaving Sections.* Research Report UCB-ITS-RR-90-16. Institute for Transportation Studies, University of California, Berkeley, 1990.
4. *Highway Capacity Manual.* Bureau of Public Roads, U.S. Department of Commerce, 1950.
5. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
6. *Transportation Research Circular 212: Interim Materials on Highway Capacity.* TRB, National Research Council, Washington, D.C., Jan. 1980.
7. R. Roess et al. *Freeway Capacity Analysis Procedures.* Final Report. Polytechnic University of New York, May 1978.
8. J. E. Leisch. *Capacity Analysis Techniques for Design and Operation of Freeway Facilities.* Report FHWA-RD-74-24. FHWA, U.S. Department of Transportation, 1974.
9. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
10. Jack E. Leisch & Associates. *Procedure for Analysis and Design of Weaving Sections.* FHWA, U.S. Department of Transportation, 1983.

11. W. R. Reilly, J. H. Kell, and P. J. Johnson. *Weaving Analysis Procedures for the New Highway Capacity Manual.* Traffic Systems Division, Offices of Research and Development, FHWA, U.S. Department of Transportation, 1984.
12. J. Fazio. *Development and Testing of a Weaving Operational Analysis and Design Procedure.* M.S. thesis. University of Illinois, Chicago, 1985.
13. K. Moskowitz and L. Newman. *Traffic Bulletin No. 4, Notes on Freeway Capacity.* Division of Highways, Department of Public Works, State of California, July 1962.
14. D. A. Wicks and E. B. Lieberman. *Development and Testing of INTRAS, A Microscopic Freeway Simulation Model.* Report FHWA/RD/80/106-107. FHWA, U.S. Department of Transportation, 1980.
15. M. J. Cassidy, A. Skabardonis, and A. D. May. *Operation of Major Freeway Weaving Areas: Recent Empirical Evidence.* Working Paper UCB-ITS-WP-88-11. Institute of Transportation Studies, University of California, Berkeley, Dec. 1988.
16. B. W. Robinson, M. A. Vandehey, G. D. Mazur, and A. D. May. *Improved Freeway Analysis Techniques: Ramp and Weaving Operations for Freeway Lane Model.* Research Report UCB-ITS-RR-92-4. Institute of Transportation Studies, University of California, Berkeley, March 1992.
17. J. R. Holmes, D. Preslar, D. Thompson, L. Leiman, B. K. Ostrom, and A. D. May. *Simulation and Empirical Relationships in Developing a Freeway Lane Model.* Technical Document UCB-ITS-TD-92-2. Institute of Transportation Studies, University of California, Berkeley, 1992.
18. B. K. Ostrom, L. Leiman, and A. D. May. *FREWEV: A Design and Analysis Model for Major Weaving Sections: Data and Assumptions.* Technical Document UCB-ITS-TD-92-1. Institute of Transportation Studies, University of California, Berkeley, 1992.
19. L. Leiman, B. K. Ostrom, and A. D. May. *FREWEV: A Design and Analysis Model for Major Freeway Weaving Sections: User's Guide.* Research Report UCB-ITS-RR-92-5. Institute of Transportation Studies, University of California, Berkeley, March 1992.

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

Revised Queueing Model of Delay at All-Way Stop-Controlled Intersections

ALAN J. HOROWITZ

An existing M/G/1 queueing model of delay at all-way stop-controlled intersections is updated to reflect recent empirical evidence of driver behavior. In particular, the queueing model has been expanded to account for turning and coordination between drivers who get priority at the same time. An exact expression for an important variable in the model, the variance of service time, could not be found. However, an approximate expression could be verified using a Monte Carlo simulation. A queueing model can produce values of delay that are similar to strictly empirical models, but that apply to a much greater range of intersection volumes.

In a series of empirical studies, Kyte has developed statistical models of delay at all-way stop-controlled (AWSC) intersections (*1,2*). Kyte's models formed the basis for the "Interim Material on Unsignalized Intersection," recently published by TRB's Committee on Highway Capacity and Quality of Service in *Transportation Research Circular 373* (*3*). The procedure presented in that circular, TRC 373, is quite simple, consisting only of the application of two equations. The first equation calculates subject approach capacity from intersection movement data; the second equation calculates delay from the approach's volume-to-capacity ratio. The most obvious drawback to the TRC 373 procedure is that its relation for capacity applies to a rather narrow range of traffic conditions.

It might be possible to build a queueing model that closely replicates the results of the TRC 373 procedure over its valid range of inputs, but permits extrapolation to all possible traffic conditions. A queueing model incorporates more features of traffic theory and is more intuitive, but it would be somewhat more difficult to calculate.

A particularly interesting queueing model was developed by Richardson (*4*), who was able to build a model consisting of four single-server queues—one for each approach—that could be iteratively solved for the mean and variance of service time at each approach. Richardson assumed exponential arrivals, but was able to avoid making any restrictive assumptions about the probability distribution of service times. With knowledge of the approach's volume, its service time, and the variance of service times, Richardson was able to calculate delay at an approach as if he were looking at a simple M/G/1 queue.

Unfortunately, Richardson's model cannot be made to replicate TRC 373. Richardson did not include any means of adjusting for more delay due to left turns or due to coordination of drivers at multilane approaches, or for less delay due to right turns. It is not obvious that such delay adjustments can be properly incorporated into Richardson's model, because the expression for the variance of the service time (necessary for delay calculation) becomes exceptionally complex.

The purpose of this paper is to create and evaluate a revised M/G/1 queueing model, generally following Richardson's framework, which encompasses the same traffic behavior as the TRC 373 procedure and can give comparable results.

Center for Urban Transportation Studies, University of Wisconsin-Milwaukee, P.O. Box 784, Milwaukee, Wis. 53201.

REVIEW OF RICHARDSON'S MODEL

Richardson hypothesized that the service time for any vehicle on the subject approach was the sum of two quantities: (a) the vehicle's minimum headway, defined as the time it takes for the vehicle to execute its maneuver once it gains priority, and (b) the amount of time it must wait for one or more vehicles on the conflicting approaches. Richardson assumed that the minimum headway was constant for all approaches. Richardson further assumed that the time necessary to wait for a conflicting vehicle was also a constant and slightly less than the minimum headway. Consequently, the only source of variation in service time at an approach would be the presence or absence of a conflicting vehicle. Richardson was able to demonstrate that his assumptions were consistent with the limited amount of AWSC intersection data available at that time.

Average service time at any approach, s, is

$$s = t_m + (t_m - t_c)\rho_c \quad (1)$$

where

t_m = minimum headway,
t_c = adjustment to conflicting vehicle minimum headway, and
ρ_c = utilization ratio and also the probability of having a conflicting vehicle (the probability that there is at least one vehicle on either of the two conflicting approaches).

$$\rho_c = 1 - (1 - \rho_{c1})(1 - \rho_{c2}) \quad (2)$$

where ρ_{c1} is the probability of having a vehicle on the first conflicting approach and ρ_{c2} is the probability of having a vehicle on the second conflicting approach. From conventional queueing theory, the probability of having a vehicle on a conflicting approach can be found from the product of the arrival rate and the service time. For the ith conflicting approach,

$$\rho_{ci} = \lambda_{ci} s_{ci} \quad (3)$$

where λ_{ci} is the arrival rate (or volume) for conflicting approach i and s_{ci} is the service time for conflicting approach i.

Where there is more than one lane at an approach, Richardson split the traffic evenly across lanes and treated each lane as if it were independent.

$$\rho_{cin} = 1 - (1 - \rho_{ci}/n)^n \quad (4)$$

where ρ_{cin} is the probability of having a vehicle at the stop line of at least one lane at an approach with n lanes. Richardson did not attempt to assign right-turning vehicles to the right-most lane or the left-turning traffic to the left-most lane.

It is important to note that a separate set of Equations 1 to 3 (and possibly Equation 4) exist for each approach and that the approaches are interdependent. These equations represent a set of nonlinear simultaneous equations, which are sufficient to solve for the average service time and the utilization ratio of each approach. Richardson found that the method of successive approximations worked very well as a solution technique.

When the service time has been found for each approach, the variance of the service time, σ^2, can be found from

$$\sigma^2 = t_m^2(2t_m - t_c - s)/(t_m - t_c) + (2t_m - t_c)^2(s - t_m)/(t_m - t_c) - s^2 \quad (5)$$

The delay can be found from

$$D = (2\rho - \rho^2 + \lambda^2\sigma^2)/[2\lambda(1 - \rho)] \quad (6)$$

where ρ is the utilization ratio and λ is the volume on the subject approach.

Richardson's model is elegant. It requires few assumptions but provides a good description of driver behavior at AWSC intersections.

IMPLICATIONS OF TRC 373

TRC 373 implies that driver behavior at an AWSC intersection is much more complex than the simple queueing relations of Richardson's original model. According to TRC 373 there are increases and decreases in subject approach delay due to turning, and there are increases in delay associated with the need for coordination among drivers at multilane approaches.

To account for turning and coordination in Richardson's model it is necessary to allow the average minimum headway, t_m, to vary at each approach according to the amount of turning and the presence of additional vehicles on the subject or opposing approaches. Thus, average minimum headway would be a complex expression

$$t_m = t_b + \sum_r t_r p_r \quad (7)$$

where

t_b = base minimum headway,
t_r = adjustment for condition r, and
p_r = probability that condition r exists.

From TRC 373 and Kyte's original data, it appears that the following conditions would be important in estimating capacity:

- One left-turning vehicle among the subject and opposing approaches,
- A left-turning vehicle at both the subject and opposing approaches,
- One right-turning vehicle among the subject and opposing approaches,
- A right-turning vehicle at both the subject and opposing approaches,
- A second vehicle on the subject approach (multilane approaches),
- One vehicle on the opposing approach, and
- Two vehicles on the opposing approach (multilane approaches).

Some of these conditions are mutually exclusive. All of these conditions apply to either the subject or opposing approach, as there is no need to deal separately with turning on conflicting approaches in this queueing model.

In addition, it is plausible that the adjustment to conflicting approach minimum headway, t_c, would be related to intersection geometry. Multilane approaches would require conflicting vehicles to travel greater distances before yielding priority to a vehicle on the subject approach.

A set of preliminary values for t_b, t_r, and t_c in units of seconds is given in Table 1. These parameters were chosen to match the capacity as given by the raw data and regression equations of Kyte and Marek (*2*). Because of the major conceptual differences between Kyte's regression models and queueing models, these parameters were obtained by trial and error and should be considered accurate to just 0.25 sec.

Some observations about the parameters given in Table 1 would help interpret the queueing model. First, the difference between the two values of t_c (1.0 sec) is roughly the amount of time it would take a vehicle to travel past the additional width of a four-lane road. Note that t_c is subtracted from the value of the average minimum headway of the conflicting approaches, so its value should become smaller as the intersection becomes wider. Second, the parameters of the first

TABLE 1 Preliminary Parameters for Enhanced M/G/1 Queueing Model for Delay at AWSC Intersections

Traffic Condition	Wait, seconds
Base Minimum Headway, t_b	3.6
Adjustments for Traffic Conditions, t_r	
1. One left vehicle	1.0
2. Two left vehicles	1.0
3. One right vehicle	-0.5
4. Two right vehicles	-1.0
5. Second vehicle on subject approach	1.0
6. One vehicle on opposing approach	0.25
7. Two vehicles on opposing approach	1.0
Adjustments to Conflicting Minimum Headway, t_c	
One-lane approach	0.5
Two-lane approach	-0.5

two conditions (left-turning vehicles) are the same. Because these conditions are mutually exclusive, the model is insensitive to having two left-turning vehicles on two approaches instead of one left-turning vehicle. Third, t_r and t_c have been rounded to the nearest 0.25 sec.

The capacity of an intersection under certain artificial conditions can be estimated directly from the parameters given in Table 1. For example, the capacity of a single-lane approach at a fully saturated intersection with no turning would be exactly 500 vehicles per hour [3,600/(3.6 + 0.25 + 3.6 + 0.25 − 0.5)]. The capacity of a two-lane approach under similar conditions would be 615 vehicles per hour [(2 × 3,600)/(3.6 + 1.0 + 1.0 + 3.6 + 1.0 + 1.0 − 0.5)]. The corresponding values from TRC 373 are 525 and 625, respectively. Some other simple comparisons are given in Table 2.

Unfortunately, many of the conditions that permit hand calculation of capacity are outside the range of TRC 373. Nonetheless, it can be seen that the M/G/1 model can produce reasonable estimates of capacity under a wide range of traffic conditions.

It has been observed that drivers exhibit different stopping behavior (full stop versus partial stop) depending on their intended movement and the presence of a conflicting vehicle. To some degree this behavior is reflected in the parameters given in Table 1. However, randomness in stopping behavior has not been accounted for in this M/G/1 model.

Richardson assigned vehicles to lanes as if they were all identical. At multilane approaches, turning vehicles should be assigned to the most appropriate lane. Thus, Equation 4 must be made more elaborate to handle the few situations in which there are insufficient through vehicles to balance flows across all lanes.

$$\rho_{ncin} = 1 - \prod_{j=1}^{n} (1 - \rho_{cij}) \tag{8}$$

where j ranges over all lanes. The lane probabilities, ρ_{cij}, at any given approach are still computed with the same mean service time (similar to Equation 3).

CALCULATION OF DELAY

The revised queueing model contains two choices for calculating delay. First, an estimate of subject approach capacity could be obtained from the average service time as found from Equations 1 to 4, then delay could be calculated from the empirical delay relation in TRC 373. Second, it is possible to use Equation 6 as provided in the M/G/1 queueing model. Regrettably, Equation 6 requires knowledge of the variance of service time, which is no longer given by Equation 5.

Finding an exact expression of the variance of service time would be very difficult. However, it is possible to find a good approximation of the variance by discarding minor sources of variation. If the variation associated with turning and coordination in Equation 7 can be ignored, the variance in service time can be approximated by

$$\sigma^2 \approx (t_m - t_o\rho_c)^2(1 - \rho_c) + [t_m + t_z - t_c + t_o(1 - \rho_c)]^2\rho_c - s^2 \tag{9}$$

where t_o is mean time necessary for coordination between vehicles on subject and opposing approaches when there is a conflicting vehicle, and t_z is the maximum of the average headways on conflicting approaches when there is a conflicting vehicle. Note that t_m now applies only to the subject approach, as each approach can have a distinctly different average minimum headway. Because Equation 9 ignores many sources of variation, its results are assured to be smaller than an exact expression.

A quick inspection of Equation 6 shows that some error in the variance might be tolerable. The question arises: How much error exists in Equation 9 and what is the effect of this error on delay?

ESTIMATING ERROR IN APPROXIMATE VARIANCE

In absence of an exact expression for the variance of service time, a Monte Carlo simulation was used to calculate the correct variance. Separate sets of simulations were performed for one-by-one, one-by-two, and two-by-two intersections. The following procedure was used.

1. Generate randomly a set of movement volumes for an intersection;
2. Determine the average service time and the utilization ratios of each approach with Equations 1 to 3, 7, and 8;

TABLE 2 Comparison of M/G/1 Capacities with TRC 373 Capacities

Traffic Condition	M/G/1	TRC 373
Single lane; no turning; all approaches at capacity	500.	525.
Double-lane; no turning; all approaches at capacity	616.	625.
Single-lane; no turning; no conflicting traffic; subject and opposing approaches at capacity	935.	950.*
Double-lane; no turning; no conflicting traffic; subject and opposing approaches at capacity	1286.	1050.*
Single-lane; no turning; only subject approach has volume	1000.	1100.*
Double-lane; no turning; only subject approach has volume	1565.	1200.*
Single-lane; all approaches at capacity, 25% left turns	446.	375.
Single-lane; all approaches at capacity, 25% right turns	535.	650.

*Outside the stated range of TRC 373 capacity relation.

3. Stop if the capacity of any approach is exceeded running the M/G/1 model;
4. Determine the service time for 2,000 separate vehicle arrivals at the subject approach with a random set of vehicles at an intersection;
5. Calculate the sample mean and the sample variance of the service time;
6. Estimate the delay from the sample mean and variance; and
7. Calculate the delay from the Equation 9 and the computed information from Step 2.

This procedure was applied 10,000 times for each of the three intersection geometries. After eliminating unfeasible volumes in Step 3, there remained 55 cases for the one-by-one intersection, 90 cases for the one-by-two intersection, and 183 cases for the two-by-two intersection. Because the random set of vehicles for each trial in Step 4 were created with the probabilities obtained in Step 2, the results of the Monte Carlo simulation should be quite consistent with the analytical model.

The results for the two-by-two intersection are illustrated in Figures 1 and 2. The results for the other two intersections are similar. The anticipated underestimate of the variance is apparent for all intersections, but the effect on the calculation of delay is small.

When interpreting Figures 1 and 2 it is helpful to note that there is no consistent relation between variance and delay or between the error in variance and delay.

If desired, the small downward bias in delay could be removed by simply adding a correction factor to the Equation 9. The correction factor would be approximately 0.7 sec^2 for a two-by-two intersection, 0.5 sec^2 for a one-by-two intersection and 0.2 sec^2 for a one-by-one intersection. The delay values shown in Figure 2 are uncorrected. Because correction factors are dependent on the chosen set of parameters, they would need to be rederived with a Monte Carlo simulation each time a new set of parameters was selected.

DELAY COMPARISON

Figure 3 illustrates the difference between the TRC 373 delay relation and the M/G/1 queueing model. The random data

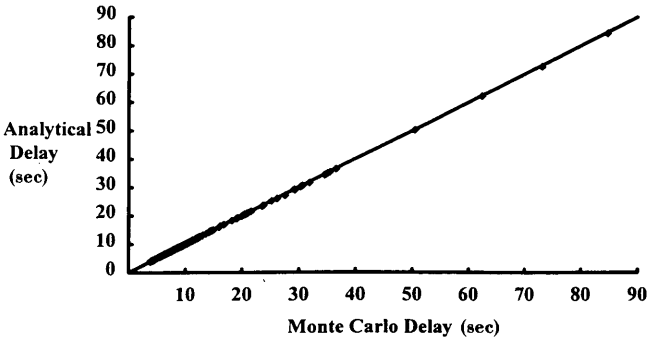


FIGURE 2 Comparison of analytical with Monte Carlo delay for two-by-two intersection.

are for the one-by-two intersection, but the other intersections look similar. The volume-to-capacity ratios were calculated from the queueing model. The two relations produce similar results for delays between 5 and 40 sec. The M/G/1 delay relation does not give any delays less than 4 sec, which is unreasonable. The TRC 373 delay relation did not estimate delays greater than 45 sec, even when the volume-to-capacity ratio approached 1.0.

QUICK COMPARISON WITH TRC 373

A comparison between the TRC 373 and the revised M/G/1 queueing models is given in Table 3. Clearly, the choice of model depends on the trade-off between ease of use versus range of applicability. TRC 373 would be preferred for cases in which hand calculation is required. The revised M/G/1 queueing model would be preferred for computerized assessment of capacity and delay.

CONCLUSIONS

Richardson's M/G/1 queueing model ignored certain aspects of driver behavior at AWSC intersections, which now are

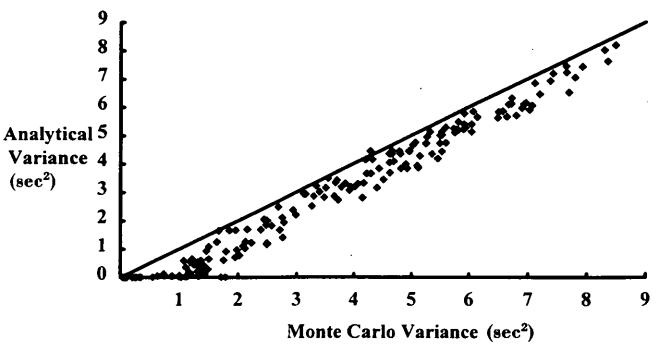


FIGURE 1 Comparison of analytical with Monte Carlo variance for two-by-two intersection.

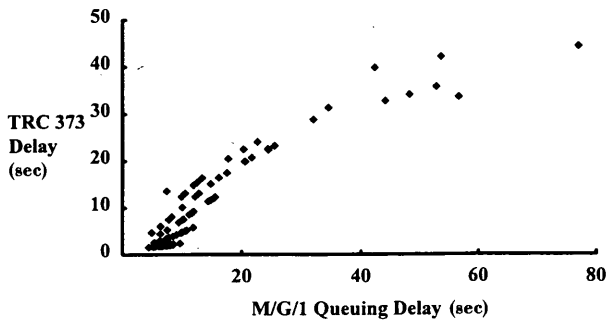


FIGURE 3 Comparison of TRC 373 delay relation with revised M/G/1 queueing model for one-by-two intersection.

TABLE 3 Summary of Model Characteristics

Attribute	TRC 373	Queuing Model
Type of model	Empirical	Theoretical
Ability to provide accurate estimates of capacity and delay over its range of applicability	Good	Good
Range of applicability	Many common situations	Most situations
Number of parameters	9	9
Physical interpretation of parameters	Some	All
Amount of required data	12	12
Hand calculation possible	Yes	No
Ease of understanding	Easy	Difficult

*Outside the stated range of TRC 373 capacity relation.

considered important. These aspects relate to turning and to coordination among drivers who get priority at the same time. A revised model can include a more comprehensive representation of driver behavior; however, it is not possible to obtain an exact formula for the variance of the service time, which is a key variable for the calculation of delay.

Fortunately, a good approximation for the variance of service time is available. The accuracy of the approximation can be verified by Monte Carlo simulation. Furthermore, it is possible to use the results of the Monte Carlo simulation to correct a bias in the approximation that tends to cause a slight underestimate of delay.

The revised queueing model can be made reasonably consistent with the relationships in TRC 373, and it applies to a wider variety of traffic conditions. The queueing model has the same number of parameters as the relations in TRC 373 and the queueing model's parameters are judged to have a stronger physical interpretation. Data requirements are identical.

The queueing model is recommended for computerized evaluation of AWSC intersections having a wide range of traffic conditions.

REFERENCES

1. M. Kyte. Estimating the Capacity of an All-Way Stop-Controlled Intersection. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 70–81.
2. M. Kyte and J. Marek. Estimating Capacity and Delay at a Single-Lane Approach All-Way Stop-Controlled Intersection. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 73–82.
3. *Transportation Research Circular 373: Interim Materials on Unsignalized Intersection Capacity*. TRB, National Research Council, Washington, D.C., July 1991.
4. A. J. Richardson. A Delay Model for Multiway Stop-Sign Intersections. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987, pp. 107–112.

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

Effect of Heavy Vehicles at Australian Traffic Circles and Unsignalized Intersections

R. J. TROUTBECK

The technique for estimating the capacity of traffic circles in Australia has employed a gap-acceptance approach. For this approach the influence of heavy vehicles was essentially ignored because the sites were typical and had about 5 percent heavy vehicles. The parameters used to quantify the gap-acceptance behavior of drivers had included the behavior of the heavy-vehicle drivers. The use of this technique has now been extended beyond a simple analytical process and is incorporated into the SIDRA analysis program. The consequences of including a high proportion of heavy vehicles needs to be estimated. The passenger car unit (pcu) technique for estimating the influence of heavy vehicles and another technique using three factors to modify the gap-acceptance parameters for drivers of heavy vehicles and drivers accepting a gap just ahead of a heavy vehicle are discussed. In both cases an analytical assessment was done. Values for these terms, which modify the gap-acceptance parameters, are evaluated from field studies and are reported. It was concluded that the pcu method is reasonable but that the modified gap acceptance approach provides greater understanding and flexibility. The values found will assist others in evaluating the influence of heavy vehicles at traffic circles and unsignalized intersections.

Drivers negotiate unsignalized intersections by accepting and rejecting gaps in opposing traffic. A driver is assumed to review the gaps in several streams as well as the queueing in some of the streams. The process is a difficult one and it is often simplified. When there is a substantial proportion of heavy vehicles at the unsignalized intersection, then capacity will be affected through changes in behavior. The behavior of drivers entering a traffic circle or an unsignalized intersection and the changes in behavior when there is a greater proportion of heavy vehicles are described in this paper.

Behavior parameters have been collected and analyzed at traffic circles, but the behavior is expected to be similar to that experienced at unsignalized intersections. The behavior at traffic circles is simple because drivers only need to give way to one or two circulating streams going in one direction. Hence they are very simple unsignalized intersections.

The numbers given in this paper are only seen to be indicative at this stage. More extensive data should be collected and studied using such intersection analysis tools as SIDRA (*1,2*).

R. J. Troutbeck, Queensland University of Technology, Two George Street, GPO Box 2434, Brisbane Q4001, Australia.

GAP-ACCEPTANCE BEHAVIOR

At unsignalized intersections, drivers are assumed to behave consistently and all drivers are assumed to behave identically. These assumptions, when applied together, have been found to give reasonable results (*3–5*). For practical purposes these assumptions will continue to be used.

Gap-acceptance behavior assumes that entering drivers accept gaps, and enter the unsignalized intersection, when the headway between the opposing vehicles is greater than the critical acceptance gap, T. It is further assumed that entering drivers will follow each other into the unsignalized intersection at headways equal to the follow on time, T_O, if there is a long headway between the opposing vehicles.

Gap-Acceptance Behavior Expected at Traffic Circles

Gap-acceptance behavior is expected to be similar at traffic circles for which the opposing traffic is the circulating traffic for each leg. Each entry leg is considered to be an independent T-intersection. The major stream, or the priority stream, is the circulating traffic that passes in front of the entering drivers. This circulating traffic excludes the drivers who exit at the same leg. Troutbeck (*6*) discusses the consequences of these assumptions.

The entering behavior of drivers was found to be dependent on the circulating or opposing flow and on the flows in the entry lanes at an approach. The greater the circulating flow the shorter are the gap-acceptance parameters—that is, the shorter are the gaps that are acceptable to the entering drivers (*7*). Although this cannot be quantified or proven by the data directly, it is hypothesized that as the speeds of the circulating vehicles are decreased as a result of greater circulating flows, drivers then feel more comfortable in accepting shorter gaps, and some circulating drivers will yield right of way to the entering drivers. This would result in the observed gap-acceptance parameters decreasing with higher circulating flows. This could result from drivers either changing their behavior when they are delayed (*8*) or sizing up the intersection before they approach, leading to lower gap-acceptance characteristics at the higher flow sites. The author favors the latter because an earlier study of overtaking behavior did not find a relationship between delay and gap-acceptance behavior (*9*). The outcome is the same, however.

Gap-Acceptance Behavior Expected at Unsignalized Intersections

A similar relationship between the speed of opposing vehicles and driver behavior has been observed at two-way stop or give-way intersections. Brilon (*10*) reported that Harders (*11*) found that gap-acceptance parameters were influenced by the speed of opposing vehicles. This speed influence has been incorporated into the *Highway Capacity Manual* (HCM), Table 10-2 (*12*).

Similarly, there are likely to be differences in the behavior of drivers in adjacent slip lanes if there is uncertainty in the path drivers will take after the slip lane. This behavior pattern is likely to be similar to that observed at entries to Australian traffic circles (*7*).

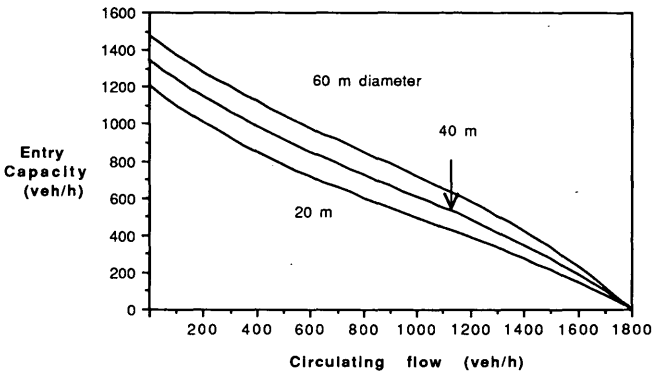


FIGURE 2 Entry capacity of single-lane traffic circles with different inscribed diameters.

USING PASSENGER CAR EQUIVALENTS FOR HEAVY VEHICLES

The influence of heavy vehicles at unsignalized intersections and traffic circles can be evaluated using a heavy-vehicle equivalent as in the 1985 HCM. The influence of heavy vehicles will be evaluated using analytical techniques to estimate the performance of Australian traffic circles. These techniques are described by Troutbeck (*7*) and are now included in Australian practice (*13*). Similar trends are expected for other forms of unsignalized intersections.

The base case will be a traffic circle with a single circulating lane, a maximum diameter between the islands at the entries (inscribed diameter) of 40 m, and a 4-m entry lane. A single-lane traffic circle is used in this analysis, although the trends are the same at larger traffic circles and at traffic circles with more than one lane. The dimensions of the traffic circle are shown in Figure 1, and the entry capacity is shown in Figure 2. In Australian guides the capacity of a traffic circle is the maximum flow that can be expected at that entry, if all other flows remain the same.

The traffic at the traffic circle sites used in the analysis of driver behavior (*7*) included about 5 percent trucks. In AustRoads (*13*), the effect of trucks was not considered necessary if there were less than 5 percent trucks. If the passenger car equivalent of heavy vehicles is changed, then the predicted entry capacity would change. Under these conditions the traffic composition factor to convert between vehicles and passenger car units (pcu) is given by

$$f_c = 1 + (p_{HV} - 0.05)(e_{HV} - 1) \quad (1)$$

where

p_{HV} = proportion of heavy vehicles,
$p_{HV} > 0.05$, and
e_{HV} = passenger car equivalent of a heavy vehicle.

Passenger car equivalents between 1.5 and 2 are acceptable values for a heavy vehicle (*12,13*).

Using Equation 1, an increase in the proportion of heavy vehicles in the circulating stream increases the pcu value of the circulating flow. This then results in a decrease in capacity. On the other hand, an increase in the proportion of heavy vehicles in the entry stream increases the traffic composition factor and reduces the arrival capacity in units of vehicles per hour (veh/hr). The effect of changing the proportion of heavy vehicles in both lanes are additive. The combined effect is to lower the predicted capacity by about the same amount for all circulating flows.

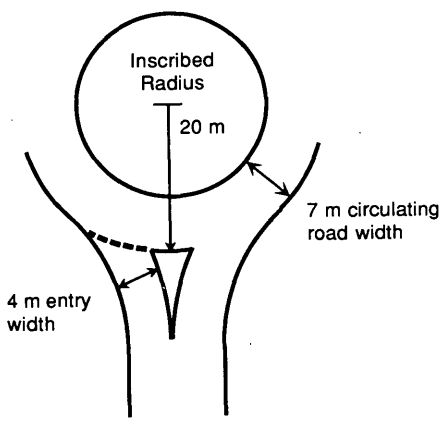


FIGURE 1 Dimensions of single-lane traffic circles.

GAP-ACCEPTANCE PARAMETERS FOR HEAVY VEHICLES

Instead of using passenger car equivalents to factor the flows to account for heavy vehicles, it is preferable to use a gap-acceptance approach. The gap-acceptance approach describes driver behavior at these intersections. This approach gives an increased understanding of the mechanisms used and is consistent with the analysis technique used by AustRoads (*13*). Entry capacity can be influenced by heavy vehicles in two ways. Drivers of heavy vehicles can display different gap-acceptance characteristics when entering a traffic circle or unsignalized intersection. Drivers may wish to accept longer gaps when entering a traffic circle in front of heavy vehicles. Both aspects are discussed in the following sections.

Different Entering Driver Behavior

As has been discussed by Harders (*14*), Catchpole and Plank (*3*), and Bennett (*15*), the entry capacity of a stream of drivers that are not homogeneous is given by harmonic means—that is, the capacity of an entry, Q, is given by

$$\frac{1}{Q} = \frac{\text{prop}_1}{Q_1} + \frac{\text{prop}_2}{Q_2} + \ldots + \frac{\text{prop}_n}{Q_n} \quad (2)$$

where Q_i is the entry capacity of the ith element of the entry stream and prop_1 is the proportion of the ith element in the entry stream.

If it is assumed for simplicity that both the critical acceptance gap and the follow-on time are increased by the same gap-acceptance parameter factor, f_{GA}, then

$$T_{HV} = f_{GA} T_{LV} \quad (3)$$

and

$$T_{0HV} = f_{GA} T_{0LV} \quad (4)$$

where T_{HV} and T_{LV} are the critical acceptance gaps for heavy vehicles and cars and T_{0HV} and T_{0LV} are the follow-on times for heavy vehicles and cars.

Heavy-vehicle drivers can be expected to require a longer gap when entering a traffic circle because their vehicles are longer and have generally lower acceleration rates (*16*). Capacity will be reduced, as shown in Figure 3. It is interesting to note that Harders (*11*) found that the gap-acceptance parameter factor for the critical acceptance gap differs from that for the follow-up time.

The results shown in Figure 3 can be used to give an equivalent pcu value for the heavy vehicles. For instance, at a circulating flow of 1,000 veh/hr, the entry capacity for cars is 615 veh/hr. If the critical acceptance parameters are 50 percent longer (gap-acceptance parameter factor = 1.5), then the entry capacity is reduced to 317 veh/hr. This implies an equivalent pcu of 1.94. In fact, a gap-acceptance parameter factor of 1.5 gives similar results to those obtained when a pcu of 2 for heavy vehicles in the entry streams for circulating flows less than 1,500 veh/hr is assumed.

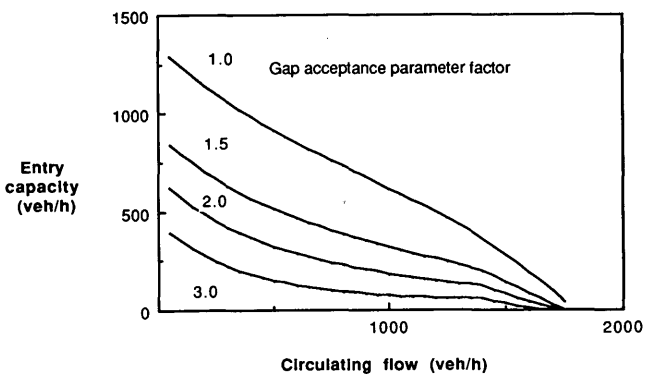


FIGURE 3 Entry capacities for drivers of heavy vehicles that have different gap-acceptance parameter factor values.

If a gap-acceptance factor of 1.5 is adopted, then the entry capacities change with the proportion of heavy vehicles. For less than 20 percent trucks, the reduction in the entry capacity is less than about 12 percent below the capacity estimated when the gap-acceptance properties of heavy vehicle drivers are assumed to be the same as cars.

Effect of Circulating Vehicle Type on Gap-Acceptance Behavior

The effect of drivers having different critical gap-acceptance characteristics that depend on the major stream vehicle types can be estimated if it is assumed that

Q_{LV} = entry capacity when major stream consists of cars (veh/sec),

Q_{HV} = entry capacity when major stream consists of heavy vehicles (veh/sec),

prop_{HV} = proportion of heavy vehicles in major stream, and

q = major stream flow (veh/sec).

Again, the major stream is the circulating stream. Assuming that there is no correlation between gap size and vehicle type, then in a given time period, H, $qH\text{prop}_{HV}$ heavy vehicles and $qH(1 - \text{prop}_{HV})$ cars will pass. For the purposes of estimating capacity, this is equivalent to a major stream consisting of a stream of cars for a period of $H(1 - \text{prop}_{HV})$ followed by a stream of trucks for a period of $H\ \text{prop}_{HV}$. The number of vehicles that can enter in each period is then $H(1 - \text{prop}_{HV})$ Q_{LV} and $H\ \text{prop}_{HV}\ Q_{HV}$. The total entry capacity, Q, is then

$$Q = \text{prop}_{HV}\ Q_{HV} + (1 - \text{prop}_{HV})\ Q_{LV} \quad (5)$$

Because heavy vehicles are generally longer than cars, it is expected that heavy vehicles present a longer "block" to the entering drivers. If it is assumed that an entering driver merges into the circulating stream so that there is an acceptable minimum headway between the circulating and entering vehicles, then a longer headway will be required between a heavy circulating vehicle and an entering vehicle. Similarly, a circulating stream that consists of only heavy vehicles would have a lower maximum flow than if the stream consisted only of cars. This implies that there should be a longer minimum headway between heavy vehicles. This interaction between the entering driver and different circulating drivers can be modeled using the following notation:

T = critical acceptance gap for drivers accepting a gap before a car;

T_0 = follow-on time for drivers accepting a gap before either a car or a heavy vehicle;

Δ = minimum headway in front of cars (all cars that are following can be assumed to have this minimum headway);

Δ_{HV} = minimum headway in front of heavy vehicles (similarly, all heavy vehicles that are following can be assumed to have this minimum headway); and

$T - \Delta + \Delta_{HV}$ = critical acceptance gap for drivers accepting a gap before a heavy vehicle.

For convenience, Δ_{HV} is considered to be a multiple of Δ. That is,

$$\Delta_{HV} = f\Delta \quad (6)$$

where f is a heavy vehicle headway factor that is greater than 1. For all practical purposes, this factor may be assumed to be equal to the ratio of the average length of heavy vehicles to cars. This factor affects the distribution of gaps in the circulating stream. It is assumed that there are $1 - \alpha$ vehicles that are following others (in bunches). Then there will be $(1 - \alpha)$ $prop_{HV}$ heavy vehicles with a headway of Δ_{HV}, and $(1 - \alpha)$ $(1 - prop_{HV})$ cars with a headway of Δ, and the remainder will have a headway greater than Δ_{HV}. The term, α, is the proportion of nonbunched or free vehicles.

Increasing the heavy-vehicle headway factor reduces the predicted capacity, but only by a few vehicles per hour. It is expected that an f-value of 1.5 gives reasonable, but unsubstantiated, results. If the proportion of heavy vehicles is greater than 10 percent, and assuming an f-value of 1.5, the error is roughly dependent on the proportion of heavy vehicles.

Another influence is that drivers may perform differently when accepting gaps in front of heavy vehicles than when accepting gaps in front of cars. The critical gap for drivers accepting a gap in front of heavy vehicles will be greater than when accepting a gap in front of cars. For convenience, the critical acceptance gap is assumed to be t_f times greater than the critical acceptance gap for cars. The follow-on times for the entering drivers are assumed to be the same when the gap is terminated by either a car or a truck as the extra time has been accounted for in the critical acceptance gap. That is,

$$T_{CHV} = t_f T_{CLV} \quad (7)$$

and

$$T_{0CHV} = T_{0CLV} \quad (8)$$

where the subscript C refers to the circulating stream and t_f is the truck factor. Changing the truck factor from 1.0 to 2.0 has only a marginal effect on the estimate of capacity.

The effect of the headway factor and the truck factor are similar, and either factor could be used to explain the influence of heavy vehicles in the opposing (circulating) streams. Their effect can then be combined.

Combined Effect of Heavy Vehicles in Circulating and Entry Streams

The influence of heavy vehicles at traffic circles described includes the following three effects:

- Heavy vehicles have slower accelerations and require a longer critical gap,
- Heavy vehicles have a greater occupancy in the circulating stream on the road system and therefore reduce the available time for merging, and
- Heavy vehicles cause entering drivers to adjust their behavior.

The combined effect of the gap-acceptance parameter factor, the truck factor, and the heavy-vehicle headway factor are additive. Figure 4 shows this combined effect and the effect of each factor when there are 20 percent trucks on both the entry and circulating streams. All factors should be incorporated into the model, although the gap-acceptance factor has the greatest influence and predominates.

The reduction in capacity from both effects resulting from an increase in heavy vehicle percentages is about 60 veh/hr for each 10 percent increase in the proportion of heavy vehicles. The likely overestimation from ignoring the presence of heavy vehicles is less than 10 percent, given that the proportion of heavy vehicles is less than 15 percent and that the circulating flow is less than 1,500 veh/hr.

TRUCK FACTOR AND GAP-ACCEPTANCE FACTOR

Evaluation

A suburban traffic circle on the Breakfast Creek and Montpelier roads in Brisbane was used for a field study to evaluate the gap-acceptance parameters to be evaluated for each variable type. This site was chosen because it had about 25 percent trucks in both the entry and the circulating streams.

The data collected consisted of the type of entering vehicle and the gaps that were rejected or accepted. Data on the type of vehicle at the end of the gap were also collected. This enabled the following data to be compiled into a data table:

- The size of the accepted gap if the gap was terminated by the presence of a car,
- The size of the accepted gap if the gap was terminated by a heavy vehicle,

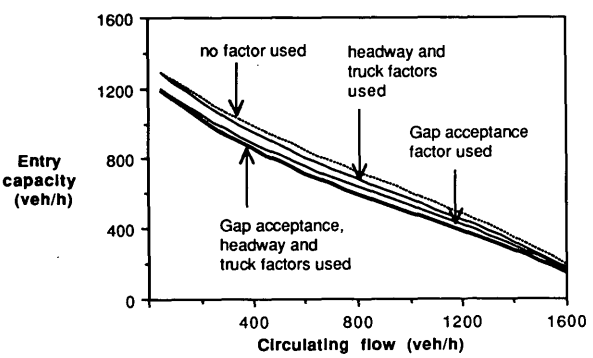


FIGURE 4 Predicted entry capacity when gap-acceptance parameter factor is applied, when heavy-vehicle headway and truck factor is applied, when no factors are used, and when all factors are applied (20 percent trucks assumed).

- The largest rejected gap in front of a heavy vehicle, and
- The largest rejected gap in front of a car.

The data were analyzed using a modified gap-acceptance method. For a conventional gap-acceptance method, the largest rejected gap and the accepted gap are used to predict the parameters for the critical gap distribution given that the form of the critical gap distribution is specified. The lognormal function has been found to be acceptable (*17*) as it has nonzero values and is skewed to the right. It is used here.

The conventional maximum likelihood technique requires that likelihood of n drivers having a largest rejected gap of x_i and an accepted gap of y_i subject to the condition that $x_i \leq y_i$, be estimated. If $y_i < x_i$ then the driver was assumed to be inattentive and x_i was set just below y_i. The likelihood is then

$$\prod_{i=1}^{n} [F(y_i) - F(x_i)]$$

The logarithm of the likelihood is then

$$L = \sum_{i=1}^{n} \ln[F(y_i) - F(x_i)] \quad (9)$$

The maximum likelihood estimators for the critical gap distribution are the mean of the logarithms of the critical gaps, μ, and the variance of the logarithms of the critical gaps, σ^2. These are given by differentiating Equation 9 with respect to both μ and σ^2. This technique is further explained in Troutbeck (*17*).

In this case the effect of trucks in the circulating stream was accounted for by assuming that a particular entering driver would accept a gap of size β in front of a car but would require a gap of βt_f in front of a heavy vehicle. The solution was obtained by finding a value of t_f that maximized the likelihood. Figure 5 illustrates the effect of the truck factor on the likelihood function for cars and trucks. The normalized values are factored so that they are 1 for truck factors of 1.0.

As the truck factor is increased above 1, the normalized logarithm of the likelihood functions moves toward a local minimum. The curves are quite shallow, and a range of values could be suitable. This graph shows that truck factors of 1.64 for trucks and 1.23 for cars are suitable. Similarly, the mean of the critical gap distribution or the critical acceptance gap is also a function of the normalized logarithm of the likelihood function. For truck drivers the critical acceptance gap is 5.13 sec; for cars, it is 5.37 sec. This is shown in Figure 6.

Figures 5 and 6 imply that minimum gap required for a car driver to enter on front of a car is 5.37 sec and 5.37 × 1.23 or 6.6 sec when entering in front of a truck. A driver of heavy vehicle would require a gap of 5.14 sec when entering in front of a car and 8.4 sec when entering in front of a heavy vehicle.

It is not impossible for a truck driver to accept a shorter gap than a car driver, particularly if the heavy-vehicle driver assumes that might is right, but it appears to be more acceptable to consider that drivers of both vehicle types have the same critical gap when accepting a gap in front of a car. Figure 6 shows that if the critical acceptance gap for heavy vehicles was increased to 5.37 to match the value for cars, then the increase in the likelihood factor is marginal. The appropriate truck factor at this critical acceptance gap is reduced to 1.37. Truck factors are now 1.23 and 1.37 for cars and trucks entering the stream. This then leads to the critical acceptance gaps given in Table 1.

There are good reasons that the truck factor differs for car drivers and heavy-vehicle drivers. Car drivers often do not fully appreciate the performance of a heavy vehicle and will often allow less-than-desirable road space for heavy vehicles. Heavy-vehicle drivers would be more considerate in this regard. Consequently, the different truck factors for different vehicle types is appropriate.

Distribution of Critical Gaps

The critical acceptance gap used in the preceding is the mean critical gap derived from the maximum likelihood process as described. The distribution of critical gaps is derived from the following data using the maximum likelihood process.

The distribution of critical gaps for drivers of cars and heavy vehicles that accept gaps in front of cars are very similar because of the revision made before. The distribution of critical gaps for drivers accepting a gap in front of a heavy vehicle are shifted to the right and have a greater skew (Figure 7). Many heavy-vehicle drivers would be expected to accept gaps shorter than those expected by drivers of cars. Similarly, many

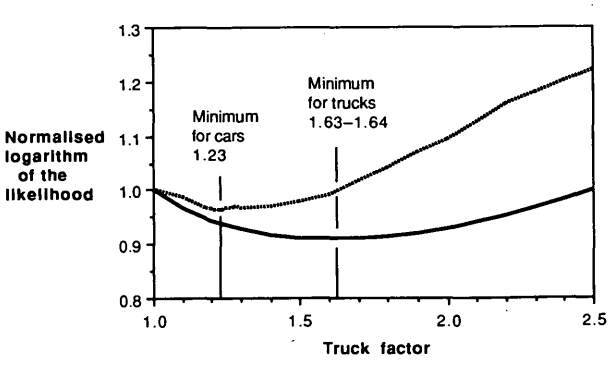


FIGURE 5 Effect of truck factor on normalized logarithm of likelihood.

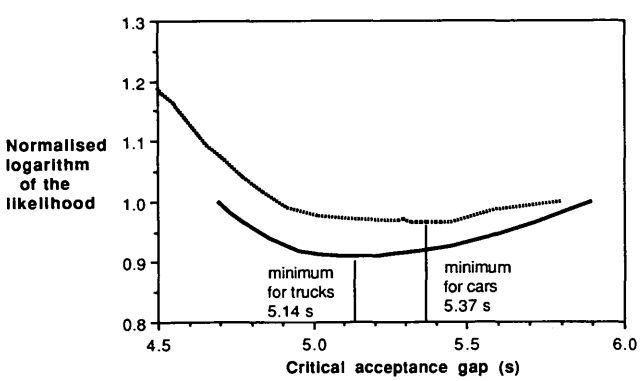


FIGURE 6 Effect of critical acceptance gap on normalized logarithm of likelihood.

TABLE 1 Critical Gap Distribution Parameters

Parameter	Entering vehicle type	
	Car	Heavy vehicle
Critical acceptance gaps for accepted gaps terminated by a car	5.37	5.37
Mean of the logarithms of drivers' individual critical gaps terminated by a car	1.557	1.586
Standard deviation of the logarithms of drivers' individual critical gaps terminated by a car	0.498	0.438
Truck factor	1.23	1.37
Critical acceptance gaps for accepted gaps terminated by a heavy vehicle	6.60	7.36
Mean of the logarithms of drivers' individual critical gaps terminated by a heavy vehicle	1.763	1.900
Standard deviation of the logarithms of drivers' individual critical gaps terminated by a heavy vehicle	0.498	0.438

drivers would accept a gap in front of a heavy vehicle that is shorter than the gaps accepted, in front of an oncoming car, by some car drivers. These drivers may believe that they can accelerate and move ahead of the slower-moving heavy vehicles in the circulating stream.

Whereas the differences in the critical gap distribution are not excessive, the results do indicate likely changes in drivers' gap-acceptance capabilities that could be expected at unsignalized intersections. From the analysis presented here, the gap-acceptance parameter that relates the mean critical gap for car drivers to truck drivers should be set to 1.0. Different truck factors of 1.23 and 1.37 were found for different entering vehicle types of cars and trucks, respectively.

Using these truck factors and the gap-acceptance factor, the influence of the proportion of trucks at the intersection affects the entry capacity only marginally. The reduction in capacity when the proportion of trucks is increased from 5 to 40 percent is still less than 10 percent for circulating flows of less than 1,500 veh/hr.

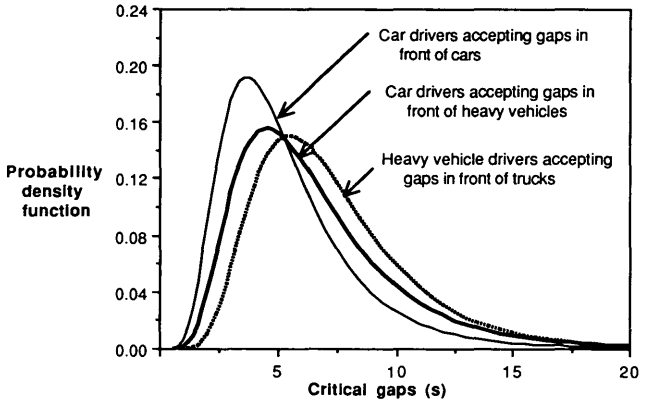


FIGURE 7 Probability density function of critical gaps of drivers of cars and heavy vehicles accepting a gap in front of a heavy vehicle.

DISCUSSION OF RESULTS

The gap-acceptance effect on capacities is similar to the effect estimated using pcu values in that an increase in the proportion of heavy vehicles results in a decreased capacity. These two techniques are different and produce different results. Both techniques have applications. The pcu method is simpler to understand but lacks the explanatory properties of the gap-acceptance approach. The latter approach is expected to offer a better behavioral description of the processes involved in the gap-acceptance behavior at traffic circles and unsignalized intersections.

Effective pcu Values

If a pcu method is used to account for the influence of heavy vehicles, then a pcu value close to 1.5 gives a good representation of the modified gap-acceptance relationships over most of the range of circulating flows. It is assumed here that the truck factors will be the same for all circulating flows. The pcu value is very similar to the one used in the HCM (*12*).

As the opposing flow increases, the pcu equivalent is reduced. This is reasonable at traffic circles because the circulating flow influences the speed. As the circulating flow increases, so the gap-acceptance terms decrease. There could also be fewer differences between the influence of oncoming (circulating) cars and trucks at these slower circulating speeds. This would imply a lower pcu value. The truck factors are only indicative, yet they do provide an estimate of the magnitude of the suitable parameters.

The observations of drivers entering traffic circles included different driver and vehicle types. The critical gap parameters evaluated from these observations are expected to account for some, if not most, of these driver differences. The drivers that accept longer gaps in front of heavy vehicles or the drivers of the heavy vehicles would be in the tail of the statistical distribution for the gap-acceptance parameters. It is therefore often convenient to leave out the effects of these heavy vehicles.

Application of Results to Other Unsignalized Intersections

Currently, many techniques [e.g., those of the HCM (*12*) and Kimber (*18*)] use a pcu method of evaluating the effect of trucks at unsignalized intersections. This is considered satisfactory for the most cases, but it does not give a full behavioral explanation of the influence of heavy vehicles.

The gap-acceptance factors explained in this paper are able to provide a better behavioral description and can be used to evaluate the performance of unsignalized intersections. The notion of a truck factor can be used at all unsignalized intersections, except that the factors could be expected to change as opposing flow increases or as speed increases. Further work to identify the appropriate factors that should be used at all types of unsignalized intersections and under all conditions is needed.

In summary, the concepts of the gap-acceptance factors described in this paper can be applied to all types of unsignalized intersections.

RECOMMENDATIONS AND CONCLUSIONS

The sensitivity of techniques and parameters used to account for the effect of heavy vehicles at traffic circles has been addressed in this paper. It was pointed out that similar effects are likely to be experienced at other unsignalized intersections. Field data were used to evaluate the likely influence and magnitude of the theoretical parameters that describe the effect of heavy vehicles. Major conclusions from this preliminary study are

- The pcu method of adjustment is seen to be a satisfactory alternative to adjust for the proportion of heavy vehicles. This technique is a simple one and is easily incorporated into a procedure.
- The gap-acceptance approach is more descriptive and is able to offer a driver behavioral interpretation on the effect of heavy vehicles.
- The gap-acceptance parameter factor, which is the ratio of the critical acceptance gap for drivers of cars to the critical acceptance gap for drivers of heavy vehicles, is the term that has the greatest influence on the estimate of capacity. However, in the limited field studies, this factor was found to be equal to 1.0. It was therefore discarded.
- The truck factors relate the critical gap distribution to the vehicle type that terminates the gap. There were different values for the different entering vehicle types, which was thought to result from the drivers' attitudes toward heavy vehicles.
- Pcu values of 1.5 applied to the circulating stream only give similar results to the modified gap-acceptance technique that used the truck factors identified in this report. As circulating factors increase, the pcu equivalents are likely to decrease.

REFERENCES

1. R. Akçelik. *Introduction to SIDRA-2 for Signal Design*. Research Report 148. Australian Road Research Board, Victoria, 1987.
2. R. Akçelik. *Calibrating SIDRA*. Research Report 180. Australian Road Research Board, Victoria, 1990.
3. E. A. Catchpole and A. W. Plank. The Capacity of a Priority Intersection. *Transportation Research B*. Vol. 20B, No. 6, 1986, pp. 441–456.
4. R. J. Troutbeck. Current and Future Australian Practices for the Design of Unsignalised Intersections. In *Intersections Without Traffic Signals* (W. Brilon, ed.), Springer-Verlag, Berlin, Germany, 1988, pp. 1–19.
5. H. Wegmann. A General Capacity Formula for Unsignalised Intersections. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, Berlin, Germany, 1991, pp. 177–191.
6. R. J. Troutbeck. Traffic Interactions at Roundabouts. *Proc., 16th Australian Road Research Board Conference*, Vol. 16, No. 5, 1990, pp. 17–41.
7. R. J. Troutbeck. *Evaluating the Performance of a Roundabout*. Special Report 45, Australian Road Research Board, Victoria, 1989.
8. W. Kittelson and M. Vandehey. Delay Effects of Driver Gap Acceptance Characteristics at Two-Way Stop-Controlled Intersections. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991.
9. R. J. Troutbeck. *Overtaking Behaviour on Australian Two-Lane Rural Highways*. Special Report 20, Australian Road Research Board, Victoria, 1981.
10. W. Brilon. Recent Developments in Calculation Methods for Unsignalised Intersections in West Germany. In *Intersections Without Traffic Signals* (W. Brilon, ed.), Springer-Verlag, Berlin, Germany, 1968, pp. 111–153.
11. J. Harders. Grenz- und Folgezeitlücken als Grundlage für die Leistungsfähigkeit von Landstassen. (Critical Gaps and Move-Up Times as a Basis of Capacity Calculations for Rural Roads.) Schrifteneihe Strassenbau und Strassenverkehrstecknik, Heft 216, 1976.
12. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
13. *Guide to Traffic Engineering Practice. Part 6—Roundabouts*. AustRoads, Sydney, Australia, 1993.
14. J. Harders. Die Leistungfähigkeit nicht signalgeregelter städtischer Verkehrsknoten. (*The Capacity of Unsignalized Urban Intersections*.) Schriftenreihe Strassenbau und Strassenverkehrstechnik, Heft 76, 1968.
15. D. W. Bennett. Elementary Traffic Flow Theory and Applications. In *Traffic Engineering Practice*, 3rd ed. (K. W. Ogden and D. W. Bennett, eds., Monash University, Melbourne, Australia, 1984.
16. J. R. Jarvis. In-Service Vehicle Performance. *Proc., Society of Automobile Engineers/Australian Road Research Board 2nd Conference on Traffic Energy and Emissions*, Melbourne, 1968, pp. 21.1–21.15.
17. R. J. Troutbeck. *Estimating the Critical Acceptance Gap from Traffic Movements*. Physical Infrastructure Centre Report 92-5. Queensland University of Technology, Australia, 1992.
18. R. M. Kimber. *The Traffic Capacity of Roundabouts*. Laboratory Report 942. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1980.

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

Capacity and Design of Traffic Circles in Germany

WERNER BRILON AND BIRGIT STUWE

Capacity and traffic safety of traffic circles in Germany have been investigated in comprehensive research projects carried out at the Ruhr University Bochum. On the basis of observations of real traffic conditions, formulas for predicting traffic circle capacity under conditions in Germany have been developed. These formulas have been specified in terms of the number of lanes in the circle and in the entry. The capacities of traffic circles measured in Germany appear to be considerably less than the values predicted by English formulas. For German traffic circles, capacity has been measured at between 0.7 and 0.8 of English values. However, there is good agreement between calculations for French and German capacity. The influence of geometric parameters was also investigated. In addition to the number of lanes in the circle and in the entry, geometric parameters turned out to be significant for estimating the capacity of traffic circles. These parameters are inscribed circle diameter, number of traffic circle arms, and distance between exit and entry of the observed arm. Traffic circles in Germany also provide a considerable increase in traffic safety when compared with normal signalized or unsignalized junctions. Before and after investigations at small, single-lane traffic circles showed that traffic circles have a lower severity of accidents, whereas the number of accidents stayed almost the same. Experience with existing traffic circles resulted in some significant recommendations for the design of traffic circles. Using the suggested design criteria, good experience has been shown in several cases and has led to a rapid increase in the number of new single-lane traffic circles. Public opinion also favors traffic circles.

Traffic circles in Germany have a tradition from the beginning of this century. Therefore, a number of traffic circles from the 1950s and early 1960s are still in operation. During the 1970s this type of intersection was completely abandoned. Traffic engineers believed that the benefits of traffic circles were not too good. In recent years, because of international experience and overwhelmingly good results of new experiments, traffic circles with a modern design are again the center of interest to traffic engineers. Therefore, new traffic circles in Germany are rapidly increasing. Many new constructions or conversions of cross junctions into traffic circles have already been carried out. The experience with both new and older types of construction during several years of practice and research are discussed in this paper.

CAPACITY

Experts around the world agree that traffic circles without traffic signals can be operated usefully only if the circulating lane has the right of way and the arriving traffic must yield. Thus, a traffic circle can be regarded as a series of T-shaped entries into a one-way circular road. This makes it possible to apply the classical priority theory (*1*). For traffic circles, however, the British regression theory appears to be more useful (*2,3*). This method ascertains capacity formulas by observing entries to traffic circles during times of continual congestion. The entering traffic and the traffic within the traffic circle are counted in short time intervals (e.g., 1 min). The measuring points are compared and represented by a regression line. Because there is continual congestion in the entry during the entire measured time, every suitable gap in the main stream is used. Therefore, the observed values are also the capacity in the observed situation.

An objection to this method is that a wide scatter of points cannot be represented precisely enough by a single line. This is incorrect. Half of the mean variation of the points results from the statistical characteristics of the 1-min interval measurements. As a result of the regression technique, the actual capacity in longer time intervals (e.g., 1 hr) is very reliably determined. This was confirmed by simulation (*4*). The other half of the mean variation results from individual driver behavior and local peculiarities of the investigated traffic circles.

In a recent study (*5*), such measurements and evaluations were carried out for 11 traffic circles in Germany. Compared with a straight line as a regression equation for the representation of measurements, a regression function according to Siegloch's equation has proven to be slightly better:

$$q_{e,\max} = A \cdot e^{-B/10{,}000 \cdot q_c} \quad (1)$$

where $q_{e,\max}$ is maximum possible traffic volume of the entry in passenger car units (pcu) per hour, and q_c is traffic volume in the traffic circle at the entry in pcu per hour.

The constants A and B in this equation have been determined separately from the measurements by regression calculation for different types of entries. The results are given in Table 1 and shown in Figure 1, which indicate that the results depend on the number of lanes in the circle and the entries. For conversion into pcu, single-unit trucks were rated as 1.5 pcu, trucks and trailers as 2 pcu, and motorbikes as 0.5 pcu (*6*).

The reliability of these values varies for different configurations. For two-lane traffic circles with two-lane entries, a sample of several thousand measuring points with large variations of q_c was developed. These equations are very reliable. Among the single-lane traffic circles, only three could be found to fulfill the conditions (considerable congestion for a long time). Only one three-lane traffic circle was available. There

Ruhr-Universität Bochum, Lehrstuhl für Verkehrswesen, Universitätsstrasse 150, 4630 Bochum, Germany.

TABLE 1 Calculating Traffic Circle Capacity—Parameters *A* and *B*

Number of lanes Circular roadway	Entry	A	B	Number of measurements
3	2	2018	6.68	295
2	2	1577	6.61	4574
2-3	1	1300	8.60	867
1	1	1226	10.77	1060

are no traffic circles with more than three circulating lanes in Germany. Further investigation is necessary, particularly for single-lane traffic circles. A corresponding research project has meanwhile been initiated by the Federal Minister of Transport in Germany.

Comparing the corresponding formulas of the most frequent case of single-lane traffic circles from other countries shows that the postulated capacities are rather close to the lower limit, and therefore on the safe side. A similar comparison for two-lane traffic circles is given by Brilon (*3*). It is remarkable that a doubling of lanes (from single-lane to two-lane) does by no means lead to a doubling capacity. The capacity increase is at best about 30 to 40 percent, because the left lane of the entry is only rarely used in Germany [see Figure 12 in the work by Brilon (*3*)]. This behavior is caused by the drivers' fear of being unable to reach the desired exit from the left circulating lane. This was also found by Troutbeck (*7*). Therefore, an extension of traffic circle measurements to more than two lanes is not expected to cause a considerable capacity increase in Germany. Troutbeck also found that capacity increased by 70 to 80 percent when the number of lanes was doubled.

In foreign countries, many investigations have shown that besides the number of lanes, other characteristics of geometric design influence capacity as well (*2,7,8*). In the data of Brilon and Stuwe (*9*), influences other than the number of lanes in the circle and the entry could not be determined. But in recent investigations, during which the volume of data was increased considerably, some geometric parameters turned out to be of significant influence (*5*). These were, in addition to the number of lanes in the entry and the circle, the inscribed circle diameter (D), the number of traffic circle arms (A), and the distance between exit and entry of the observed arm (EM). The differences resulting from these when compared with the earlier single variable equation are important. Nevertheless, these results are preliminary.

A further point for discussion is whether the traffic leaving the traffic circle at an exit has an influence on the next entry (*2*). An influence of this kind could not be ascertained using German data. Simon (*10*) found the same results from data for Switzerland. Both questions, however, will be investigated further on the basis of a larger sample.

The formulas given in Equation 1 should not be overinterpreted. If the resulting capacity is 600 pcu/hr, this can mean that in fact values of 630 or 570 pcu/hr can occur as well. Therefore, the results of the equation should be regarded as a realistic estimate of the magnitude of capacity. For single-lane traffic circles, the following problem can arise from the equations: even with larger q_c values, L is greater than 0. Alternatively, it is obvious that downstream of the entry on a single-lane traffic circle, no more than about 1,800 pcu/hr can pass the traffic circle. Therefore, the curve has to be broken off pragmatically, as shown in Figure 1. Theoretically, this should also be considered for traffic circles with more than one lane per direction. This question, however, is less relevant for these cases than for a single-lane traffic circle, because situations with such traffic loads hardly ever occur.

The calculated capacity should not be directly used in practice as limit of traffic load, because it is associated with considerable congestion. Therefore, to achieve an acceptable traffic quality, the limits for practical use have to be lower than the values obtained from Equation 1. The following values result from the theory of unsignalized intersections (*6*): the mean delay is less than 40 sec/vehicle if a capacity reserve of 100 pcu/hr is subtracted according to Equation 1 [i.e., the practical usable limit of capacity, L, is 100 (*11*)]. This is a considerable simplification. For a closer look, the mean delays can be calculated according to the universal delay formula by Kimber and Hollis (*12*). Their formula takes into account the effect resulting from actual temporal changes in traffic volume. All

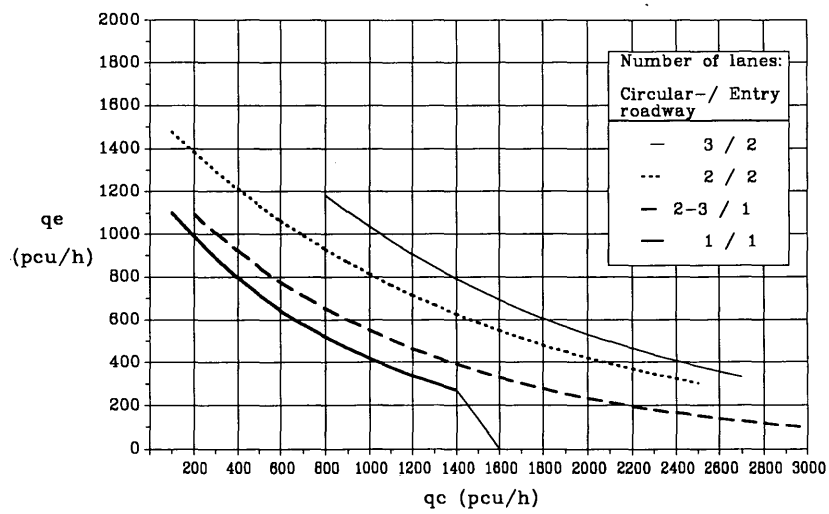


FIGURE 1 Regression curves for determination of traffic circle capacity.

calculations can be made with a multilingual computer program provided by Brilon and Wu (*13*).

The results of these calculations for a large number of cases make it possible to reach the general conclusion regarding capacity that is given in Table 2. In Table 2, ADT means the average total traffic passing the traffic circle per day. If the left-turning streams are very strong, capacity can be less than the minimum, M. The upper limit of capacity, U, may be exceeded only if the share of right-turning streams is very large or if extreme congestion is accepted. If the ADT values are between M and U, the capacity should be calculated using the previously introduced technique. These statements are valid for traffic circles with four or more legs. For three-legged traffic circles, the limits for the overall load can also be lower.

Of course, a comparison of capacity of traffic circles with other intersection types is quite interesting. First of all, it can be stated that traffic circle capacity will always surpass that of a four-way intersection or T-junction without traffic signals. Therefore, the real competitors are traffic circles and signalized intersections. If the signals are two-phase-controlled, they can be more favorable than the traffic circle. A four-phase-controlled signal, however, which is very popular in Germany because of improved safety, is in general associated with longer delays than the traffic circle. The investigations show that traffic circles are an alternative solution especially for smaller multiphase-controlled intersections. Capacity for multiphase-controlled intersections can surpass a traffic circle only if the intersection is widened by multilane operation of the traffic streams.

TRAFFIC SAFETY

Traffic circles are very popular in many countries, particularly because of their high traffic safety. In the United Kingdom, they are regarded to be by far the safest type of intersection (*14*). In France conversions of intersections into traffic circles have led to a considerable increase in traffic safety inside (*15*) and outside (*16*) of towns. The traffic safety advantages of traffic circles mainly result from the decrease in accident severity. In Germany, there are some older traffic circles. Harder and Kinzel (*17*) reported very high accident rates on the order of six accidents per 10^6 vehicles on older traffic circles, but also found relatively low costs per accident. Because of this type of experience, there is still a lot of skepticism among German practicians concerning traffic circles.

To collect more recent data, Brilon and Stuwe (*5,9*) investigated 14 traffic circles and 14 junctions near the traffic circles. The results are summarized in Table 3. These results are explained in detail elsewhere. Special attention was paid to the individual analysis of every intersection in order to rule out the influence of individual peculiarities on the overall results.

TABLE 2 Minimum and Upper Limits of Capacity

		Roundabout	
		1-lane ADT	2-lane ADT
no problems regarding capacity	M	< 15000	< 18000
upper limit of capacity	U	25000 - 28000	35000 - 40000

TABLE 3 Accident Rate, Accident Cost, and Mean Accident Cost

ROUNDABOUTS				JUNCTIONS			
	AR	ACR	MC		AR	ACR	MC
medium and large roundabouts	6.58	24.90	3.78	junctions with traffic signals	3.35	21.73	6.49
smaller roundabouts	1.24	4.67	3.77	junctions without traffic signals	1.00	11.96	11.96

Note: AR = accident rate (10^6/veh), ACR = accident cost rate (DM/10^3 veh), MC = mean accident cost (1,000 DM/accident), and DM = deutsche mark (1.7 DM = $1.00, 1993 U.S. dollars)

The investigations show that traffic circles can be categorized by accident clearly into two groups as follows:

- Multilane traffic circles with an inscribed circle diameter of 40 to 142 m that always had an old design, and
- Small single-lane traffic circles with inscribed circle diameters of 28 to 40 m and modern design.

All accident parameters indicated that traffic safety in both groups differed considerably in scale. The older, larger traffic circles have accident rates similar to those reported by Harder and Kinzel. In this sample, however, the modern, small traffic circles have proved to be the safest intersection, especially if the accident cost rate is regarded as the central criterion.

It is also interesting to make a before and after comparison. This could be done for seven traffic circles (Table 4). More traffic circles with the necessary data could not be found. Because of the lack of suitable data, some very short after periods and different seasons had to be included. The sample size is too small, however, to draw strong conclusions. Nevertheless, if these data are considered together, the overall changes are significant and speak well for the traffic circle. Although the number of accidents (expressed by the accident rate) stayed the same, the accident severity (expressed by the accident cost rate) was clearly reduced after transformation into a new traffic circle.

These examples give only a preliminary impression. The investigations of traffic safety must therefore be continued. A general conclusion can be made, however. In Germany the installation of properly designed traffic circles appears to increase traffic safety, especially small, single-lane traffic circles. These can be regarded as a type of intersection with a high degree of traffic safety.

PRINCIPLES OF PROPER TRAFFIC CIRCLE DESIGN

If it were intended to derive design elements for traffic circles with a perfect scientific meticulousness, it would be necessary to experiment on a range of different intersection forms. Hardly anyone would fund such a project, and no town will agree to have experimental intersections constructed. Nevertheless, indications can be derived from foreign experiences, particularly on the basis of experiences with other types of intersection. These indications are derived from the four standard demands for a safe design of an intersection: clarity of the

TABLE 4 Before and After Comparisons

	JUNCTION						ROUNDABOUT					
	from to	AR	Sl	Se	F	ACR	from to	AR	Sl	Se	F	ACR
Bochum	'82 - '85	1.63	13	1	-	16.96	6.87 - 8.90	0.73	3	-	-	1.50
Brühl	6.87 - 7.89 (signalized)	1.26	7	3	-	21.47	12 mon. prov. 8.90 -16.5.91	2.24 1.73	4 4	1 -	- -	20.9 6.17
Ispringen	'87 - '88	1.74	2	7	2	168.9	7.90 -17.3.91	1.47	-	-	-	2.53
Kamen	1985	0.75	1	-	-	13.71	'86 - '89	0.81	1	-	-	2.44
Königsfeld	'88 - 8.89	2.16	2	-	-	13.16	9.89 - '90	2.52	-	-	-	4.33
Leimen	'87 - '89	0.62	3	1	-	7.32	'90 - 4.91	0.96	1	-	-	6.14
Sinnersd.	'85 - 11.89	0.06	-	-	-	0.25	12.89 -16.5.91	0.64	1	-	-	4.30
sum	225 months		31	16	2		169.5 months		14	1	-	
weighted average		1.00				23.52		1.22				5.78

AR = accident rate [accidents/ 10^6 veh]
ACR = accident cost rate [DM / 10^3 veh]

Sl = slightly injured persons
Se = seriously injured persons
F = Fatal accidents

situation for the approaching driver, visibility between road users at the inner intersection areas, comprehensibility of traffic operations, and passability for the largest permitted vehicles (such as trucks or articulated buses). Furthermore, a safe intersection is designed on the basis of the principle of harmony between design, construction, and operations. In other words, traffic regulation must be adjusted to the structural design and vice versa. Every violation of this principle creates the risk of an accident. This principle—applied to traffic circles—results in the requirement that entries must lead to the circulating lane in a vertical (i.e., radial) way. An acute-angled or tangential entry causes rear-end collisions in the approach and priority conflicts. This has been proved by Leutzbach and Ernst (*18*).

Particularly for the smaller single-lane traffic circles, further basic questions of design are associated with passability. In general, it can be assumed that traffic circles are intended to be used without restrictions by all vehicles that are licensed for road traffic. According to German regulations, every vehicle must be able to pass a circle with an outer radius of 12.5 m and a width of 7.2 m. Until October 1988 these standards were 12 and 6.7 m. Vehicles using these new standards are still the exception. Because these standards can be reached only at a very low velocity by heavy-duty vehicles or by articulated buses, an additional space of 0.40 to 0.50 m on each side is recommended for practical use in road traffic. Therefore 26 m can be regarded as the lowest value for the outer diameter of a traffic circle. This leads to a very broad circulating lane with a very small central island. Because an increase in outer radius is associated with a decrease in the width of the circulating lane, as shown in Figure 2, a larger radius is recommended.

Traffic circles that have been designed by the authors show that a minimum outer diameter of 30 m can lead to satisfactory results for both traffic engineering aspects and design. As shown in Figure 2, a circulatory highway width of 6.2 m must be assumed for the largest licensed vehicle. Considering a sufficient clearance, the resulting width of the circulating lane is 7 m. This combination of diameter and lane width leads to the effect that the central island of the traffic circle makes it impossible to look through the intersection (non-transparent traffic circle). A circulating lane with such an extreme width can cause problems with the appearance of the road space. Moreover, it could be passed at a relatively high velocity, if a free use of the whole width was possible. Therefore, the inner circle has been paved in several cases. For reasons of durability, this pavement should be laid very tight, bordered by a curb, and raised further up to the central island, which can be 2 to 3 cm higher than the asphalt surface of the circulating lane.

Experience has shown that this construction makes all passenger car drivers and motorcyclists use the outer area of the circulating lane. The paved zone is only sometimes used by the inner rear wheel of trucks and trailers. Some doubts about this construction have been raised—motorcyclists may have problems when touching the raised edge and snowplows may have difficulties. Previous experience, however, has not confirmed these doubts. Thus, the outer diameter of 30 m is the minimum value recommended when space is limited. If suf-

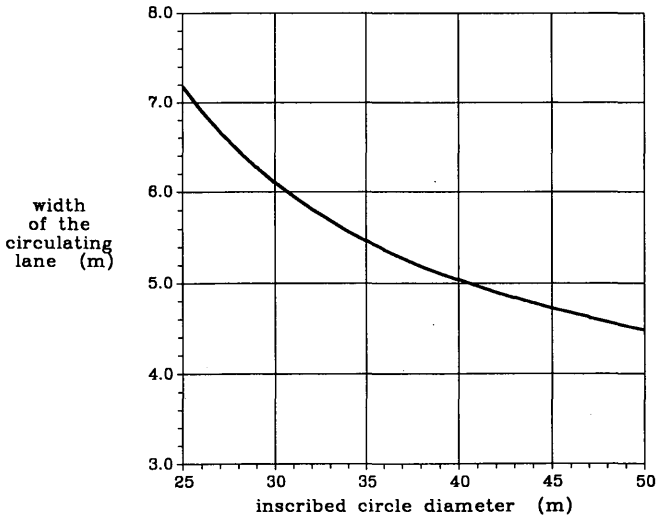


FIGURE 2 Width of circulating lane (without room to move) as function of outer radius.

ficient space is available and the traffic circle can be fitted into the environment, a larger outer diameter of about 40 m with a lane width of 6 m is recommended. A further increase in diameter facilitates higher velocities and, therefore, does not contribute to increased traffic safety.

Passability must be guaranteed in the entries and exits as well. Mistakes in design are often made in the exits. In most cases, it is sufficient to round off the corners according to the swept paths in the German guideline for intersection design (*19*). For an even more detailed examination of passability, Schult (*20*) developed a computer procedure. Except for the necessary flare at the corners, the entries and exits have the same widths as the neighboring road sections. Widening in the vicinity of the traffic circle is not recommended because it would lead to an unwelcome acceleration. If an entry is wider than 3.5 m, it should be reduced to this measure in a sufficient distance from the circulating lane. If necessary, separate lanes for buses can be added to the normal lane and carried to the outer border of the circulating lane. Thus, the traffic circle facilitates a simple way of giving priority to public transport. This design has been used in Switzerland (*21*).

Another construction detail worth mentioning is the cross slope. It should never be directed inward because this would lead to an unwelcome traffic acceleration in the traffic circle. Moreover, an inward curve would cause a negative cross slope for entering vehicles at the kinematically most problematic point of the traffic circle: the road section leading from the entry onto the circulating lane. Therefore, the cross slope should be directed outward. This also improves the visibility and clarity for drivers approaching the traffic circle. The cross slope should be 1.5 to 2.5%, which is necessary for drainage. If the whole area of the traffic circle slopes, an extreme negative cross slope can result in some sections of the circulating lane. Definitive results are not available yet. According to recent insights, however, negative cross slopes of more than 5 percent in traffic circles should not be tolerated. If this is not possible, a different type of intersection should be chosen. Of course, the drainage system should be adjusted exactly to the slope of the lane surfaces. In most cases, a level surface plane is recommended.

High-quality standards must be demanded for the asphalt surface of the circulating lane and the entries. Braking friction in the entries and lateral friction, especially caused by articulated trucks with several axles, means that the surface construction is heavily used. All curbs in the traffic circle and in the entries—especially at the pedestrian islands—should be made of a permanent white material. Reflectors can be attached to the edges of the dividing islands. These actions would improve visibility for drivers approaching the traffic circle. The same effect could be achieved by one or more signposts on the approaches to the intersection (at least outside of towns where there is no lighting) showing the traffic circle symbol. To improve visibility, the central island should be planted or equipped in a nontransparent way, that is, free sight from the entry through the traffic circle to the opposite exit is obstructed. It must, however, be guaranteed that the sight distance while driving on the circulating lane is sufficient. This is achieved primarily by the paved inner circle. It is particularly important that the central island be free of any objects that could be dangerous obstacles if hit by a vehicle, especially trees and poles, but also low bordering walls. For safety reasons, sloping curbs are the recommended construction for all curbs. These considerations have to be taken very seriously, because the central island can be hit because of inattentiveness (e.g., caused by drunkenness). With the proscribed measures, such an accident is likely to be harmless. Otherwise, it will cause serious personal injuries. A wave-like transition from the pavement to the border of the island has proven to be a suitable solution. Traffic circles within towns should be lighted. Placing light poles in the four corners outside the traffic circle is the most favorable construction. In this position, the poles are less hazardous to errant vehicles. Moreover, all islands and curbs are lighted favorably. For the reasons mentioned, an arrangement of light poles on the central island is less favorable.

Pedestrians

For safety reasons, pedestrians should be kept off the circulating lane and the central island, except if the traffic volume of the traffic circle is very low. This can be achieved, for example, by planting the area between the circulating lane and footways. Pedestrians, however, cross the entries and exits of the traffic circle. These pedestrian crossings should have a distance of one passenger car length (6 m) to the circulating lane. This distance enables the drivers to treat the conflicts with pedestrians and priority vehicles in the traffic circle separately. Experience has shown that this more relaxed situation is mainly used to let pedestrians cross the lane without problem. Moreover, in the exits this space is used for queueing in front of the pedestrian crossing.

An island should be installed between entry and exit that can be used by pedestrians as a refuge (splitter island). Without such a deflection island, pedestrians can barely see the traffic situation on the opposite side of the road, particularly for the clockwise walking direction around the traffic circle. With a deflection island, pedestrians only have a very short (3- to 5-m) conflict zone with the vehicular traffic that, moreover, comes from only one direction at a time. This makes it easier for pedestrians to make decisions, and the situation is controllable for children and senior citizens as well. Thus, behavior is more considerate to drivers. Wherever there is pedestrian traffic, the deflection islands should not be left out. The crossing itself can be either constructed as zebra crossing, which gives the absolute right of way to pedestrians in Germany, or constructed without any marking. To make the crossing visible, it is possible to pave it with formed concrete elements. This solution, however, leads to higher noise levels. Moreover, problems with the construction and maintenance of the road surface must be considered. The German guidelines for zebra crossings (*22*) can be used for decisions about the installation of zebra crossings. These guidelines require a minimum 100 pedestrians per hour crossing the road in the peak hour to justify a zebra crossing.

Taking the aforementioned design principles into account, one should, however, think of appropriate solutions that give passengers priority without constructing zebra crossings. If high volumes of pedestrian traffic occur at the crossings, a capacity reduction for motorized vehicles can result. This effect can be computed with a procedure by Griffith (*23*). Marlow and Maycock (*24*) also ascertain congestion lengths. In

the calculation procedure presented previously (*13*), this step is included as well. However, given a German traffic circle with a high volume of pedestrians, cyclists, and motorized vehicles (e.g., Münster Ludgeriplatz), the reduction of capacity by pedestrians was considerably smaller than predicted by Marlow and Maycock [see work by Brilon et al. for example (*9*)]. In any case, the procedure by Marlow and Maycock is on the safe side. Further investigations are being prepared.

The capacity reduction resulting from pedestrian traffic becomes particularly obvious in the exits. Here, the pedestrian crossings can cause congestion of vehicles on the circulating lane. If this congestion lasts long, the whole traffic circle can be blocked. Therefore, at least 90 percent of the time, congestion from pedestrian crossings at the exit into the traffic circle should be avoided. In critical cases, this can be achieved by increasing the distance between pedestrian crossing and traffic circle to more than 6 m. In most cases, pedestrians can cross the lanes without noticeable delays. This has been proved by many analyses.

These extensive remarks on pedestrian traffic are by no means intended to give the impression that this is a particularly critical subject regarding traffic circles. Normally, capacity aspects become important only in cases of extremely high volumes of pedestrian traffic, which rarely occur in practice. If the aspects mentioned are considered, traffic safety is generally very high for pedestrians at traffic circles because there are only very narrow conflict zones with the motorized traffic, which, moreover, passes these zones at very low speeds. The detour that pedestrians have to make as a result of this design is small, especially compared with a signal-controlled intersection. Of course, the footway design in the area of the traffic circle must have the effect that pedestrians do not regard it as a detour. In general, however, this aspect turns out to be unproblematic as well.

Cyclists

Bicycle traffic at traffic circles requires special attention in Germany. In general, there are four solutions for guidance on this issue:

1. Grade-separated guidance outside the traffic circle. This solution should be preferred unless general objections to undercrossings or overbridges prevail.
2. Guidance on the circulating roadway without separate lanes for cyclists. This solution is quite advantageous, especially for small traffic circles. Cyclists indicate their desired direction by cycling either on the right side or in the middle of the circulating lane. In the small traffic circles (diameter up to 40 m), they get along well in the motorized traffic without safety risks because cars and bicycles drive at almost the same speed. There are no overtakings in the traffic circle. Of course, cyclists decrease the traffic circle capacity. This can be considered in calculations by counting a cyclist as 0.5 pcu in the earlier equations.
3. Guidance of cyclists on separate bicycle paths. If the roads leading to the traffic circle are equipped with separate cycle paths, these can be continued around the traffic circle. Cyclists then cross the entries and exits on paths on the inner side of the pedestrian crossings. This solution has been implemented successfully at some of the locations mentioned. At small traffic circles, cyclists normally do not have problems when crossing the lanes, with regard to neither safety nor delays. However, one must reckon with cyclists using the cycle paths in both directions. Moreover, bicycle traffic often splits into cyclists needing protection (e.g., children, who use the separate cycle paths) and fast, sporty cyclists who stay on the circulating lane. Considering the aspects mentioned in Guidance 2, there are no objections to this behavior.
4. "Fixed tracks" for cyclists. In the past in Germany cycle paths were often marked, sometimes with a colored surface, at the outer side of the circulating lane. This solution seems to be unfavorable. According to a Swiss investigation (*25*), this solution is even regarded as dangerous. Because of the forced cycling on the right side, cyclists are unable to indicate their desired driving direction. The authors have observed that cyclists often stretch out their left arm while using the fixed track in order to prevent motorized vehicles from cutting in front of them. Moreover, the fixed track systematically leads the cyclists to the conflict zones at the entries. This problem became obvious during the accident investigations: 13 of the 26 seriously injured persons in the sample (see Table 3) were cyclists, 8 of whom were using fixed tracks. Therefore, this solution of a cycle track on the circulating lane must be rejected. If the approaching lanes are equipped with cycle tracks, the tracks should end 20 to 30 m in front of the traffic circle, which is then used by the cyclists according to Guidance 2.

These remarks cannot be regarded as conclusive rules. Rather, the problems of cyclists in traffic circles demand a more extensive investigation.

DESIGN OF ENTIRE TRAFFIC CIRCLE

The entire design of a traffic circle should be adjusted to the urban environment as much as possible. This design can be supported by including circular elements and radial alignments in the planting and building activities around the traffic circle (*26*). The circular central island offers such possibilities for individual design as plantings, memorials, or other objects of local importance. By this, traffic circles become an experience for road users and especially for residents. They can become unique points of orientation in the townscape. Aspects of architectural design, however, should not predominate over traffic engineering requirements. Traffic circles that mainly represent works of modern art are unfavorable, especially from the point of traffic safety.

Even if architectural design is not considered foremost, the traffic engineering elements of design by no means result in sterile technological constructions. Instead, the traffic circles are pleasant in appearance. After new construction of or conversion into traffic circles, road users and residents begin appreciating this type of intersection very quickly. Opinion polls show agreement of more than 95 percent. Wherever there are older traffic circles, citizens and politicians are interested in their preservation and improvement.

Small traffic circles that are designed according to the principles presented in this paper can harmonize different objec-

tives that often cause such conflicts in traffic engineering as capacity, economy, environmental benefits, safety, and design.

REFERENCES

1. W. Brilon. Der Kreisverkehr—eine vergessene Knotenpunktform? (Roundabouts—A Forgotten Type of Intersection?) *Strassenverkehrstechnik*, No. 6, Germany, 1984.
2. G. Louah. *Recent French Studies on Capacity and Waiting Times at Rural Unsignalised Intersections* (W. Brilon, ed.). Springer-Verlag, Germany, 1988.
3. W. Brilon. Leistungsfähigkeit von Kreisverkehrsplätzen. (Capacity of Roundabouts.) *Strassenverkehrstechnik*, No. 5, Germany, 1988.
4. W. Brilon and H. Ohadi. Empirische Überprüfung der Leistungsfähigkeit von Kreisverkehrsplätzen. (Empirical Investigation of Roundabout Capacity.) *Arbeitsblätter AVV Nr. 4*, Arbeitsgruppe für Verkehrstechnik und Verkehrssteuerung der Ruhr-Universität Bochum, Germany, 1988.
5. B. Stuwe. *Untersuchung der Leistungsfähigkeit und Verkehrssicherheit an deutschen Kreisverkehrsplätzen. (Investigation of Capacity and Safety at German Roundabouts.)* Dissertation. Lehrstuhl für Verkehrswesen, Ruhr-Universität Bochum, Germany, 1992.
6. W. Brilon and M. Grossmann. The New German Guideline for Capacity of Unsignalized Intersections. In *Intersections without Traffic Signals II*. Springer-Verlag, Germany, 1991.
7. R. J. Troutbeck. *Evaluating the Performance of a Roundabout*. Australian Road Research Board, 1989.
8. R. M. Kimber. *The Traffic Capacity of Roundabouts*. Laboratory Report 942. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1980.
9. W. Brilon, B. Stuwe, et al. *Einsatzmöglichkeiten von Kreisverkehrsplätzen und aufgeweiteten Knotenpunkten unter besonderer Berücksichtigung ausländischer Erfahrungen. Teil A: Kreisverkehrsplätze. (Possibilities to Use Roundabouts and Widened Intersections with Special Consideration of Foreign Experiences.)* Forschungsbericht FE 77198, 87 BMV, No. 7, Germany, 1990.
10. M. Simon. *Roundabouts in Switzerland—Recent Experiences, Capacity, Swiss Roundabout Guide*. Emch & Berger Zürich AG, Switzerland, 1990.
11. W. Brilon. *Intersections without Traffic Signals*. Springer-Verlag, Germany, 1988.
12. R. M. Kimber and E. M. Hollis. *Traffic Queues and Delay at Road Junctions*. Laboratory Report 909. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1979.
13. W. Brilon and N. Wu. *Handbuch zum Programm KREISEL. (Manual for the Software Package KREISEL.)* Lehrstuhl für Verkehrswesen, Ruhr-Universität Bochum, Germany, 1990.
14. R. D. Hall and R. A. J. Surl. Accidents at Four-Arm Roundabouts and Dual-Carriageway Junctions—Some Preliminary Findings. *TEC* No. 6, 1982.
15. J. Le Coz. Quimper: Carrefour giratoire à anneau prioritaire. *TEC*, No. 50, 1982.
16. J. M. Gambard. Safety and Design of Unsignalised Intersections in France. In *Intersections Without Traffic Signals*. Springer-Verlag, Germany, 1988.
17. G. Harder and O. Kinzel. Möglichkeiten der Einbeziehung von Unfallkosten im städtischen Bereich in die RAS-W. (Possibilities to Integrate Accident Costs in Urban Areas into the RAS-W.) *Strassenbau und Strassenverkehrstechnik*, No. 471, Germany, 1986.
18. W. Leutzbach and R. Ernst. Verkehr an Kreisplätzen. (Traffic at Roundabouts.) *Strassenverkehrstechnik*, No. 2, Germany, 1969.
19. *German Guideline for Intersection Design* (in German). Forschungsgesellschaft für Strassen-und Verkehrswesen, Cologne, Germany, 1988.
20. R. Schult and J. Holzwarth. Der Flächenbedarf repräsentativer Lastzug- und Sattelzugkombinationen im Strassenverkehr. (Space Required by Representative Truck and Trailer Combinations in Street Traffic.) *Strassenverkehrstechnik*, No. 1, Germany, 1988.
21. H. Bovy. Les giratoires: développements en Suisse romande et élaboration d'un guide suisse. *Strasse und Verkehr*, No. 9, Switzerland, 1990.
22. *German Guideline for Zebra Crossings* (in German). Forschungsgesellschaft für Strassen- und Verkehrswesen, Cologne, Germany, 1984.
23. J. D. Griffiths. A Mathematical Model of a Nonsignalized Pedestrian Crossing. *Transportation Science*, Vol. 15, No. 3, 1981, pp. 223–232.
24. M. Marlow and G. Maycock. *The Effect of Zebra Crossing on Junction Entry Capacities*. Special Report 724. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1982.
25. *Veloverhalten im Kreisel. (Bicycle Behaviors in Roundabouts.)* Arbeitsbericht Nr. 1. Emch & Berger AG, Aarau, Switzerland, 1987.
26. Schnüll. *Einsatzmöglichkeiten von Kreisverkehrsplätzen und aufgeweiteten Knotenpunkten unter besonderer Berücksichtigung ausländischer Erfahrungen. Teil B: Fahrverhalten an kleinen Kreisverkehrsplätzen und städtebauliche Integration (cf. 11, Part B: Driving Behavior at Small Roundabouts and Integration into Urban Planning.)* Forschungsbericht FE 77198/87 an den BMV, No. 7, Germany, 1990.

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

Capacity and Design of Traffic Circles in Australia

R. J. TROUTBECK

The Australian design standards for traffic circles were developed during the last decade. Early standards borrowed heavily from British research and techniques as well as the Australian practices for analyzing and designing unsignalized intersections. Recently, techniques for analyzing the performance of a traffic circle have been developed from Australian empirical relationships for gap-acceptance parameters as a function of traffic circle geometry. This has improved the ability to account for differences resulting from the geometric design. The design of traffic circles in Australia is a function of their use which may be different from other countries. Locations where and reasons why traffic circles are likely to be effective are described. In addition, the Australian technique for analyzing traffic circle performance by estimating capacity and average delay is described. The main geometric design principles are outlined as well. Traffic circles are considered to be safe intersection control devices; accident experience at traffic circles is also described.

The Australian road system differs from the road system in North America and in many parts of Europe. For instance, the Australian road system places a much higher emphasis on the use of arterial roads that are not grade-separated from other roads. Traffic engineers must be aware of different driver attitudes because they will influence the acceptability of a road-element type by another country. Small differences in the design of road elements often will be necessary to accommodate the different roles of road-element types in other countries and to allow for different driver behavior. Consequently, Australian traffic circles cannot be expected to be best for all other countries.

WHY ARE TRAFFIC CIRCLES EFFECTIVE?

Traffic circles are effective because all vehicles are slowed to a reasonable speed and decisions are simple and separated. If either of these factors is compromised, then the performance of the traffic circle will be degraded.

At a cross intersection, there are 32 decision points for which drivers are required to cross, merge, or diverge from another stream. At a traffic circle there are eight. These decision points are illustrated in the Figure 1. The traffic circle, with fewer decision points, is therefore likely to have fewer accidents. The throughput could also be increased as drivers have to do fewer tasks.

Drivers are slowed as they travel around a curved path. Their speed can be adjusted using horizontal curves of different radii for the approach and the entry. The maximum radius of the driver's path is governed by deflection, which will be described later. Suffice it to say that traffic circles must be designed to encourage drivers to travel slowly. This is also influenced by the ability of drivers to recognize that they are approaching a traffic circle and that they will be required to slow down. This requirement correctly suggests that traffic circles on high-speed roads need to be designed with a great deal of care.

Queensland University of Technology, Two George Street, GPO Box 2434, Brisbane Q4001, Australia.

WHERE SHOULD TRAFFIC CIRCLES BE USED?

Traffic circles can be used satisfactorily at a wide range of sites for which the intersection roads have roughly the same classification and purpose. These include on local and collector roads in urban areas, on arterial roads in urban areas, on rural roads, and at freeway terminals. However, traffic circles are best suited to subarterial roads for which each road has about the same importance.

Driver delays can be reduced at traffic circles that replace some intersections with two-way stop control, give-way (yield) control, or signalization. Examples of such intersections include

- Intersections that have more than four legs, and
- Intersections for which there are high proportions of left-turning traffic. Unlike most other intersection treatments, traffic circles operate most efficiently for high volumes of left-turning vehicles. For example, a left turner from the north would stop the through movement from the south, thus allowing traffic from the east to enter the traffic circle. These vehicles from the east would then stop the through movement, preventing traffic from the north from entering the traffic circle. This action reduces delays.

Traffic circles can also offer a safer intersection control as relative vehicle speeds are contained and because vehicle paths merge and not cross. Accident experience has been found to be better.

- At cross intersections of local, collector, or both local and collector roads for which a disproportionately large number of accidents occur that involve either crossing traffic or turning movements.
- At rural cross intersections (including those in high-speed areas) at which there is an accident problem that involves crossing traffic.
- At arterial-roads intersections for which traffic speeds are high and left-turning traffic flows could also be high. A traffic

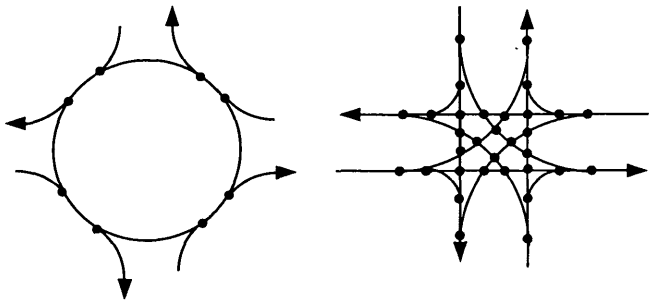


FIGURE 1 Decision points at a traffic circle (*left*) and at a conventional cross intersection (*right*).

circle could have fewer left-turn-opposed accidents than over traffic-signal-controlled intersections.

- At major roads that meet at Y- or T-intersections.
- At T- or cross-intersections for which the major traffic route turns through a right angle. Here, the turning movements are major ones.

Traffic circles can improve the amenity of a local street or residential street network. In these cases priority is not given to either road. Traffic circles can also enable flexible design. Traffic circle performance adjusts as traffic patterns or volumes change.

Traffic circles can be inappropriate in some cases.

- If a satisfactory geometric design cannot be provided owing to insufficient space or unfavorable topography;
- If traffic flows are unbalanced with high volumes on one or more approaches, and some vehicles would experience long delays (this could be improved by metering with traffic signals);
- If a major road intersects a minor road (a traffic circle would cause delay and deflection to all traffic, whereas control by signs or T-junction rule would result in delays to the minor road traffic only);
- If there is considerable pedestrian activity or if high traffic volumes would make it difficult for pedestrians to cross either road;
- If at an isolated intersection in a network of linked traffic signals, a similarly linked signalized intersection would generally provide a better level of service;
- If peak-period reversible lanes may be employed; or
- If traffic flows leaving the traffic circle would be interrupted by a downstream traffic control, which could result in queueing back into the traffic circle (a high pedestrian movement at a nearby pedestrian crossing could cause queueing into the traffic circle).

The use of traffic circles at these sites should not be completely discounted, but their effectiveness must be closely evaluated.

ELEMENTS OF A TRAFFIC CIRCLE

The major elements of traffic circles are shown in Figure 2. The definitions are self-evident from this figure.

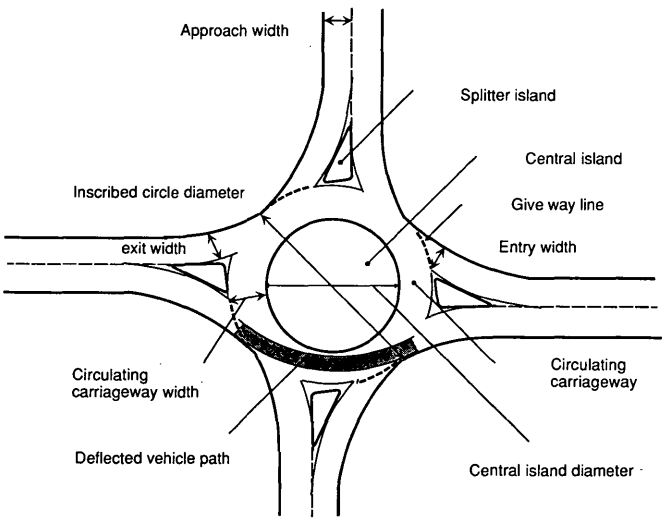


FIGURE 2 Major elements of a traffic circle.

AUSTRALIAN METHOD OF ANALYZING TRAFFIC CIRCLES

The Australian method of predicting the performance of a traffic circle is presented by AustRoads (*1*). The capacity of a traffic circle is evaluated as a series of T-intersections. This is a standard technique used in most design standards around the world. The circulating traffic is the traffic flow past the entering vehicles that opposes their entry. The entry-lane flows are the other traffic inputs. The capacity of the T-subintersection is calculated using gap-acceptance techniques.

The gap-acceptance theory can be considered to have two elements. The first is a measure of the usefulness of a gap, t sec, to an entering driver. The second element is an estimation of the frequency of acceptable gaps of duration t in the opposing traffic streams (*2*). These two elements will be discussed subsequently.

Usefulness of a Gap

The usefulness of a gap is measured by gap-acceptance parameters; that is, by the critical acceptance gap and the follow-up time. The shorter these parameters, the more useful the gaps are to entering drivers. In the AustRoads guide (*1*), the critical acceptance gap parameters are functions of the circulating flow and the road geometry.

From field studies, it has also been found that drivers in different entry lanes, at one approach, performed differently. The differences in behavior were introduced by the concepts of dominant and subdominant streams (*2*). The dominant stream is defined as the stream that has the greatest entry flow. Drivers in this stream had lower critical gap parameters, which resulted in a higher entry lane capacity. Drivers in the other entry streams (the subdominant streams) at the same leg had larger critical gap parameters. These streams also had a lower capacity. There is only one dominant stream at each entry, but there may be many subdominant streams. If there is only one entry stream, it will be a dominant stream (*3*).

A critical acceptance gap and follow-up time were calculated for each lane (*1,2*). The follow-up time for the dominant stream, t_{fdom}, was calculated first using the equation

$$t_{fdom} = 3.37 - 0.000394Q_c - 0.0208Di + 0.0000889Di^2 - 0.395n_e + 0.388n_c \quad (1)$$

where

Q_c = circulating flow (veh/hr)
Di = inscribed diameter, the largest diameter that can be drawn inside traffic circle (m),
n_e = number of entry lanes, and
n_c = number of circulating lanes.

The follow-up times for the subdominant stream, t_{fsub}, were a function of the dominant stream follow-up time values, t_{fdom}, and the ratio of the dominant stream entry flow, Q_{dom}, to the subdominant stream entry flow, Q_{sub}.

$$t_{fsub} = 2.149 + 0.5135t_{fdom}\frac{Q_{dom}}{Q_{sub}} - 0.8735\frac{Q_{dom}}{Q_{sub}} \quad (2)$$

The larger the dominant stream follow-up time, the larger is the subdominant stream follow-up time. The dominant stream follow-up time also increases with larger variations in the lane entry flows.

The critical gap is dependent on the follow-up time, the circulating flow, the number of circulating lanes, and the average entry lane width $\bar{e}$. An expression for the ratio of the critical gap, t_a, to the follow-up time, t_f, was found to decrease with an increased circulating flow, the number of circulating lanes, and the average entry lane width. This equation was applied to the conditions in all entry lanes.

$$t_a/t_f = 3.6135 - 0.0003137Q_c - 0.3390\bar{e} - 0.2775n_c \quad (3)$$

Number of Useful Gaps

The number of useful gaps in the circulating traffic stream is influenced by the degree of bunching in the circulating stream. The distribution of gaps in the circulating stream are modeled by Cowan's M3 model (*4*). This model provides a good description of the headways between the nonbunched (or free) vehicles. It models the headways between bunched vehicles rather poorly because it considers that all bunched vehicles have the same short headway, τ. This deficiency is not a concern because headways between bunched vehicles are not accepted by entering drivers.

The entering drivers were found to give way to all circulating vehicles regardless of whether the circulating vehicle was in the inner or outer circulating lane. The characteristics of the circulating traffic were developed on the basis of a single-lane flow. Troutbeck (*3,5*) has indicated that it is acceptable to model the combined influence by a single lane.

The two terms used to define the circulating stream characteristics are the proportion of nonbunched (free) vehicles, α, and the minimum headway between bunched vehicles, τ. For multiple lanes, τ is the average headway between closely following vehicles in all lanes. Because τ and α are interrelated, values of τ were chosen on the basis of the number of circulating lanes. Expressions for α were then derived from field data for traffic circles with single- and multiple-circulating lanes. The regression equations were found to be similar and the following generalized equation was developed:

$$\alpha = 0.75(1 - \tau Q_c/3{,}600) \quad (4)$$

where τ equals 1.0 if there are more than one circulating lanes and 2 if there is one circulating lane.

Entry Capacity

Entry capacity is defined as the maximum entry flow for a particular lane, intersection geometry, and other opposing flows. Entry capacity depends on the opposing flows: as the opposing flows increase, the entry capacity decreases. This definition is similar to the British definition (*6*) but is quite different from the definition used by Kyte et al. (*7*), who defined capacity in terms of when flows in all streams were increased, and the intersection had reached the maximum level of performance. Kyte's definition is appropriate for all-way stop intersections.

The entry capacity–circulating flow curves have the shape shown in Figures 3 and 4 for single-lane and multilane traffic circles. The shape of these curves are similar to the curves obtained by Kimber (*6*).

The entry capacity for each entry lane is given by

$$C = \frac{\alpha q_c e^{-\lambda(t_a - \tau)}}{1 - e^{-\lambda t_f}} \quad (5)$$

where

C = absorption capacity of an entry lane (veh/sec)
α = proportion of free vehicles in circulating streams,
q_c = flow of vehicles in circulating streams (veh/sec) or $Q_c/3600$,
t_a = critical acceptance gap,
t_f = follow on time, and
τ = minimum headway in circulating streams, and these are related by

$$\lambda = \alpha q_c/(1 - \tau q_c) \quad (6)$$

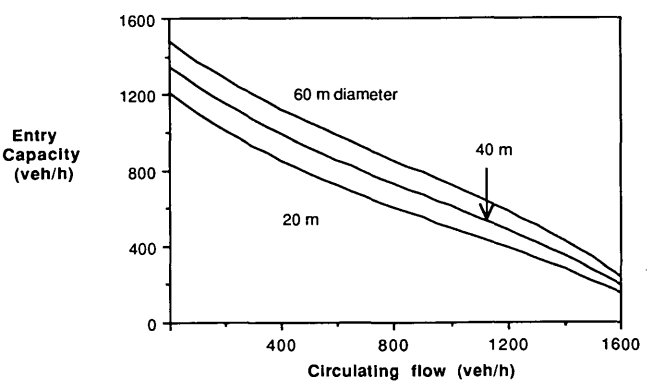


FIGURE 3 Effect of increasing inscribed diameter of traffic circles with a single circulating lane (entry lane width is 4 m).

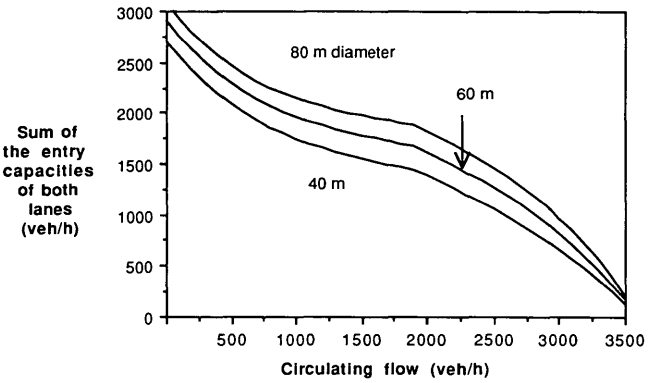


FIGURE 4 Effect of increasing inscribed diameter of traffic circles with multiple circulating lanes (entry lanes are 4 m wide).

The inscribed diameter and the average entry-lane width had the greatest influence on the estimates of capacity. The larger the inscribed diameter, the easier the entry turn is likely to be. This will translate into shorter critical-gap parameters and a larger capacity. Figures 3 and 4 illustrate the effect of the inscribed diameter on the capacity of a single circulating lane and a multiple circulating lane traffic circle. The effect is significant. A 20-m increase in the inscribed diameter results in an increase of between 10 and 25 percent in the capacity.

Estimates of Delay

Two types of delay should be considered. The first is queueing delay and is the delay for drivers while they wait for an acceptable gap. The second delay is geometric delay, which is the delay incurred as drivers slow to approach a traffic circle or stop at the end of a queue. Geometric delay also includes the delay traveling around the traffic circle and accelerating to the departure speed.

Queueing Delay

Queueing delay is a function of gap-acceptance parameters and flows on circulating and entry lanes. AustRoads (*1*) introduced the Akçelik and Troutbeck nonsteady state solution (*8*). This solution requires the user to estimate the duration of the peak period. It assumes that the flow is zero before the peak period, and the average delay is estimated for all vehicles arriving in this period.

An equation for the average delay per vehicle is

$$D = D_m + 900T\left[x - 1 + \sqrt{(x-1)^2 + \frac{D_m x}{450T}}\right] \quad (7)$$

where

D = average delay per vehicle in seconds;
D_m = average delay when the entry stream flow is low, that is, when x is close to zero;
T = duration of the peak flow period in hours (typically 1 hr); and
x = degree of saturation.

$$D_m = \frac{e^{\lambda(t_a-\tau)}}{\alpha q_c} - t_a - \frac{1}{\lambda} + \frac{\lambda\tau^2 - 2\tau(1-\alpha)}{2(\lambda\tau + \alpha)} \quad (8)$$

$$x = q_e/C \quad (9)$$

where q_e is entry flow.

Equation 7 is always less than the steady-state delay equation.

$$D = \frac{D_m}{1-x} \quad (10)$$

This is illustrated in Figure 5. Note that the average queueing delay is strongly dependent on the chosen value of T. The designer must be careful to select the appropriate value of T.

The equations presented in this paper give estimates of capacity and delay. These equations are now included in the computer routines SIDRA 4 (*8*) and KREISEL (*9,10*).

Geometric Delay

The average geometric delay is dependent on the proportion of drivers stopped and the distance traveled around the traffic circle at the slower negotiation speed. The average delay was calculated using the assumption that drivers accelerated at 1.2 m/sec^2 and decelerated at 1.8 m/sec^2. The negotiation speed was a function of the drivers' turn radius in the traffic circle. Geometric delays can be substantial.

The chances of being stopped at an intersection depends on two factors. The probability of being stopped increases as the traffic on the major road increases. However, the probability of being stopped is also proportional to the degree of saturation on the minor road. If the degree of saturation is high, there would be heavy queueing and the probability of being stopped also increases (*11*).

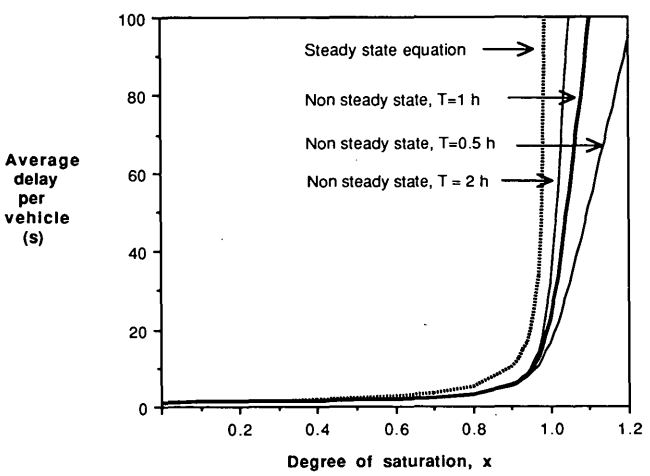


FIGURE 5 Average delay per vehicle as a function of degree of saturation for a traffic circle entry when w_h term is equal to 2 sec.

Estimates of the proportion of stopped vehicles of saturation were obtained from a simulation program for gap-acceptance behavior. Given that the geometric delays to vehicles stopped and not stopped were not well defined, it was decided to calculate the proportion of stopped vehicles from either a traffic circle with two circulating lanes (a 60-m inscribed diameter and two 4-m entry lanes), or a single-lane traffic circle with a 40-m inscribed diameter. The expected proportions of stopped vehicles are shown in Figure 6 for a multilane traffic circle. The results for a single-lane traffic circle are similar (*12*).

The average geometric delay, d_g, for each turn type and from each approach is given by the equation

$$d_g = P_s d_s + P_n d_n \quad (11)$$

where

P_s = probability of being stopped,
P_n = probability of not being stopped = $1 - P_s$,
d_s = average delay if stopped, and
d_n = average delay if not stopped.

Performance Measures

The performance measures are the average delay, including both the geometric and the queueing delay, and the degree of saturation, which is the entry flow divided by the entry capacity. There is no limiting value for the average delay, but it is recommended that the degree of saturation be less than 0.85.

DESIGN OF TRAFFIC CIRCLES

Deflection Through Traffic Circles

Adequate deflection of the paths of vehicles entering a traffic circle is the most important factor to influence their safe operation. Traffic circles should be designed so that the speed of all vehicles within the intersection is as slow as reasonable but certainly less than 50 km/hr. This is done by ensuring that through-vehicle paths are significantly deflected by positioning a suitably sized central island, using splitter islands, and adjusting the alignment of the entrances and exits. The splitter island is the small, roughly triangular island that is in the entry road and separates the entering vehicles from the exiting vehicles.

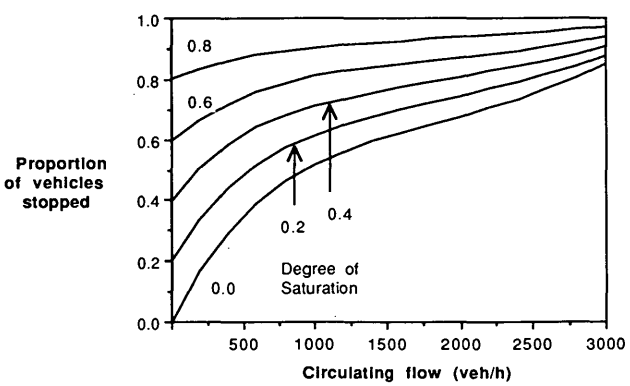


FIGURE 6 Proportion of stopped vehicles based on the simulation results for a traffic circle with multiple circulating lanes, an inscribed diameter of 60 m, and two 4-m entry lanes.

The importance of achieving adequate deflection cannot be overemphasized. It is required to reduce vehicle speeds. British research (*13*) has indicated that the frequency of casualty accidents is increased at sites for which the deflection has been reduced. Australian practice also confirms the importance of adequate deflection and geometric layout, to control vehicle speeds on the entry to a traffic circle.

The maximum desired design speed is obtained if no vehicle path (assumed 2 m wide, with the vehicle using all the available road width) has a radius greater than 100 m. This degree of curvature corresponds approximately to 50 km/hr with a sideways acceleration of 0.2 g. Here the central island size and the approach geometry are the controlling factors.

Central Island

Circular central islands are preferable, otherwise the driving task demand changes as the driver negotiates the traffic circle. The size of the central island is determined principally by the space available and the need to obtain sufficient deflection to reduce through-vehicle speed. Traffic circles with a larger inscribed diameter, and consequently a larger central island, have a slightly larger entry capacity. In areas in which drivers are likely to be unfamiliar with traffic circle operation, the central island should have a diameter of at least 4 m and preferably more than 10 m.

Width of Circulating Highway

The width of the circulating highway depends on the number of entry lanes and the radius of vehicle paths within the traffic circle. The width is generally greater than the widths of the approach lanes simply because vehicles must turn. For instance, a single-lane traffic circle with a 10-m central island needs a 7-m circulating highway width to enable an articulated vehicle to turn right.

Splitter Islands and Entrance and Exit Curves

Splitter islands should be provided on all traffic circles installed on arterial roads. They provide shelter for pedestrians, guide traffic into the traffic circle, and deter wrong movements. The Australian approach is to design splitter islands that provide a smooth curve with the circulating highway. On arterial road traffic circles the splitter island should be at least 8 m^2 to present a reasonable target to be seen by approaching drivers.

Entry and exit lane widths are generally in the range of 3.4 to 4.0 m. Exceptions are for curbed single-lane entrances and exits where a minimum of 5.0 m between curbs is required to allow traffic to pass a disabled vehicle. The exit from a traffic circle should be as easy to negotiate as practicable. After having been slowed by the curved entry path, vehicles

should be able to accelerate out of the circulating highway in the exit lane.

The approach curves to traffic circles are upstream of the entry curve and should be used if the upstream approach speed is more than 10 km/hr greater than the entry speed defined by the entry radius. It is important to advise drivers that they are approaching a traffic circle and to slow down. This is best achieved with horizontal curves.

Sight Distance

Several sight-distance criteria should be applied to both the vertical and horizontal geometries at traffic circles. These criteria also affect the positioning of signs and plantations and so on. The approaching driver must have a good view of the splitter island, the central island, and desirably the circulating highway. Adequate approach–stopping-sight distance should be provided, preferably to the give-way (yield) lines and, at an absolute minimum, to the nose of the splitter island.

A driver, stationary at the yield line, should have a clear line of sight to approaching traffic for a distance representing at least the travel time equal to the critical-acceptance gap. A value of 5 sec is typical for traffic circles operating with low circulating flows, and 4 sec would be acceptable at the traffic circles with higher flows but slower speeds.

When drivers reach 40 m before the yield line on approach to a traffic circle, they should be able to see other entering vehicles, although in urban areas it may not always be possible.

Visibility Considerations

At any traffic circle, designers must provide the sight distance just described, but they must ensure that the traffic circle is conspicuous. The driver must be readily able to assess the driving task. To enhance the prominence of the traffic circle, the curbs on both the splitter island and central island should be light colored or painted white. As with other types of intersections, it is better to position a traffic circle in a sag vertical curve. The circulating highway must be conspicuous. This is best achieved with negative crossfall; that is, by sloping the crossfall away from the central island.

Superelevation of Circulating Highway

Normal curve superelevation through the traffic circle is generally not necessary as speeds are constrained, and drivers tolerate higher values of the coefficient of sideways friction traveling through an intersection. It is important, however, that the layout of the traffic circle is clearly visible to approaching drivers and provides for adequate drainage. As a general design practice, a minimum pavement crossfall of 0.025 to 0.03 m^2 should be adopted for the circulating highway. A crossfall as low as 0.02 m^2 has been found to allow for pavement drainage and also provides additional driver comfort. Designing superelevation to slope away from the central island often simplifies the detailed design of pavement levels and avoids drainage pits around the central island.

Exceptions to this requirement include larger traffic circles with an inscribed diameter more than 100 m and traffic circles on a slope. These latter traffic circles have a crossfall across the whole of the traffic circle rather than a crossfall away from the central island. Traffic circles should not be used on grades greater than about 3 percent.

Pavement Markings at Entry and Exit

The linemarking used at the give-way point consists of a 300-mm striped line across the entry. The markings follow the circumference of the traffic circle to allow the curb side drivers to see past entering vehicles on their left. There are no lines across the exits.

There is divided opinion on whether lane lines should be used to delineate circulating lanes within a traffic circle. There is no definitive research on the subject. AustRoads (*1*) recommends that lane marking in the circulating highway be used only in the shadow of the splitter islands on larger traffic circles (inscribed diameter of 50 m or more). The absence of linemarking between the entry and exit, where most lane changing occurs, minimizes confusion about the drivers' requirement for lane change signaling.

Direction arrows are not necessary in the entry lanes on the approach to the yield line, except when exclusive left-turn lanes (right turns in the United States) are provided. Some arrows may mislead some drivers into making incorrect turns before the central island.

Traffic Signs

An inverted triangular regulatory sign is used to identify a traffic circle. On the splitter islands, signs are used to advise drivers to drive on the correct side of the island. For large splitter islands, particularly for traffic circles in high-speed areas, hazard boards are used to emphasize the curved approach and deflection into the traffic circle.

SAFETY RECORD OF TRAFFIC CIRCLES

The safety performance of traffic circles has been documented in a number of Australian and U.K. studies. Before and after accident studies carried out at intersections that involve a wide range of site and traffic conditions, at which traffic circles have been constructed, indicate very significant reductions in casualty rates.

The good safety record of properly designed traffic circles can be attributed to the following factors:

- Traffic circles result in general reduction in conflicting traffic speeds of all vehicles (limited to less than 50 km/hr).
- Traffic circles eliminate high angles of conflict, ensuring low relative speeds between conflicting vehicles.
- Traffic circles provide for relative simplicity of decision making at the point of entry.
- Long splitter islands provide good advance warning of the presence of the intersection on individual roads in high-speed areas.
- Splitter islands provide refuge for pedestrians and permit them to cross traffic from one direction at a time.

- Traffic circles always require a conscious action on the part of all drivers passing through the intersection, regardless of whether other vehicles are present or not.

In 1981 VicRoads carried out a before and after study to assess the safety performance of 73 traffic circle sites throughout Victoria (*14*). The form of control during the before period was either Stop or Give Way sign controls, or in one case, police control. The sites were primarily in urban areas, although some rural sites were included.

The study results indicated that the average casualty accident rate decreased by 74 percent after traffic circle installations. Sites were grouped according to entering traffic volumes. All groups showed statistically significant reductions in accident rates. The sites with lighter traffic volumes had a 95 percent reduction, and the moderate to heavy sites had a 59 percent reduction. There was a 32 percent reduction in property-damage accidents recorded at the study sites. This is inconclusive because all property-damage accidents are not reported. In the 3 yr following the installation of two traffic circles on high-speed roads in Victoria (with 100 km/hr speed limits), there were no casualty accidents. The cross intersections had been controlled by Give Way or Stop signs. There was a 68 percent reduction in casualty accidents per year that involved pedestrians following traffic circle installation for all sites combined. This result is encouraging. However, owing to the low number of pedestrian accidents, the reduction was not statistically significant at the 0.10 level. Pedestrian and cyclists are often more at risk at traffic circles. Jordan (*15*) reported that pedestrian accidents were reduced at traffic circles, although traffic circles generally present increased difficulty to crossing pedestrians. Similarly, cyclists also encounter difficulty when riding through traffic circles. Jordan reports that the accident rate for cyclists is increased at traffic circles. Methods to improve traffic circle safety are being investigated.

Arndt is currently investigating the influence of traffic circle geometry on accident rates (*16*). Arndt's preliminary results have revealed several useful trends.

- Traffic circles with no left-hand (right-hand in the United States) approach curvature and high approach speeds have an above-average number of accidents with entering and circulating vehicles colliding.
- Traffic circles with no through deflection provided have a large number of accidents that involve collisions between entering and circulating vehicles.
- Traffic circles with a small approach radius and high approach speeds produce single-approaching-vehicle accidents.
- Traffic circles for which there is poor recognition of the central island from approach legs will lead to accidents between approaching vehicles and other approaching vehicles. Single-approaching-vehicle accidents will also occur.
- Traffic circles that are large in diameter, elliptical, and located in high-speed areas with adverse crossfall on circulating lanes lead to instability for larger vehicles.
- Traffic circles that have exit legs with very small splitter islands and a small radius will lead to some accidents that involve exiting vehicles that collide with approaching vehicles.

CONCLUDING REMARK

After many thousands of traffic circles have been constructed in Australia, traffic circles have been demonstrated to be very useful devices in controlling traffic at road intersections. Major advantages include the provision of adequate throughput and driver safety, basically through slower vehicle speeds and low conflict angles. Their use in other parts of the world is encouraged.

REFERENCES

1. *Guide to Traffic Engineering Practice. Part 6—Roundabouts.* AustRoads, Sydney, Australia, 1993.
2. R. J. Troutbeck. *Evaluating the Performance of a Roundabout.* Special Report 45. Australian Road Research Board, Victoria, 1989.
3. R. J. Troutbeck. Traffic Interaction at Roundabouts. *Proc., 15th Australian Road Research Board Conference*, Vol. 15, No. 5, 1990, pp. 17–42.
4. R. J. Cowan. Useful Headway Models. *Transportation Research*, Vol. 9, No. 6, 1975, pp. 371–375.
5. R. J. Troutbeck. Unsignalised Intersections and Roundabouts in Australia: Recent Developments. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, Berlin, Germany, 1991, pp. 238–257.
6. R. M. Kimber. The Traffic Capacity of Roundabouts. Laboratory Report 942. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1980.
7. M. Kyte, J. Zageer, and K. B. Lall. Empirical Models for Estimating Capacity and Delays at Stop-Controlled Intersections. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, Berlin, Germany, 1991, pp. 335–361.
8. R. Akçelik and R. J. Troutbeck. Implementation of the Australian Roundabout Analysis Method in SIDRA in Highway Capacity and Level of Service. *Proc., International Symposium on Highway Capacity* (U. Brannolte, ed.), Karlsruhe, Germany, July 1991.
9. W. Brilon and N. Wu. *Handbuch zum Programm KREISEL.* (*Manual for the Software Package KREISEL.*) Lehrstuhl für Verkehrswesen, Ruhr-Universität Bochum, Germany, 1991.
10. W. Brilon and B. Stuwe. Capacity and Safety of Roundabouts in West Germany. *Proc., 15th Australian Road Research Board Conference*, Vol. 15, No. 5, 1990, pp. 17–42.
11. R. J. Troutbeck. The Characteristics of the Times Drivers are Stopped at Unsignalised Intersections. To be presented at the 12th International Symposium on Transportation and Traffic Theory, Berkeley, Calif., July 1993.
12. R. J. Troutbeck. Changes in the Analysis and Design of Roundabouts Initiated in the AustRoads Guide. *Proc., 16th Australian Road Research Board Conference*, Vol. 16, No. 5, 1992, pp. 245–262.
13. *Highway Directorate—Technical Memorandum on Roundabout Design.* Technical Memorandum H2/75, Department of the Environment, London, England, 1975.
14. E. Richardson. Experience with Road Humps and Roundabouts in Perth. Seminar on Area Traffic Management 1982, Local Government Engineers Association of Western Australia, 1982.
15. P. W. Jordan. Pedestrians and Cyclists at Roundabouts. *Proc., 3rd National Local Government Engineers Conference*, 1985, pp. 196–205.
16. O. Arndt. *Roundabout Safety Study: Effect of Geometry on Accident Rate.* Report DSB01. Transport Technology Division, Queensland Department of Transport, Australia, 1991.

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

Left-Turn Adjustment for Permitted Turns from Shared Lane Groups: Another Look

ELENA SHENK PRASSAS AND ROGER P. ROESS

Initial research on the left-turn adjustment factor for shared, permissive left-turn lane groups resulted in the recommendation that the theoretical model presented in the 1985 *Highway Capacity Manual* (HCM) be replaced by regression models. After several meetings and spirited debates, the Signalized Intersection Subcommittee of the Highway Capacity and Quality of Service Committee recommended that the regression-based models not be adopted for the HCM. The weaknesses in the 1985 HCM are discussed. Illustrations are provided to show how improved accuracy can be achieved with a hybrid model that uses parts of regression-based models inserted into the theoretic framework of the HCM model. The hybrid model is compared with the 1985 HCM and with other suggested methods. The hybrid model was found to be best for predicting saturation flow on the basis of field data for 267 data periods of 15 min each. At a workshop in summer 1992, the committee concluded that the hybrid model will be incorporated into the HCM.

In early 1990 the results of the FHWA-sponsored effort *Levels of Service in Shared, Permissive Left-Turn Lanes at Signalized Intersections* were reported in both a Final Report (*1*) and a TRB paper (*2*). The recommendations of the report included replacing the theoretic model for left-turn factors for shared, permissive lane groups in the 1985 *Highway Capacity Manual* (HCM) (*3*) with regression-based models. These models more accurately predicted field-measured values of saturation flow rate from 25 intersections in four urban areas across the United States.

The regression models were simpler than those of the 1985 HCM but lacked the clear logic of the theoretic approach. For single-lane approaches, the conclusion that the left-turn adjustment factor did not depend on the opposing flow rate was also counterintuitive, although rational explanations could be constructed.

Over the following 3 years, several meetings and discussions by the Signalized Intersection Subcommittee of the TRB Committee on Highway Capacity and Quality of Service led to the recommendation that the regression models not be adopted for inclusion in the HCM. The full committee directed the subcommittee to organize and conduct a workshop in summer 1992 to reexamine all aspects of the controversy and come to a positive conclusion on what to incorporate into the HCM.

This paper is the result of one of the analyses conducted for that workshop. It presents an argument for merging some key aspects of the regression models developed in the research effort into the general theoretic framework of the 1985 HCM. Some key modifications to that framework are also recommended to address important omissions.

At the workshop in summer 1992, the committee approved this hybrid method of merging the regression models into the 1985 HCM framework, and it is to be incorporated into the next printing of Chapter 9 of the HCM. At the same time, it was decided that the ideal saturation flow rate for an intersection should be increased from 1,800 to 1,900 per cars per hour of green time per lane (pcphgpl). This was one of the recommendations made in the initial FHWA-sponsored research and was developed on the basis of the measurements observed for ideal saturation flow rate at the 25 intersections used in the data base.

The recommended hybrid model given in this paper is compared with other methods that have been recommended for analyzing permitted left turns. All of the methods are compared using 1,900 pcphgpl as the ideal saturation flow, because this number has already been approved as the new ideal saturation flow rate for the HCM.

1985 HCM MODEL: A THEORETIC FRAMEWORK

The 1985 HCM model for the left-turn adjustment factor applied to permitted left turns made from shared lane groups is

$$f_m = (g_f/g) + (g_u/g)\,[1/1 - P_L(E_L - 1)] + (2/g)\,(1 + P_L)$$

$$f_{lt} = [f_m + (N - 1)]/N$$

where

- f_m = left-turn adjustment factor applicable only to lane from which left turns are made,
- f_{lt} = left-turn adjustment factor applicable to the entire lane group from which left turns are made,
- g_f = green time until arrival of first left-turning vehicle in subject lane group,
- g_u = unsaturated portion of green phase; effective green remaining after clearance of opposing queue,
- P_L = proportion of left-turning vehicles in left lane, and
- E_L = through-car equivalent of a left-turning vehicle.

The model for f_m presumes a rigid structure of the green phase into three component portions. In addition to g_f and g_u, there is also g_q, defined as the green time consumed in clearing the opposing queue of waiting vehicles from the intersection.

Polytechnic University, Six Metrotech Center, Brooklyn, N.Y. 11201.

When the light turns green, the opposing queue takes g_q sec to clear. During this time, it is assumed that no left turns can be made, as the clearing queue effectively presents an opposing flow without gaps. The remaining time, $g - g_q$, is by definition, g_u. During g_u, left turns are made through gaps in an unsaturated opposing flow. There is, however, the initial period, g_f, during which flow from the shared lane proceeds uninterrupted, because no left turns are present. The 1985 HCM presumes that g_f should never be more than g_q, an assumption that will be discussed later. The basis of HCM model is therefore the following:

- During g_f, there are no left turns present. Thus, the left-turn factor for this portion of the green should logically be 1.00. The term g_f/g is therefore multiplied by 1 in the model.
- Assuming that the first left-turning vehicles arrives before the clearance of the opposing queue, there is a period $g_q - g_f$, during which the first left-turning vehicle blocks the shared lane while waiting for the unsaturated portion of the green phase. Because there is no flow in the shared lane during this period, the left-turn factor during this period is logically 0.00. There is no term addressing this period in the model for this reason.
- During g_u, left turns are made through the unsaturated opposing flow. The friction of the left-turn conflict with opposing vehicles, however, restricts the rate at which these can be made. An adjustment factor of between 0.00 and 1.00 logically applies here. The HCM computes a factor based on E_L that is computed as 1,800 (1,400 $- v_o$), which is the ratio of the ideal saturation flow rate to the saturation flow rate of left turns made through an opposing flow.
- The last term of the HCM model accounts for "sneakers," who complete their left turns during the clearance interval. Each sneaker effectively adds 2 sec to the effective green—hence, 2/g. The model posits that between 1 and 2 such vehicles will exist on the basis of the proportion of left turners in the shared lane, P_L.

In the HCM model, f_m is extended to f_{lt} by assuming that there is no impact of left turns in the shared lane on adjacent through lanes in multilane approaches. In effect, an average adjustment factor is computed for all lanes, assuming that f_m applies to the shared lane, and that 1.00 applies to all others in the lane group.

DIFFICULTIES WITH HCM MODEL

There are three principal difficulties with the HCM model as it is currently constituted. The first is the assumption that g_f should not be permitted to be more than g_q. In practice, of course, the first left-turning vehicle can arrive either before or after the clearance of the opposing queue. In the data base developed for the FHWA study, a significant portion of the periods studied had $g_q < g_f$. If the HCM rule is followed in this case, an adjustment factor of 1.00 is applied only through g_q. For the period $g_f - g_q$, the adjustment factor for g_u is applied. Thus, for some portion of the green, a factor of less than 1.00 is applied when there are no left turns present. Figure 1 illustrates this anomaly and the recommended correction. A factor of 1.00 should be applied until the arrival

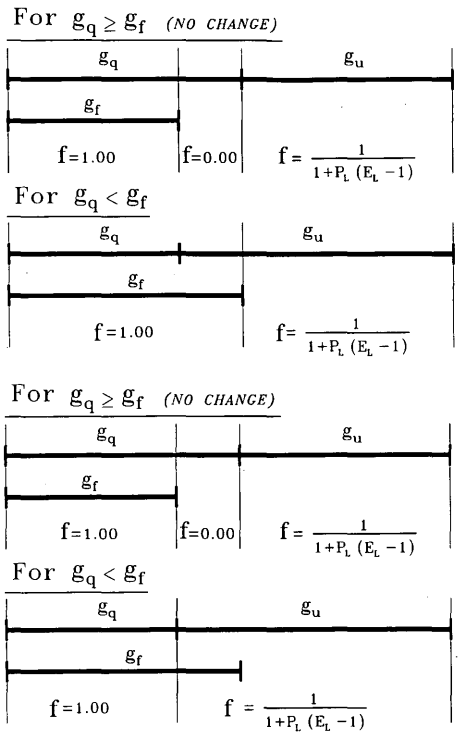


FIGURE 1 Interpretations of g_f, g_q, and g_u: *top*, recommended model; *bottom*, 1985 HCM model.

of the first left-turning vehicle, regardless of whether this is before or after clearance of the opposing queue. Thus, g_f should not be constrained. Then, g_u would be redefined as follows:

$$g_u = \begin{cases} g - g_q & g_q \geq g_f \\ g - g_f & g_q < g_f \end{cases}$$

With this revision, an adjustment factor less than 1.00 is only applied to the unsaturated portion of the green for which left-turn demand exists.

The second major difficulty with the HCM model occurs when it is applied to single-lane approaches. In these cases, the assumption that no left turns can be made during the period $g_q - g_f$ (for those cases in which $g_q > g_f$) is incorrect. When single-lane approaches are opposed by single-lane approaches, left turns in one direction open a gap for left turns in the other to be made, as illustrated in Figure 2. Thus, if the opposing queue of vehicles includes left-turning vehicles, some left turns will be made through gaps opened in the opposing queue. Thus, the adjustment factor applied during the period, $g_q - g_f$ should not be 0.00 for the case of opposing single-lane approaches. A recommendation for what it should be will be presented.

Opposing left turns that open gaps for each other is a unique characteristic and is one reason that regression models did not indicate any sensitivity of left-turn factors to opposing flow for such cases. Left turns will be made regardless of opposing flow on the basis of the balance of left-turning vehicles in the opposing and subject flows. Whereas the HCM model attempts to address this by eliminating left turns from

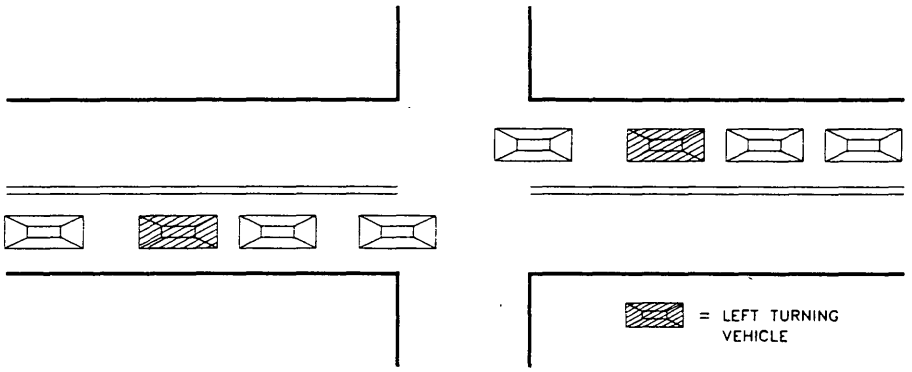


FIGURE 2 Opposing left turns on a single-lane approach.

a single-lane approach from the computed opposing flow rate, this is an arithmetic approach that does not explicitly account for left turns made through the opposing queue and the resulting reduction in the blockage of the shared lane.

The third difficulty is the assumption (for multilane approaches) that left-turning operations in a shared lane have no impact on adjacent lanes in the subject lane group. In multilane situations, left-turners create delay in the shared lane that other drivers seek to avoid by moving into adjacent lanes, causing additional turbulence in adjacent lanes. It would be logical to assume that there is some impact in these lanes as well, perhaps at some reduced level. Results of the research demonstrated that there was such an impact.

There are two additional problems with the HCM model, which are more practical than theoretical. The HCM model, in predicting g_u and g_q, does not consider the effect of signal progression and and the platooning of vehicles on the queue clearance time. Thus, for example, changing the offset at an intersection that is part of an arterial will have no effect on the queue or the queue clearance time in the current model. Second, the HCM model necessitates an initial "guesstimate" of saturation flow rate in the opposing direction before the saturation flow rate is computed for the subject direction. This creates a circular logic that should most properly be iterated, which is not done in the HCM.

An additional problem was uncovered during tests of the models using varying ideal saturation flow rates. It was found that reducing the ideal saturation flow rate yielded more accurate results. This was perplexing, particularly because the FHWA study included direct observations of this parameter and suggested that 1,900 pcphgpl is a more realistic value to use. This is explained by a simple fact—the studies of left-turn adjustment factors assume that all other factors given in the HCM are correct. That better results are achieved using lower ideal saturation flow rates indicates that the other factors do not account for all of the negative impact of other nonideal conditions.

RECOMMENDATIONS FOR IMPROVEMENTS

Adopt Regression Models for g_q, g_f, and g_u

The HCM model starts with a "guesstimate" of opposing saturation flow rate because one is needed to establish g_f, g_u, and g_q analytically. Use of regression models for these values avoids this need and removes the need to iterate a circular procedure in which opposing values of saturation flow rate are interdependent. The FHWA research produced models for predicting these values that were more accurate than the HCM. The models for g_f are summarized by the following:

$$g_f = \begin{cases} G \exp(-0.860\ \mathrm{LTC}^{0.629}) - t_L & \text{single-lane approaches} \\ G \exp(-0.882\ \mathrm{LTC}^{0.717}) - t_L & \text{multilane approaches} \end{cases}$$

where

G = actual green time for the phase;
LTC = average number of left turns per cycle, which may be fractional; and
t_L = total lost time per phase (sec).

Figure 3 illustrates the accuracy of these models in predicting g_f compared with the HCM model for 305 study periods, each 15 min long, at 25 intersections in four major urban regions: New York; Chicago; Austin, Texas; and Los Angeles.

The models for g_q are

$$g_q = \begin{cases} 4.943\ v_{olc}^{0.762}\ qr_o^{1.061} - t_L & \text{single-lane approaches} \\ 9.532\ v_{olc}^{0.560}\ qr_o^{0.819} - t_L & \text{multilane approaches} \end{cases}$$

where

v_{olc} = opposing flow expressed in units of per lane and per cycle,
t_L = total lost time per phase, and
qr_o = opposing queue ratio.

The models for g_f and g_q were calibrated from the beginning of the actual green time for the given movement. Because it is the effective green time that needs to be partitioned into periods of use for left turns, the model values are adjusted by the lost time, t_L, to account for the difference between effective green time and actual green time.

The opposing queue ratio, qr_o, is defined as the proportion of opposing flow originating in the opposing queue (i.e., during g_q). For most cases, this can be taken as the same as the proportion of vehicles arriving on red, or 1 − (proportion of vehicles arriving on green). Because the proportion of vehicles arriving on green is needed to implement new models for delay estimation, this recommendation does not introduce another variable into the overall intersection methodology.

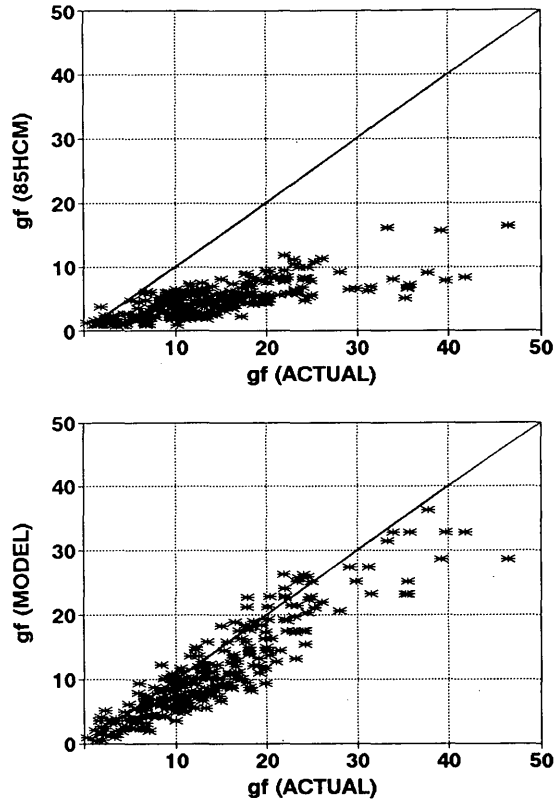


FIGURE 3 Comparative accuracy in prediction of g_f: *top*, 1985 HCM predictions; *bottom*, recommended model predictions (*1*).

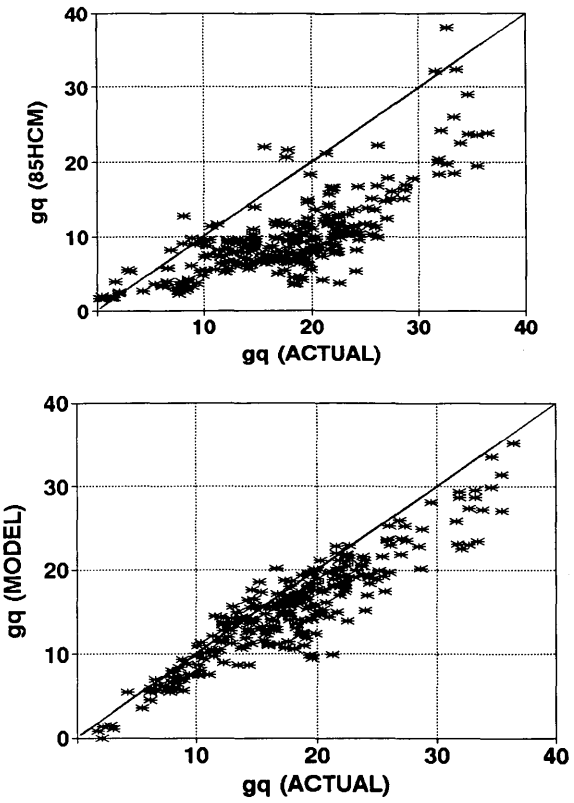


FIGURE 4 Comparative accuracy in prediction of g_q: *top*, 1985 HCM predictions; *bottom*, recommended model predictions (*1*).

It might be argued that a moving platoon without gaps has as much impact on left turns as the clearance of a standing queue. The blocking effect of a moving platoon of vehicles through a well-coordinated signal progression can be handled by including these vehicles in the opposing queue ratio.

Figure 4 illustrates the relative accuracy of these models in predicting g_q. When g_f and g_q are estimated, g_u is deduced as previously recommended.

$$g_u = \begin{cases} g - g_q & g_q \geq g_f \\ g - g_f & g_q < g_f \end{cases}$$

Develop Adjustment Factor for Left Turns Made Through Opposing Queue on Single-Lane Approaches

The theoretical structure of the HCM model does not allow for left turns made within g_q, the time during which the opposing queue clears the intersection. For single-lane cases, this is not correct, because left turns may be made through gaps created by opposing left turns. Thus, during the critical interval between the arrival of the first left turn and the clearance of the opposing queue (i.e., $g_q - g_f$), for which the HCM model assumes no flow in the shared lane, left turns may in fact occur, and flow in the shared lane is therefore not only possible, but probable.

It is therefore necessary to construct a model for an adjustment factor to be applied during this period. During g_u, the adjustment factor is

$$f_1 = 1/[1 + P_L(E_L - 1)]$$

where E_L is the through-car equivalent of a left-turning vehicle.

What is needed for single-lane cases is a model for E_{L2}, a through-car equivalent for a left-turning vehicle during the $g_q - g_f$ period. Such a model may be developed relatively simply. At g_f, there is, by definition, a left-turning vehicle waiting for service. This vehicle will wait until a gap in the opposing queue occurs as a result of an opposing left-turning vehicle. The expected value of the waiting time for this vehicle is found by considering the probability of the first, second, third, and so on opposing vehicle being a left turner. In the limit, no opposing left turner occurs during the $g_q - g_f$ period. The expected waiting time of the left-turning vehicle arriving at g_f is

$$\sum_{i=0}^{n} [2\ i\ P_{THO}^{i-1} P_{LTO}] + [(g_q - g_f)\ P_{THO}^{n}]$$

where

P_{LTO} = proportion of left-turning vehicles in opposing flow;
P_{THO} = proportion of through vehicles in opposing flow;
n = maximum number of opposing vehicles that arrive during period $g_q - g_f$, taken as $(g_q - g_f)/2$; and
2 = assumed saturation headway for opposing vehicles.

The first term of the equation accounts for the probabilities of an opposing vehicle arriving among the n opposing vehicles during $g_q - g_f$. The second term accounts for the probability that all opposing vehicles during this time are through vehicles.

A value of E_{L2} is estimated by dividing this expected waiting time by the saturation headway of 2 sec. Doing this and algebraically clearing the series results in

$$E_{L2} = (1 - P_{THO}^{n})/P_{LTO}$$

which can easily be converted to a multiplicative adjustment factor. This simple approach ignores the impact of left-turning vehicles arriving after g_f, that is, a second, third, or fourth left turner. These would have a smaller impact per vehicle because their maximum wait time is reduced. Thus, applying this single value of E_{L2} to all of the left turns occurring during the $g_q - g_f$ period is a conservative approach.

Implement Regression Model for f_{LT}

One of the key conclusions of the FHWA study was that there was a nonnegligible impact of left turns from a shared lane on adjacent lanes in a multilane group. Rather than arbitrarily assigning a value of 1.00 for the left-turn factor to adjacent through lanes, the study suggested the following:

$$f_{LT} = [f_m + a(N - 1)]/N$$

where N is the number of lanes in the lane group, and a is the factor applied to through lanes in the lane group.

Whereas the study attempted to establish an algorithmic expression for a on the basis of the proportion or number of left turns in the shared lane, the best fit to data from 192 study periods, each of 15 min, on multilane sites occurred when $a = 0.91$, a constant.

Remove Sneakers from Model

During the research for FHWA by Polytechnic, it was observed that the number of sneakers completing the turn during the clearance interval was much smaller than that predicted by the 1985 HCM. In addition, these vehicles were being double counted because vehicles are counted as they cross the stopline. For the recommended model, therefore, the procedure does not represent sneakers, and the sneaker term is dropped from the equation.

Use Messer-Fambro Left-Turn Equivalent

E_L, in the current manual is calculated on the basis of the ratio of the ideal saturation flow rate and a simple model for the saturation flow rate of left turns through an unsaturated opposing flow, 1,800 (1,400 $- v_o$). This model assumes that the left turn equivalent is the same for a left turning vehicle crossing a given volume per hour, whether that volume is in one, two, or three or more lanes. In a paper by Messer and Fambro (*4*), more realistic values for E_L were developed that account for the effect of the number of opposing lanes. Table 1 gives these values and has been incorporated into the procedures.

Increase Heavy Vehicle Equivalency

An additional change recommended is to the heavy vehicle equivalency value used in determining F_{HV}. The manual now uses a heavy vehicle equivalency of 1.5. The authors have suggested changing this value to 2.0. Zegeer (*5*) conducted surveys to study the effects of heavy vehicles on headways that suggested a heavy vehicle passenger car equivalent of 1.92.

TABLE 1 Through-Car Equivalents, E_L, for Permitted Left Turns

No. of Signal Phases	Type of Left Turn Lane	No. of Opposing Lanes	E_L by Opposing Flow, V_o[1]				
			200	400	600	800	1000
2-PHASE	Shared	1	2.0	3.3	6.5	16.0*	16.0*
		2	1.9	2.6	3.6	6.0	16.0*
		≥ 3	1.8	2.5	3.4	4.5	6.0
	Exclusive	1	1.7	2.6	4.7	10.4*	10.4*
		2	1.6	2.2	2.9	4.1	6.2
		≥ 3	1.6	2.1	2.8	3.6	4.8
Multiphase	Shared	1	2.2	4.5	11.0*	11.0*	11.0*
		2	2.0	3.1	4.7	11.0*	11.0*
		≥ 3	2.0	2.9	4.2	6.0	11.0*
	Exclusive	1	1.8	3.3	8.2*	8.2*	8.2*
		2	1.7	2.4	3.6	5.9	8.2*
		≥ 3	1.7	2.4	3.3	4.6	6.8

* Generally indicates turning capacity only available at end of phase - sneakers only.

[1] For the purpose of determining v_o and N_o, opposing right and left turns from exclusive lanes are not included in v_o, nor are the exclusive lanes in N_o.

Summary of Model for f_{LT}

Table 2 gives a summary of the recommendations and illustrates the model recommended for implementation.

COMPARING RESULTS OF SEVERAL METHODOLOGIES

Using 192 data periods (each 15 min) for two-lane sites and 102 data periods (each 15 min) for single-lane sites from the FHWA study, the relative accuracy of the recommended hybrid model was compared with that of the 1985 HCM model and several other suggested models.

The models tested include

- The recommended model, described in this paper,
- The 1985 HCM Model,
- Modifications to the 1985 HCM suggested by Messer,
- A model recommended by Lin in a paper presented at TRB in 1992 (*6*),
- A model recommended by Levinson in a paper presented at TRB in 1992 (*7*).

The Messer modifications have not been published, but they include revision of the starting assumption of opposing saturation flow rate to reflect nonideal conditions and carrying forward this modification to every use of either the opposing saturation flow rate or the opposing saturation headway in the model. A second modification involves the assumption that multilane opposing flow is not uniformly distributed across available lanes. A third modification adjusts the equation for g_f to account for no left turns arriving during g_q.

Table 3 gives the average error in predicting saturation flow rate and the percent error for each of the models. The data indicate that using an ideal saturation flow rate of 1,900 pcphgpl, the recommended model yields the best results for both single- and two-lane approaches.

Another finding of note is that the best of these models results in an average error of somewhat less than 20 percent of the observed saturation flow rate. The plots in Figure 5 help explain this. The variation in measured ideal saturation flow rates from the FHWA study are shown. Even when averages over 15-min intervals are considered, the variation in this value is extremely large. This phenomenon was evident in the study, with variations occurring from site to site, and even at the same site over time. For single-lane sites, average ideal saturation headways ranged from 1.8 sec/vehicle to more than 3.4 sec/vehicle. For multilane sites, values ranged from 1.6 sec/vehicle to more than 2.3 sec/vehicle.

Because all of the model structures considered in this paper operate on the basis of modifying a constant ideal saturation flow rate, the considerable actual variation in the base variable more than explains the approximately 20 percent average errors in predictions of prevailing saturation flow rate.

TABLE 2 Recommended Model

ITEM	Single Lane Approaches	Multilane Approaches
g_f	$G\, e^{-0.860\ LTC^{0.629}} - t_L$	$G\, e^{-0.882\ LTC^{0.717}} - t_L$
g_q	$4.943\ V_{olc}^{0.762}\ qr_o^{1.061} - t_L$	$9.532\ V_{olc}^{0.560}\ qr_o^{0.819} - t_L$
g_u	$g - g_q$ for $g_q \geq g_f$ $g - g_f$ for $g_q < g_f$	$g - g_q$ for $g_q \geq g_f$ $g - g_f$ for $g_q < g_f$
Period	Factor Applied - Single lane	Factor Applied - Multilane
$\frac{g_f}{g}$	1.00	1.00
$\frac{g_q - g_f}{g}$	$\frac{1}{1 + P_L(E_{L2} - 1)}$ $E_{L2} = \frac{1 - P_{THO}^n}{P_{LTO}}$	0.00
$\frac{g_u}{g}$	$\frac{1}{1 + P_L(E_L - 1)}$ where E_L is found in Table 3	$\frac{1}{1 + P_L(E_L - 1)}$
f_{LT}	$f_{LT} = f_m$	$f_{LT} = \frac{f_m + 0.91(N - 1)}{N}$

FINAL THOUGHTS AND RECOMMENDATIONS

Because of the base variation in ideal saturation flow rate, it is unlikely that any model developed by modification will result in much better accuracy than those examined in this paper. Given, then, the ballpark nature of the computation of prevailing saturation flow rate, the great complexity of

TABLE 3 Average Errors in Predicting Saturation Flow Rate for Various Left-Turn Models

Ideal Sat Flow (S_o)	Model	Single Lane Approaches (Avg Sat = 1300 vphg)		Two-Lane Approaches (Avg Sat = 2600 vphg)	
		Avg Error	% Error	Avg Error	% Error
1900	1985 HCM	355	27%	569	22%
	Messer	401	31%	509[1]	20%
	Lin	357	27%	966	37%
	Levinson	354	27%	760	29%
	RECOM	289	22%	496	19%

[1] Application of multilane factor "a" shown here; without "a" error = 590 vph

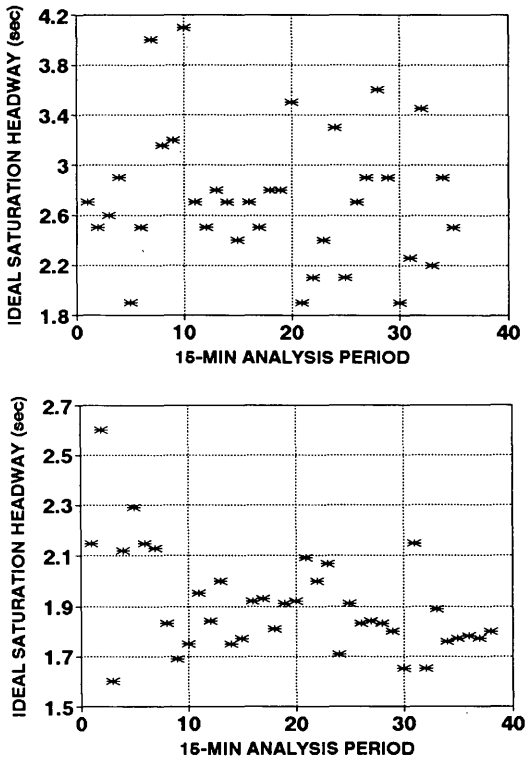


FIGURE 5 Observed variation in ideal saturation flow rate: *top*, single lane approaches; *bottom*, multilane approaches (*1*).

some models is not justified. Unless complexity yields significant benefits in accuracy, it becomes a liability in users' ability to comprehend and properly apply the procedure.

The procedure presented in this paper has the following advantages:

- The circular logic of the HCM, requiring a "guesstimate" of saturation flow rate in one direction to compute in the other is eliminated.
- The regression models for g_f and g_q yield significantly better estimates of these variables than the HCM model, and therefore introduce more appropriate sensitivities to these variables into the overall model.
- The recommended model results in saturation flow estimates that are better than that produced by other models.
- The use of the recommended model simplifies computations for f_{LT}.
- The recommended model explicitly corrects several difficulties in the HCM model, particularly for single-lane approaches.
- The recommended model retains the basic theoretic structure of the HCM model, and is a rational depiction of driver behavior in permitted left-turn situations.

REFERENCES

1. R. P. Roess, V. Papayannoulis, J. M. Ulerio, and H. Levinson. *Levels of Service in Shared-Permissive Left-Turn Lane Groups*. Transportation Training and Research Center, Polytechnic University, New York, Sept. 1989.
2. R. P. Roess, J. M. Ulerio, and V. Papayannoulis. Modeling the Left-Turn Adjustment Factor for Permitted Left Turns Made from Shared Lane Groups. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
3. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
4. C. Messer and D. B. Fambro. Critical Lane Analysis for Intersection Design. *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977.
5. C. Zegeer. Field Validation of Intersection Capacity Factors. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986.
6. F. B. Lin. Left-Turn Adjustment Factors for Saturation Flow Rates of Shared Permissive Left-Turn Lanes. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992.
7. H. Levinson. The Capacity of Shared Left Turn Lanes: A Simplified Approach," Presented at the 71st Annual Meeting of the Transportation Research Board, National Research Council, Washington, D.C., 1992.

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

Overflow Delay at a Signalized Intersection Approach Influenced by an Upstream Signal: An Analytical Investigation

ANDRZEJ TARKO, NAGUI ROUPHAIL, AND RAHMI AKCELIK

Some initial formulations of general overflow delay formulas that can be applied to isolated intersection approaches as well as those contained in a signalized arterial network are discussed. The main issues addressed are the deficiencies in models of arrival that have been developed to account for the filtering effect of upstream signals, which use the Poisson arrival process as their foundation. A cycle-by-cycle simulation model is the investigation tool. This level of modeling allows for the estimation of the uniform and overflow delay components separately. The results of the model are first tested for the isolated intersection scenario. The simulation model results are then extended to a two-intersection system, and a generalized delay model is calibrated to encompass both the isolated and the system cases. The initial work is limited, however, to fixed-time signal control with no platoon dispersion and no secondary or midblock flows between the two approaches. Only single-lane flow cases are considered. It was found from the simulation results that the inclusion of a parameter X_o (a value of the degree of saturation below which the overflow delay is negligible) in the delay model is justified. The study also confirmed that progression quality has no effect on the overflow delay estimate. Two competing overflow delay model forms were investigated; one used the variance-to-mean ratio of upstream departures and the other used the capacity differential between intersections to reflect the filtering effect. Whereas the first form has been widely quoted and advocated in the literature, the second approach provided better predictions in this study.

Vehicular delay at signalized intersections is a critical component of travel time in an urban road network. Delay contributes to vehicle operating costs and is intimately affected by the traffic engineer's decisions. The *Highway Capacity Manual* (HCM) (*1*) uses delay as the sole criterion for determining level of service at signalized intersections and indirectly for level-of-service evaluation on urban arterials.

Of the earlier delay formulas, Webster (*2*) applied steady-state queueing theory augmented by simulation to evaluate the overflow delay component. Several general formulas have since been proposed that use an index of the dispersion of the arrival and departure processes [see e.g., Darroch (*3*) and Gazis (*4*)]. None of the models, however, is directly related to coordinated signalized intersections.

To overcome the difficulties encountered with steady-state queueing models (i.e., infinite delays at capacity), time-dependent delay models were originally conceived by Robertson (*5*) and were elaborated further by Kimber and Hollis (*6*) using the so-called coordinate transformation method. Although there is no rigorous theoretical basis for this approach (*7*), empirical evidence indicates that these models yield reasonable results. A number of time-dependent formulas have found their way into the capacity guides of several countries, including the United States (*1*), Canada (*8*), and Australia (*9*).

This brief overview points to the importance of providing rational delay estimates for internal signals in a network. The work reported on in this paper follows on from the initial work reported in Rouphail and Akcelik (*10*). Initially, the assumptions are kept simple to be effective. Hence, the model will be investigated for a single traffic stream flowing between two closely spaced intersections (i.e., no platoon dispersion or minor flows). The two signals operate at a fixed-cycle length, but could vary in splits and offsets.

The paper is organized as follows. First, a brief discussion of the delay model form is presented and current attempts to model overflow delay at isolated and coordinated signals are discussed. Next, the simulation methodology is described, along with validation results from the single intersection case. The extension to the two-intersection system is discussed next. Two general model forms are then presented to describe the filtering effect of the upstream signal on downstream delay: the first uses the variance-to-mean ratio of upstream departures (equivalent to downstream arrivals), I; the second uses a function of the capacity differences at the two intersections. Both model forms are compared and evaluated. The paper concludes with possible extensions of the concept to multiphase, multistream cases, and further research needs in that area are briefly highlighted.

ANALYTICAL DELAY EXPRESSIONS

The average approach delay per vehicle at a signalized intersection approach can be expressed as the sum

$$d = d_1 + d_2 \quad (1)$$

where d_1 is the uniform delay component, which refers to the average vehicle delay experienced assuming that traffic demand is the same for all signal cycles, and d_2 is the overflow delay component, which consists of both the random and time-dependent oversaturation effects for a sustained time period, T.

A. Tarko and N. Rouphail, Urban Transportation Center, University of Illinois at Chicago, 1033 West Van Buren, Suite 700 South, Chicago, Ill. 60607. R. Akcelik, Australian Road Research Board, P.O. Box 156, Nunawading 3131, Australia.

The first delay component d_1 is known as Webster's first term and is given by

$$d_1 = \frac{C(1 - \lambda)^2}{2(1 - \lambda \cdot X)} \quad (2)$$

where

λ = effective green to cycle ratio,
X = volume-to-capacity ratio, and
C = cycle length.

Several delay models have been derived with the assumption of steady-state conditions (*2*). These models estimate the queue length under stochastic equilibrium conditions. The queue equilibrium can be attained only if the prevailing traffic volume and capacity are stationary for an indefinite period of time (in practice, however, it need only be sufficiently long) and if the expected flow rate is below capacity.

The coordinate transformation technique converts steady-state queueing models to time-dependent models (*9*). This process allows the relaxation of the steady-state model restrictions and the extension of the delay model to oversaturated conditions. Akcelik (*11*) proposes a generalized time-dependent expression in the form

$$d_2 = 900TX^n \left[X - 1 + \sqrt{(X-1)^2 + \frac{m(X - X_o)}{QT}} \right] \quad (3)$$

where

Q = capacity [vehicles (veh)/hr]
X_o = volume-to-capacity ratio below which overflow delay is negligible.

and

$$X_o = a + bsg$$

where

s = saturation flow rate (veh/sec),
g = effective green, and
a, b = parameters.

The delay model parameters a,b,n, and m depend on the distribution of arrivals and departures.

Model parameters as presented by Arcelik (*11*) are given in Table 1.

CURRENT STATUS

Apart from the Australian model given in Table 1, all other expressions are applicable solely to isolated intersections. The Australian model includes the platooning effect in a rather rough way, which is partly supported by empirical evidence gathered by Hillier and Rothery (*12*). They found that for internal signals (a) overflow delay was insensitive to offset variations (see section on *Offset Impact on Overflow Delay*) and (b) delay was consistently lower than that observed at isolated intersections. Although a delay formula that reflected these findings was used in earlier versions of the TRANSYT model, the current TRANSYT model allows for the first finding, but not for the second. In the 1985 HCM, the overflow delay term is allowed to vary according to progression quality, which is contrary to the empirical evidence. Recently, empirical work by Fambro and Messer (*13*) and theoretical analysis by Olszewski (*14*) have independently confirmed the fact that progression and platooning effects are limited to the first term, d_1. Interestingly, Fambro and Messer and Chodur and Tracz (*15*) also found that the 1985 HCM formula appears to overestimate the observed overflow delays in coordinated signal cases. This was further confirmed with recent empirical data in South Africa by Van As (*16*).

Along the line of previous theoretical work by Darroch (*3*), Gazis (*4*), and Hutchinson (*17*), Rouphail and Akcelik (*10*) developed and applied a simulation model for calibrating an overflow delay model for a two-intersection signal system. The model incorporates fixed-saturation flow rates and an index of dispersion for the arrival process (I = variance-to-mean of arrivals per cycle). As shown by Olszewski (*14*) the assumption of fixed-saturation flow is acceptable for unopposed vehicle streams. Van As developed an approximate method for estimating I.

A recent paper by Newell (*18*) proposes an interesting hypothesis. In it, the author questions the validity of using overflow delay expressions derived for isolated intersections at internal signals in an arterial system. Newell goes on to suggest that the sum of overflow delays at all intersections in an arterial system with no turning movements is equivalent to the overflow delay at the critical intersection, assuming that it is isolated. In essence, Newell proposes that the use of delay formulas designed for isolated intersections on arterials results in a gross overestimation of such delays on the main road. If signals are timed to minimize delays, then the resulting controls may be very inefficient, particularly for side streets.

It is evident that an increasing body of research suggests that a fundamental difference exists between overflow delay

TABLE 1 Calibration Parameters for d_2 in Selected Capacity Guides

Method	Model Parameters[a]			
	n	m	a	b
1985 HCM	2	4	0	0
Australian	0	12(6)[b]	0.67	1/600
Canadian	0	4	0	0
HCM (Akcelik)	0	8	0.50	0

[a]Source: Akcelik (11).
[b]Value 6 When Platooning Occurs.

estimates at isolated and internal signals. What is lacking, however, is a systematic analysis of the relationship between signal system design parameters (such as cycle, upstream capacity, or offset) and overflow delay at a downstream approach. The goal of this research was to investigate the feasibility of a generalized overflow delay model applicable for both internal and external links in a system.

METHODOLOGY

A cycle-by-cycle simulation model developed by Rouphail and Akcelik (*10*) has been used to generate the data base to be used for model development. Subsequently, multiple regression techniques (linear and nonlinear) were applied for model calibration in the form given by Equation 3. The simulation model is described next, then some issues regarding analytical model calibration are discussed.

Simulation Model Description

A discrete, macroscopic cycle-by-cycle simulation approach was adopted for modeling a two-intersection system. Vehicles are represented as individual entities, but delays are computed for groups of vehicles that have the same headway properties (hence the macroscopic designation). Delay calculations extend the original work of Staniewicz and Levinson (*19*). Following is the list of assumptions and limitations of the model:

1. Assumptions regarding the arrival process at the upstream intersection:
   - *a*. The number of arrivals per cycle is generated according to the Poisson distribution; and
   - *b*. The arrival pattern within a cycle is strictly uniform, with the arrival headway equal to the cycle length divided by the number of generated arrivals (which varies each cycle).
2. Assumptions regarding the departure process at the upstream intersection:
   - *a*. There are no departures during the effective red time;
   - *b*. The vehicles in the queue (earliest departure time exceeds arrival time) depart at a constant saturation headway;
   - *c*. The vehicles arriving after the queue has discharged cross the stopline at the arrival headway; and
   - *d*. If a cycle is overloaded, overflow vehicles are released first-come, first-served in subsequent cycles.

The arrival process at the downstream intersection is equivalent to the departure profile at the upstream intersection shifted by the cruise travel time between stoplines. This is the consequence of the following assumptions:

- All vehicles travel between stoplines at the same speed, immediately accelerating and decelerating at the stopline when delayed and
- There are no midblock vehicle or pedestrian flows.

The departure process at the downstream intersection follows the same principles as those given for the upstream approach. Model parameters for signal control are cycle length, effective green times, and signal offset. Parameters for traffic flow are traffic volume, saturation flow rates, and cruise time.

Figure 1 is a schematic representation of the model. The structure of the model allows for the estimation of all delays for two arrival types—random arrivals upstream and platooned arrivals downstream. More important, the simulation provides an option to generate a fixed number of arrivals in each cycle, which is concomitant with the uniform delay component, d_1 (Equation 1). Note that for the upstream approach, d_1 is equivalent to Webster's first term (Equation 2). Downstream uniform delays on the other hand are very much tied to progression quality (offset). The ability to generate each delay component separately at each intersection is a key strength of the simulation model.

Delay Estimation Process

All runs were performed assuming steady-state queueing conditions (i.e., T = infinity in Equation 3). This required that (a) a sufficiently large number of cycles be simulated and (b) the v/c ratio not approach 1. The closer v/c is to 1, the longer the simulation time period required. The simulation model was run for 100 cycles (2 to 3 hr), which allowed the investigators to neglect the initial nonsteady conditions and to cut off the overflow delay at the end of the simulation period. The resulting bias is estimated at less than 0.4 percent for the assumed maximum value of v/c = 0.9. After a Poisson run was completed, a tandem run was initiated using the average demand from the first run. In fact among the 100 simulated runs in the second case, some had n arrivals, while others had $n + 1$ arrivals, to produce the desired average per cycle. The difference in the average vehicle delay produced by the two tandem runs is, by definition, the overflow delay component, as defined in Equation 1.

The coordinate transformation technique was subsequently used to enable model evaluation by comparison with other time-dependent formulas currently used in various countries.

Model Calibration

A generalized overflow delay model in the steady-state form, based on Akcelik (*11*) was calibrated from simulation. The model form is

$$d_2 = \frac{k[X - (a + bsg)]I}{Q(1 - X)} \tag{4}$$

where

X = degree of saturation,
s = saturation flow (veh/sec),
g = effective green (sec),
I = variance-to-mean ratio of an arrival process,
Q = capacity (veh/sec), and
k, a, b = model parameters (in Equation 3, $m = 8k$).

For the purpose of calibration using regression analysis, Equation 4 is rewritten as

$$d_2 = k\frac{XI}{Q(1 - X)} + ka\frac{-I}{Q(1 - X)} + kb\frac{-sgI}{Q(1 - X)} \tag{5}$$

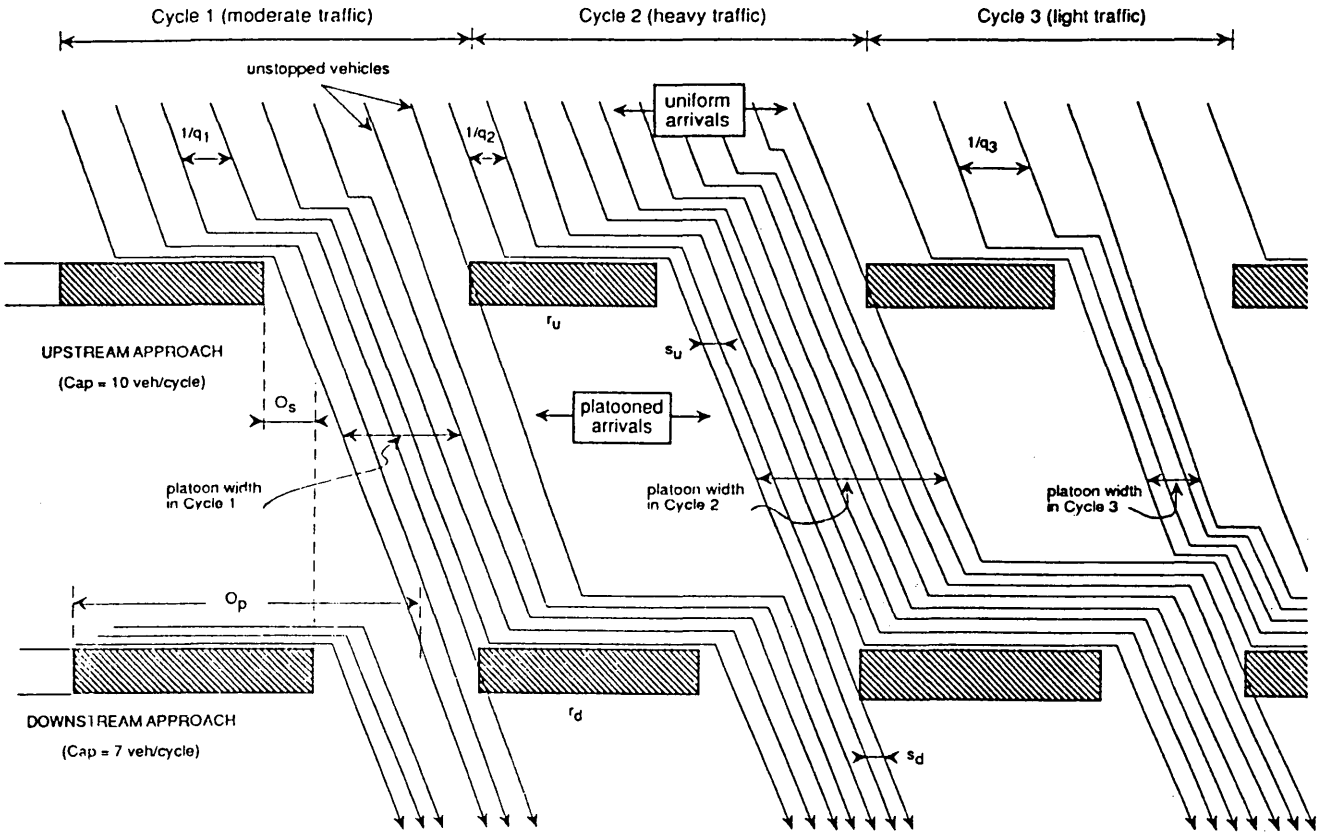


FIGURE 1 Schematic of simulation model concept.

in which the terms k, ka, and kb are regression parameters. The calibration data set for the isolated case consisted of 480 simulation runs of the following combinations:

- Cycle length (C): 80, 100, and 120 sec;
- Effective green-to-cycle ratio (g/C): 0.1, 0.3, 0.5, and 0.7; and
- Degree of saturation (X): 0.5, 0.7, 0.8, and 0.9.

The simulation data set corresponds to the reasonably wide range of single lane capacities (sg = 4 to 42 veh/cycle and s = 0.5 veh/sec).

Model Evaluation

The two simulated delay components in Equation 1 were evaluated independently. The simulated uniform component is shown in Figure 2 against Webster's first term. For all practical purposes, the differences are minimal. A systematic bias of about 0.5 sec was found to result from the discrete nature of the simulation model.

The calibrated overflow delay component based on Equation 4 is expressed by the model (R^2 = .912)

$$d_2 = \frac{0.456I[X - (sg/100)]}{Q(1 - X)} \quad (6)$$

Note that the I value in the simulation varies about 1.0 from one run to the next. For practical purposes, I = 1 for isolated intersections. Hence

$$d_2 = \frac{0.456[X - (sg/100)]}{Q(1 - X)} \quad (7)$$

A comparison of simulation results with Equation 7 estimates is shown in Figure 3. The resultant overflow delay model has parameters that are different from those listed in Table 1. Because of multicollinearity effects among parameters, how-

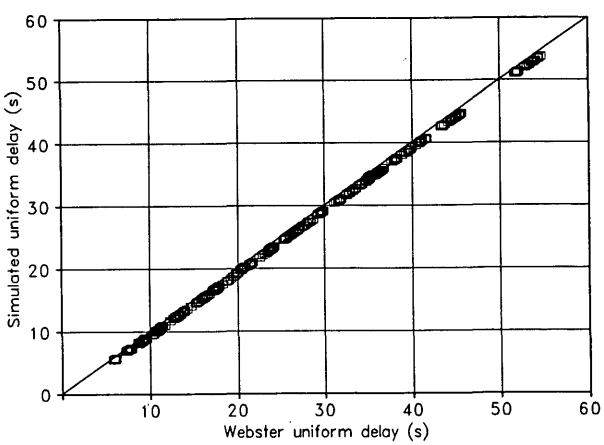


FIGURE 2 Comparison of uniform delays—simulated and Webster first term.

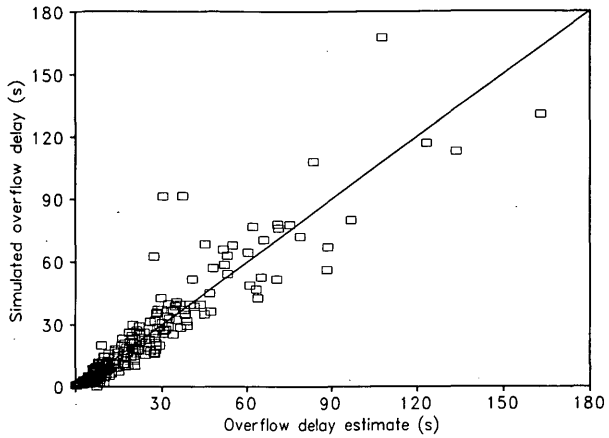


FIGURE 3 Comparison of simulated overflow delays and model M3 estimates—steady-state conditions.

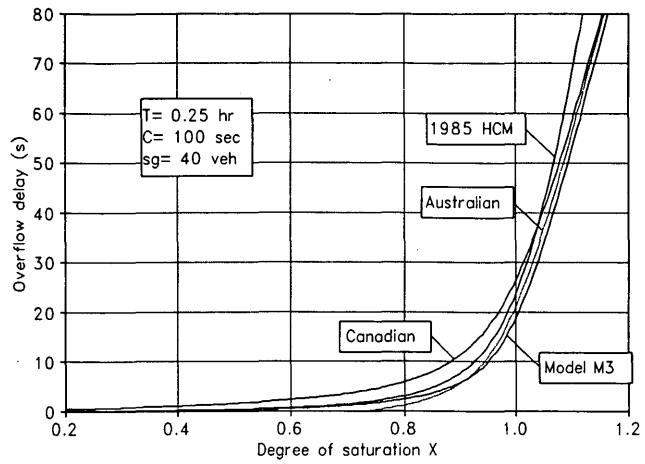


FIGURE 5 Comparison of model M3 estimates with existing models—high cycle capacity.

ever, a better test is to directly compare the time-dependent form of the delay model in Equation 7 (herein termed the M3 model) with those listed in Table 1 (except for Akcelik's HCM formula). This is shown in Figures 4 and 5 for cycle capacities (sg) of 10 and 40 vehicles, respectively. The graphs indicate that when both cycle capacity and degree of saturation are low, the results given by the M3 model are almost identical to the HCM formula. The differences between the same two methods become more noticeable as X increases. For high X's and low cycle capacity, the model tracks the Canadian formula. In Figure 5 (high cycle capacity) the model tracks the Australian formula. By virtue of the coordinate transformation method, the M3 model (along with the Australian and Canadian models) becomes asymptotic to the deterministic delay model, $d = 450\,(X - 1)$. Overall, the model calibrated from the simulation compares very favorably with well-established overflow delay models cited in the literature.

EXTENSION TO TWO-INTERSECTION SYSTEM

Two additional factors have been included in this extension—signal offset between the two approaches and variations in upstream and downstream capacity. The first is related to the cyclic arrival pattern, and the second is concerned with the magnitude of arrivals per cycle as well as cycle-to-cycle fluctuations (see Figure 1). The calibration data set is described as follows:

- Cycle length (C): 80, 100, and 120 sec;
- Downstream approach degree of saturation (X): 0.6, 0.8, and 0.9;
- Downstream effective green times (g): 24 and 36 sec for C = 80, 44 sec for C = 100, and 36 and 54 sec for C = 120 (sg = 12 to 27 veh/hr); and
- Upstream effective green g_u varied from g up to a maximum value that is cycle-dependent.

The data reflect the tendency of coordinated phases to dominate the green allocation between coordinated and uncoordinated movements. The obvious case of zero delay when the upstream capacity is lower than downstream is excluded from the data set (e.g., $sg_u < sg$).

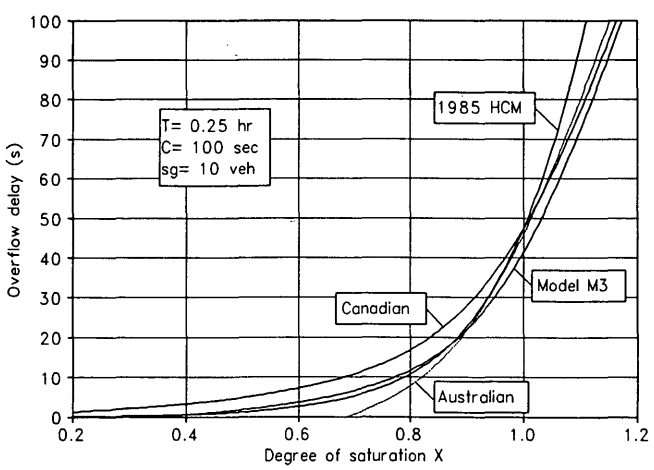


FIGURE 4 Comparison of model M3 estimates with existing models—low cycle capacity.

Offset Impact on Overflow Delay

The literature is replete with models that relate progression quality to uniform delay [see Fambro (*13*)]. To confirm previous findings reported in the literature, which show no correlation between overflow delays and offsets (*12*), a separate simulation experiment was conducted. In it C = 100 sec, g = 44 sec, g_u = 52 sec, cruise time = 20 sec, and X = 0.90. The model was run for each offset between zero and cycle length in intervals of 10 sec. A sample of the results is shown in Figure 6. As suspected, the overflow delay was highly variable (note X = 0.90) and independent of offset. On the other hand, the uniform term was very sensitive to offset selection. The minimum delay was attained at and around the travel time offset of 20 sec.

Steady-State Delay Model Calibration

Using Departing Stream I Ratio

By applying the same concepts described in the section, *Model Evaluation*, for the downstream approach as suggested in the

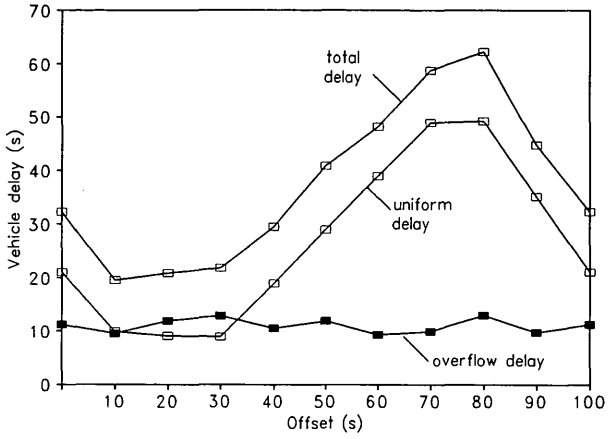


FIGURE 6 Simulated delays against signal offset.

literature (*17,19,20*), the steady-state form of the overflow delay model in this case (note $R^2 = 0.738$) is

$$d_2 = \frac{1.02\{X - [0.55 + (sg/300)]\}I}{Q(1 - X)} \quad (8)$$

The differences between the models (Equations 6 and 8) are partly explained by the different calibration data sets. The isolated intersection case included some data points with very low cycle capacity (4 to 6 veh/cycle) and consequently yielded much higher overflow delays than in other cases. Twelve vehicles per cycle is the lowest capacity value considered for coordinated intersections. Further, the upstream signal impact may not have been appropriately represented through the single index, I.

In general, the R^2 in the model (Equation 8) is smaller than for the isolated case, although the exact value of I (simulated) was used. This has been one of the reasons for seeking another form of the overflow delay model. Comparison of the calculated delays in the model (Equation 7) with the simulated values revealed some bias for small values of delay.

Figure 7 illustrates the variation of overflow delay and the coefficient I with the difference between upstream and downstream capacities ($sg = 20$ veh/cycle, $C = 120$ sec, and $X = 0.85$). As suspected, overflow delay vanishes when the upstream capacity is less than or equal to the capacity at the downstream intersection. Clearly, this is the result of imposing an upstream constraint, which prevents the downstream intersection from cycle failure. On the other hand, overflow delay is found to be virtually fixed when the capacity difference is large (as the upstream capacity increases). In this case the downstream intersection can be considered to be isolated (from an overflow delay standpoint). The "smooth" transition from the coordinated to the isolated case (at high capacity differentials) implies that distinguishing between the two cases is artificial, especially because the offset impact is nonexistent.

Recall that attempts to express the upstream constraint by means of the coefficient I in the overflow delay formula have been undertaken earlier by Hutchinson (*17*) and Van As (*16*). As shown in Figure 7, the coefficient I and overflow delay reach a zero value at different coefficients on the x-axis, although their trends are similar. For traffic arriving downstream, the coefficient I approaches zero when the upstream approach is close to saturation. On the other hand, overflow

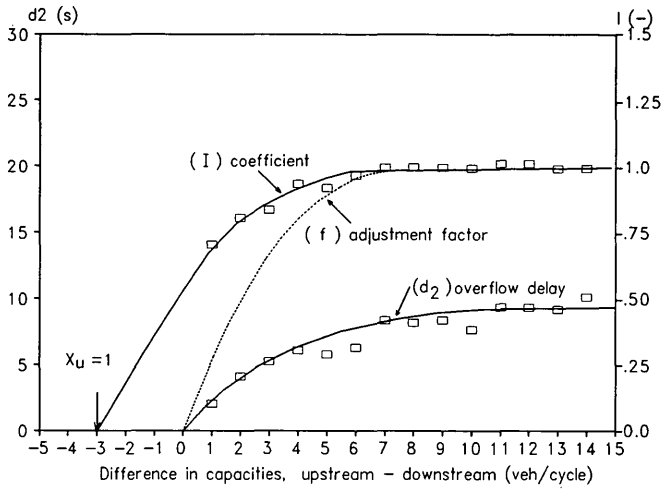


FIGURE 7 Index I, overflow delay, and adjustment factor f versus differential capacity—steady-state conditions.

delay vanishes when the upstream capacity is less than the downstream capacity. Figure 7 illustrates the situation when the downstream degree of saturation is less than unity and the coefficient I approaches zero ($X_u = 1$) on the left side of the point at which overflow delay vanishes. Clearly this is not desirable for estimating a valid d_2 value using Equation 8.

Using Capacity Differential Method

Taking into account these facts and the practical difficulties with the estimation of I, another more general overflow delay formula is suggested. The form of the analytical model for overflow delay (Equation 4) is retained, but without the coefficient I, as follows:

$$d_2 = \frac{k(X - X_o)}{Q(1 - X)} \quad (9)$$

where

$k = k_o f$, with k_o = slope parameter for the isolated case (k in Equation 4, and f = adjustment factor for upstream conditions that is expressed as a function of the difference between the upstream and downstream capacities, and
$X_o = b(sg)$ according to results obtained for the isolated case (Equation 6).

The function $f = f[(sg)_u - (sg)]$ shown in Figure 7 should meet the requirements

$f = 1$ when $(sg)_u >> (sg)$,
$0 < f < 1$ when $(sg)_u > (sg)$, and
$f = 0$ when $(sg)_u < (sg)$

where $(sg)_u$ and (sg) are the upstream and downstream capacities in vehicles per cycle, respectively.

The following functional form, which is consistent with Figure 7, has been selected for calibration:

$$f = \begin{cases} 1 - e^{-\alpha \cdot [(sg)_u - (sg)]} & (sg)_u > (sg) \\ 0 & \text{elsewhere} \end{cases} \quad (10)$$

where α is a calibration parameter. To achieve the generalized model form, the data base for model calibration included simulated data from both isolated and coordinated cases. Because of limited computer memory, one-half the number of data points were randomly drawn from the original data sets before combining. The calibration data set thus contained 738 observations.

Multiple linear regression was used to determine the best estimates of parameters k_o and b. Because of the nonlinear form of the function given by Equation 10, the parameter α cannot be calibrated directly. The optimal value $\alpha = 0.5$ was determined from a series of tests which maximizes the R^2 value ($R^2 = 0.819$). The resulting model has the form

for $(sg)_u > (sg)$ and $X > (sg)/100$, and 0 elsewhere.

Equation 11 provides calibrated values of all delay parameters as follows:

$$d_2 = \frac{0.408\{1 - e^{-0.5[(sg)_u - (sg)]}\}[X - (sg/100)]}{Q(1 - X)} \quad (11)$$

where

$k_o = 0.408$,
$f = 1 - \exp[-0.5(sg_u - sg)]$,
$X_o = sg/100$.

A comparison of the analytical delay estimates with simulated values is shown in Figure 8 in which f values calculated from Equation 10 and $\alpha = 0.5$ are compared with the simulated values of I. It is clear that the factor f provides a more appropriate means for estimating the overflow delay when compared with the coefficient I.

Time-Dependent Delay Model

Transformation of the steady-state model (Equation 11) into a time-dependent formula is carried out by setting $n = 0$ and using the relationship $m = 8k$ in Equation 3 (*11*) with $k = k_o f$. The parameters k_o, f, and X_o are given by Equation 11. Furthermore, m_c was substituted for $(sg)_u$

$$d_2 = 900T\left[X - 1 + \sqrt{(X-1)^2 + \frac{3.3[1 - e^{-0.5(m_c - sg)}][X - (sg/100)]}{QT}}\right] \quad (12)$$

where

T = time period (hr),
X = degree of saturation,
Q, sg = lane capacity (veh/hr, veh/cycle, respectively), and
m_c = maximum number of arrivals per cycle.

Because of the imposed constraint on demand represented in Equation 12 by m_c, the observed degree of saturation at the analyzed approach cannot exceed value $m_c/(sg)$. Examples of values given by the time-dependent formula are shown in Figure 9 ($Q = 900$ veh/hr, $C = 80$ sec, and $T = 0.25$ hr). The impact of the constrained demand disappears rapidly with an increase in the difference between upstream and downstream capacity. When the difference approaches 6 veh/cycle, the resulting overflow delays approach those occurring at an isolated intersection.

IMPLEMENTATION CONSIDERATIONS

The analytical investigation of overflow delay models presented was aimed at finding a relatively simple model that could incorporate the upstream signal impact. The results obtained should be considered preliminary results. Simplifying assumptions in the simulation model limit the direct implementation of the analytical model (Equation 12) to cases for which an unopposed single vehicle stream is controlled by

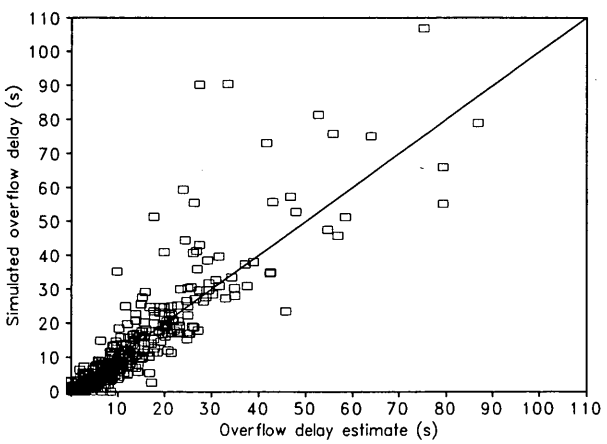


FIGURE 8 Simulated and analytical model overflow delay estimates—steady-state form.

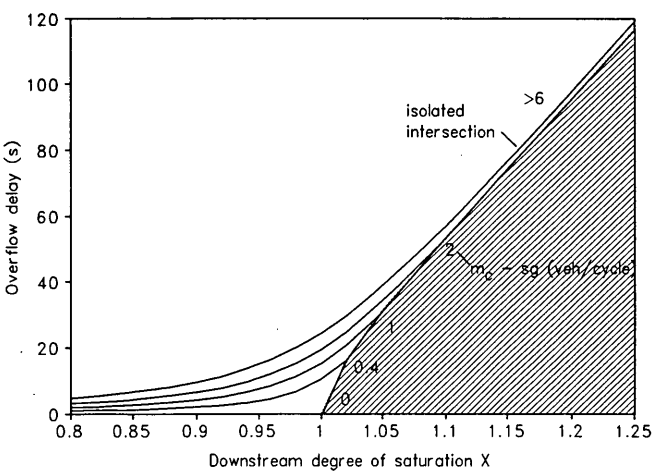


FIGURE 9 Time-dependent overflow delay estimates for various capacity differentials and degrees of saturation (overflow delay is zero when the downstream capacity is larger than upstream capacity).

a downstream pretimed signal and filtered upstream by a bottleneck. The bottleneck is located close to the signal and is characterized by a fixed capacity (expressed in vehicles per downstream signal cycle). This may describe a pedestrian signalized crossing or an unsignalized bottleneck (e.g., traffic lane closure or parking vehicles) located upstream to the subject approach at a signalized intersection.

Model expansion to more general cases requires the investigation of the following issues:

- How turning movements at the upstream signals and unsignalized midblocks contribute to the total constraint on arrivals at the downstream signal;
- How variations in capacity of minor streams (e.g., filter left turns) should be incorporated in the model (Equation 11) (a substantial short-term variability in capacity of opposed streams precludes the use of its mean value in the strongly nonlinear function, Equation 11); and
- To what extent does cruise time variance influence the distribution of vehicle arrivals at the downstream intersection, particularly with regard to the maximum number of arrivals.

These factors should be incorporated into the model through the adjustment of the maximum number of arrivals, m_c, or the factor, f, itself.

CONCLUSIONS

The analysis has shown that X_o, proposed by Akcelik for the overflow delay model, is justified. In particular, its use is justified for the case of low capacity, as in a lane-by-lane analysis in SIDRA (*21*).

Overflow delay is independent of progression quality, even in the case of no platoon dispersion when the effect is expected to be the strongest. Thus, there is justification for adjusting only the first component of the HCM delay model with a *PF* factor, as suggested by Fambro (*13*). The delay overestimation with the HCM formula at degrees of saturation higher than 1.2 has been confirmed.

When the upstream signal meters traffic flows, the overflow delay at the downstream intersection can decrease dramatically. Although this metering effect is reflected in the variance-to-mean ratio of arrivals at the downstream signal, its use can cause significant delay overestimation. The model proposed in this paper, which uses differential capacity measure, appears to be more appropriate for overflow delay estimation.

Traffic flow metering implies the existence of a maximum degree of saturation that can be observed at a downstream intersection. In the control optimization methods for signalized networks (e.g., TRANSYT), this fact is unfortunately neglected.

The model developed should be considered a preliminary model. It needs to be verified in field studies. Moreover, the inclusion of additional factors into the delay model (e.g., turning movements, midblock flow sources, or cruise time variations) would enhance its usefulness for practical implementation.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. F. V. Webster. *Traffic Signal Settings.* Technical Paper 39. Road Research Laboratory, HMSO, London, England, 1958.
3. J. N. Darroch. On the Traffic-Light Queue, *Annals of Mathematical Statistics*, Vol. 35, 1964, pp. 380–388.
4. D. C. Gazis. *Traffic Science.* Wiley-Interscience, New York, N.Y., 1974, pp. 148–151.
5. D. I. Robertson. Traffic Models and Optimum Strategies of Control. *Proc., International Symposium on Traffic Control Systems*, 1979, pp. 262–288.
6. R. Kimber and E. Hollis. *Traffic Queues and Delays at Road Junctions.* Laboratory Report 909. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1979.
7. V. F. Van Hurdle. Signalized Intersection Delay Models—A Primer for the Uninitiated. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984, pp. 96–104.
8. S. Teply. Quality of Service in the New Canadian Capacity Guide. *Proc., International Symposium on Highway Capacity*, Karlsruhe, Germany, July 1991, pp. 377–386.
9. R. Akcelik. Traffic Signals: Capacity and Timing Analysis. Research Report 123. Australian Road Research Board, 1981.
10. N. M. Rouphail and R. Akcelik. *Paired Intersections: Initial Development of Platooned Arrival and Queue Interaction Models.* Report WD TE91/010. Australian Road Research Board, 1991.
11. R. Akcelik. The Highway Capacity Manual Formula for Signalized Intersections. *ITE Journal*, Vol. 58, No. 3, 1988, pp. 23–27.
12. J. A. Hillier and R. Rothery. The Synchronization of Traffic Signals for Minimum Delay. *Transportation Science*, Vol. 2, May 1967, pp. 81–93.
13. D. Fambro and C. Messer. Estimating Delay at Coordinated Signalized Intersections. *Proc., International Symposium on Highway Capacity*, Karlsruhe, Germany, July 1991, pp. 127–144.
14. P. S. Olszewski. Traffic Signal Delay for Nonuniform Arrivals. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
15. J. Chodur and M. Tracz. Effects of Progression Quality and Traffic Flow Non-Stationarity in Delay Models at Signalized Intersections. *Proc., International Symposium on Highway Capacity*, Karlsruhe, Germany, July 1991, pp. 91–98.
16. S. C. Van As. Overflow Delay at Signalized Networks. *Transportation Research*, Vol. 25A, No. 1, 1991, pp. 1–7.
17. T. P. Hutchinson. Delay at a Fixed Time Traffic Signal-II, Numerical Comparisons of Some Theoretical Expressions. *Transportation Science*, Vol. 6, No. 3, 1972, pp. 286–305.
18. Newell. Stochastic Delays on Signalized Arterial Highways. *Proc., 11th International Symposium on Transportation and Traffic Flow Theory* (M. Koshi, ed.), Elsevier, New York, N.Y., 1990, pp. 589–598.
19. J. Staniewicz and H. Levinson. Signal Delay with Platoon Arrivals. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 28–32.
20. J. A. Sosin. Delays at Intersections Controlled by Fixed Traffic Signals. *Traffic Engineering and Control*, Vol. 21, Nos. 8/9, 1980, pp. 407–413.
21. R. Akcelik. *Calibrating SIDRA.* Research Report ARR 180. Australian Road Research Board, 1990.

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

Effects of U-Turns on Left-Turn Saturation Flow Rates

JOHN CLIFTON ADAMS AND JOSEPH E. HUMMER

As more U-turning vehicles use a left-turn lane, the saturation flow rate of the lane may become significantly lower. However, the 1985 *Highway Capacity Manual* (HCM) does not account for U-turns in calculating the capacity of a left-turn lane group at a signalized intersection. To determine whether a U-turn factor should be included in a revised HCM analysis method, a preliminary study was conducted at North Carolina State University. The study team selected four intersections with exclusive left-turn lanes and protected signal phasing and recorded saturation flow rates and U-turn percentages for 198 queues during weekday midday peaks. The data analysis showed that a saturation flow reduction factor appears necessary for left-turn lanes that have large percentages of U-turns. *T*-tests and regression models indicated that saturation flow rates were significantly lower when queues had more than 65 percent U-turns. However, the analyses also showed no correlation between saturation flow and the percentage of U-turns for queues with 50 percent or fewer U-turns. The analysis was inconclusive between 50 and 65 percent U-turns because of small samples. The results suggest tentative saturation flow reduction factors of 1.0 for U-turn percentages below 65, 0.90 for U-turn percentages between 65 and 85, and 0.80 for U-turn percentages exceeding 85. A follow-up investigation should focus on intersections that have high percentages of U-turns, restrictive geometry, or high percentages of U-turning heavy vehicles.

As traffic volumes continue to increase, states construct more roads that are divided by medians. One of the primary purposes of a median is to improve road safety by redirecting large volumes of left turns into driveways. Arterials lined with restaurants, stores, and other businesses tend to experience many accidents involving vehicles that attempt left turns across heavy traffic. By dividing a road with a median, some potential customers must proceed to the next crossover or intersection and make a U-turn. As a result, U-turn volumes are increasing at signalized intersections.

As more U-turning vehicles use a left-turn lane, the saturation flow rate for the lane may become significantly lower. However, the 1985 *Highway Capacity Manual* (HCM) (*1*) does not account for U-turns in calculating the capacity and level of service of a left-turn lane group at a signalized intersection. Because major revisions are planned for the operational model and method of analyzing signalized intersections in the HCM, the question arises: Should a U-turn factor be included when adjusting for left turns?

To help answer that question, a team from the Department of Civil Engineering at North Carolina State University (NCSU) conducted a preliminary study to determine the impact of U-turns on the saturation flow rate of left-turn lanes. A secondary objective of the study was to develop tentative U-turn adjustment factors derived from the percentage of U-turns at an intersection approach, should the need for an adjustment factor be evident. The team selected four intersections in Raleigh, North Carolina, for study. Each intersection had an exclusive left-turn lane with protected signal phasing and significant U-turn volumes. The study team recorded saturation flow rates and the percentage of U-turns for 198 queues. These data were then statistically analyzed to determine the effects of U-turning vehicles on the saturation flow rates of left-turn lanes. In this paper the current resources available to analyze the problem are briefly discussed, the study method is described, and the results from the data analysis are presented.

J. C. Adams, Post, Buckley, Schuh & Jernigan, Inc., 1560 Orange Avenue, Suite 700, Winter Park, Fla. 32789-5544. J. E. Hummer, Department of Civil Engineering, North Carolina State University, Box 7908, Raleigh, N.C. 27695-7908.

CURRENT RESOURCES

The 1985 HCM uses saturation flow rates for lane groups to determine the capacity and level of service at signalized intersections. The analysis of left-turn movements is one of the most important components of the signalized intersection analysis procedure. The 1985 HCM determines saturation flow by modifying a suggested ideal saturation flow rate. Currently, the manual recommends 1,800 passenger cars per hour of green time per lane (pcphgpl); however, it is likely that this value will be revised upward. Analysts modify this ideal by applying adjustment factors that describe nonideal traffic and roadway conditions. The left-turn adjustment factor accounts for the fact that left-turn movements are not made at the same saturation flow rate as through movements. There are eight left-turn adjustment factors, which are categorized by the number of turn lanes, type of phasing, and type of lane (exclusive or shared). For exclusive single and dual left-turn lanes with protected phasing, the adjustment factors are constant values of 0.95 and 0.92, respectively. The 1985 HCM does not include comment on how these adjustment factors were derived. However, these values were most likely developed from observed saturation headways of vehicles operating in exclusive left-turn lanes with protected phasing. The HCM does not give adjustment factors for left-turn lanes that accommodate a large number of U-turning vehicles. The 1985 HCM encourages users to measure saturation flow directly in the field when unique conditions are encountered. Unfortunately, many local agencies do not have the resources to conduct saturation flow studies, and high U-turning volumes are not unique in many areas.

An on-line search of the Transportation Research Information Service, along with on-line and manual searches of the NCSU library, revealed no published studies that directly address the effects of U-turns on protected left-turn saturation flow rates for single or dual lanes. *A Policy on Geometric Design of Highways and Streets* (*2*) provides guidance for the geometric design of intersections to accommodate U-turns but does not discuss U-turn capacity.

STUDY METHOD

To gain a preliminary understanding of the relationship between U-turns and left-turn saturation flow, the study team collected data at four intersections for which significant volumes of U-turns are made from exclusive left-turn lanes controlled by protected signals. Saturation flow rates and the percentages of U-turns were collected for individual queues of traffic. The study team also collected general information describing intersection layout at the study sites. The study needed data from the field because current traffic simulation packages are inadequate for analyzing the effect of U-turns on left-turn lanes. Exclusive left-turn lanes with protected signals make up the vast majority of locations in North Carolina where significant volumes of U-turns are found. If this preliminary study convinces the profession of the need for an adjustment factor developed from the percentage of U-turns, other signal and lane conditions may also be of interest in other states.

The measure of effectiveness used in this study was the saturation flow rate for the left-turn lane. The study team adapted a data collection sheet from the forthcoming *Manual of Transportation Research Studies* (*3*) for collecting saturation flow manually using a stopwatch. The study team adopted this method because it provided quick, accurate measurements of saturation flow and percentages of U-turns. The sheet was modified so that observers could record which vehicles in the queue made U-turns. Figure 1 displays the data collection sheet used in the study. For each observation, the observer circled the number on the form that corresponded to the queue position of each vehicle that made a U-turn. The observer started the stopwatch when the rear axle of the fourth vehicle in the queue crossed the stop bar and stopped the watch when the rear axle of the last stopped vehicle in the queue crossed the stop bar. Extensive studies of left-turn lanes have shown that headways stabilize after the fourth vehicle in the queue. Vehicles that joined the back of the queue after the signal had turned green were not included in the data collection. Observers did not record queues of six or fewer vehicles. For queues of more than 10 vehicles, the watch was stopped as the 10th vehicle crossed the stop bar. This reduced the chance that the green left-turn signal would expire before observers could record a time. Because times were always recorded when vehicles had green left-turn signals, ending lost time was not an issue.

The study team considered more than a dozen Raleigh-area sites for data collection. Prospective sites needed to have a median dividing the roadway, exclusive left-turn lanes, protected signal phasing, an approach grade of nearly 0 percent, an intersection angle of about 90 degrees, at least two lanes to receive the U-turns, limited left-turn and U-turn truck volumes, and adequate left-turn and U-turn volumes. Initially, the team wanted to collect data at both single and dual exclusive left-turn lanes with protected phasing. Unfortunately, no dual left-turn lanes examined satisfied the study requirements. The study team conducted short-turning movement counts at eight potential sites during peak volume periods. After these counts, the team selected the best four intersections for data collection. These intersections included Western Boulevard and Kent Road, Glenwood Avenue and Duraleigh Road, Capital Boulevard and Millbrook Road, and Capital Boulevard and Spring Forest Road.

Table 1 provides descriptions of the chosen sites. The study team measured the width of receiving areas for Intersections 2, 3, and 4 from the right edge of pavement to the yellow line that marked the inside shoulder. Medians were measured from yellow line to yellow line. Intersection 1 had a curbed concrete median, so the receiving area and median widths were measured from curb to curb. The sites were in suburban areas, and the major roads at each site were lined with businesses that generate U-turning traffic. Each intersection had a left-turn-lane storage length capable of holding 10 or more vehicles. Bus operations were not an issue at any of the intersections. Right turns on a red signal were legal from the minor streets at all intersections. Each intersection had three lanes available to receive U-turning traffic except for Intersection 1. However, Intersection 1 had a 3.66-m (12-ft) paved right shoulder that created a receiving width comparable to those of Intersections 2, 3, and 4. All the intersections were generally flat with the exception of Intersection 2, which had an approach grade of approximately −2 percent (i.e., downhill) at the stop bar.

The team collected data during eight weekdays in July 1992 at the midday peak between 11:00 a.m. and 2:30 p.m. The midday peak periods had large volumes of left-turning and U-turning traffic at the sites. The weather on data collection days was clear and pavements were dry. Recorded queues consisted only of passenger vehicles and light trucks. Observers ignored queues that had buses, trucks with more than four tires, motorcycles, or vehicles with trailers. Also, queues in which unusual events occurred, such as a right turn on red from the minor street that blocked U-turns or an uncharacteristically slow vehicle, were not considered. Adams collected most of the data, although Hummer collected some data at the intersection of Western Boulevard and Kent Road. The observer stood close enough to the intersection to clearly see the lane being observed, the stop bar, and the signal indication, but was generally not visible to turning drivers. Upon completion of the data collection, the field information was transferred to a computer data file that was analyzed using Version 6.07 of the Statistical Analysis System software (*4*).

RESULTS

During data collection, the observers noted some general perceptions concerning left-turn and U-turn traffic. First, the observers noted that few conflicts occurred between right-on-red movements from the minor street and U-turn vehicles, even though there were usually vehicles ready to turn right on red. On a few occasions, large trucks turned right on a red signal and obstructed U-turn traffic. These observations

Intersection: _____

Approach: _____

Lane Type: _____

Time: _____ Date: _____

Obs. No.	Time (seconds) between 4th vehicle and ...				U-turn positions
	7th veh.	8th veh.	9th veh.	10th veh.	
1					1 2 3 4 5 6 7 8 9 10
2					1 2 3 4 5 6 7 8 9 10
3					1 2 3 4 5 6 7 8 9 10
4					1 2 3 4 5 6 7 8 9 10
5					1 2 3 4 5 6 7 8 9 10
6					1 2 3 4 5 6 7 8 9 10
7					1 2 3 4 5 6 7 8 9 10
8					1 2 3 4 5 6 7 8 9 10
9					1 2 3 4 5 6 7 8 9 10
10					1 2 3 4 5 6 7 8 9 10
11					1 2 3 4 5 6 7 8 9 10
12					1 2 3 4 5 6 7 8 9 10
13					1 2 3 4 5 6 7 8 9 10
14					1 2 3 4 5 6 7 8 9 10
15					1 2 3 4 5 6 7 8 9 10
16					1 2 3 4 5 6 7 8 9 10
17					1 2 3 4 5 6 7 8 9 10
18					1 2 3 4 5 6 7 8 9 10
19					1 2 3 4 5 6 7 8 9 10
20					1 2 3 4 5 6 7 8 9 10

FIGURE 1 Data collection sheet used to record field observations.

TABLE 1 Intersection Descriptions

Intersection	Major St.	Minor St.	Intersection Angle (degrees)	Width of Receiving Area (m)	Curb and Gutter (yes/no)	Approach Lane Width (m)	Median Type	Width of Median (m)	% Grade of Approach	Dist. Stop Bar from Nose (m)
#1	Western Blvd.	Kent Rd.	90	10.97	yes	3.66	concrete	2.44	0	0
#2	US70	Duraleigh Rd.	90	11.28	yes	3.66	grass	6.10	-2	0
#3	Capital Blvd.	Millbrook Rd.	90	14.02	yes	3.66	grass	5.79	0	0
#4	Capital Blvd.	Spring Forest Rd.	90	14.02	yes	3.66	grass	5.49	0	0

1 m = 3.28 ft

were interesting but not kept in the data base. Second, pedestrians did not conflict with left-turning or U-turning traffic during data collection. Intersection 1 had the most pedestrian traffic and was the only intersection with pedestrian-crossing indicators. Intersection 1, however, produced only one pedestrian conflict that disrupted left-turn traffic flow during about 5 hr of data collection. This observation was not included in the data base. Finally, observers noted that U-turning trucks significantly influenced the saturation flow of left-turn lanes. The observers only recorded queues consisting of passenger vehicles for this study; however, left-turn lanes with significant U-turn truck traffic may warrant further investigation.

The first step of the analysis was to calculate the saturation flow rate and the percentage of U-turns for each observed queue. Saturation flow was easily calculated by converting the recorded times into vehicles per hour. The study team had several options available to determine which vehicles to include when computing the percentage of U-turns, however. The team could have computed the percentage of U-turns using the vehicles in queue Positions 1 through the last recorded vehicle in the queue (n), Positions 1 through $n-1$, Positions 4 through n, or Positions 4 through $n-1$. To determine the preferred option, the team first addressed the issue of whether U-turns made in the first three positions of the queue significantly affected saturation flow measured for Vehicles 4 through n. The study team calculated the mean saturation flows associated with queues that had zero, one, two, or three U-turn movements in the first three queue positions. Table 2 gives these means by intersection. For cases in which sample sizes allowed meaningful comparisons within an intersection, the team conducted t-tests between groups of observations with different numbers of U-turns. The results of these t-tests indicated that no two means were significantly different from each other at the 95 percent confidence level. Therefore, the study team chose to define the beginning of the queue with the fourth vehicle.

The next question in deciding how to calculate the percentage of U-turns was whether to include the last vehicle recorded. A U-turning vehicle could affect saturation flow by slowing before it reaches the stop bar or by causing the vehicle behind it to slow. If only the latter effect is important, the last vehicle in the queue should not be included when computing and analyzing the percentage of U-turns. The study team believed, on the basis of field observations, that both effects are important and that the last vehicle should be included. To verify this belief, the study team calculated mean saturation flows for each intersection for queues where the last vehicle did or did not make a U-turn, as shown in Table 3. The mean saturation flow is lower at all four intersections when the last vehicle made a U-turn, and t-tests at the 95 percent confidence level showed a significant difference between the means for Intersection 4. The study team therefore computed the percentage of U-turns from Vehicle 4 through the last observed vehicle.

The study team observed a total of 198 queues at the four intersections. Figures 2 through 5 show scatter diagrams of saturation flow versus the percentage of U-turns for each intersection. Table 4 provides a summary of key left-turn lane statistics by intersection. A mean saturation flow of 1,589 pcphgpl at Intersection 1, which is lower than other sites, may be attributed to a curbed median, a narrower side street, more local traffic, and other reasons. Mean saturation flow rates at Intersections 2, 3, and 4 do not differ greatly from each

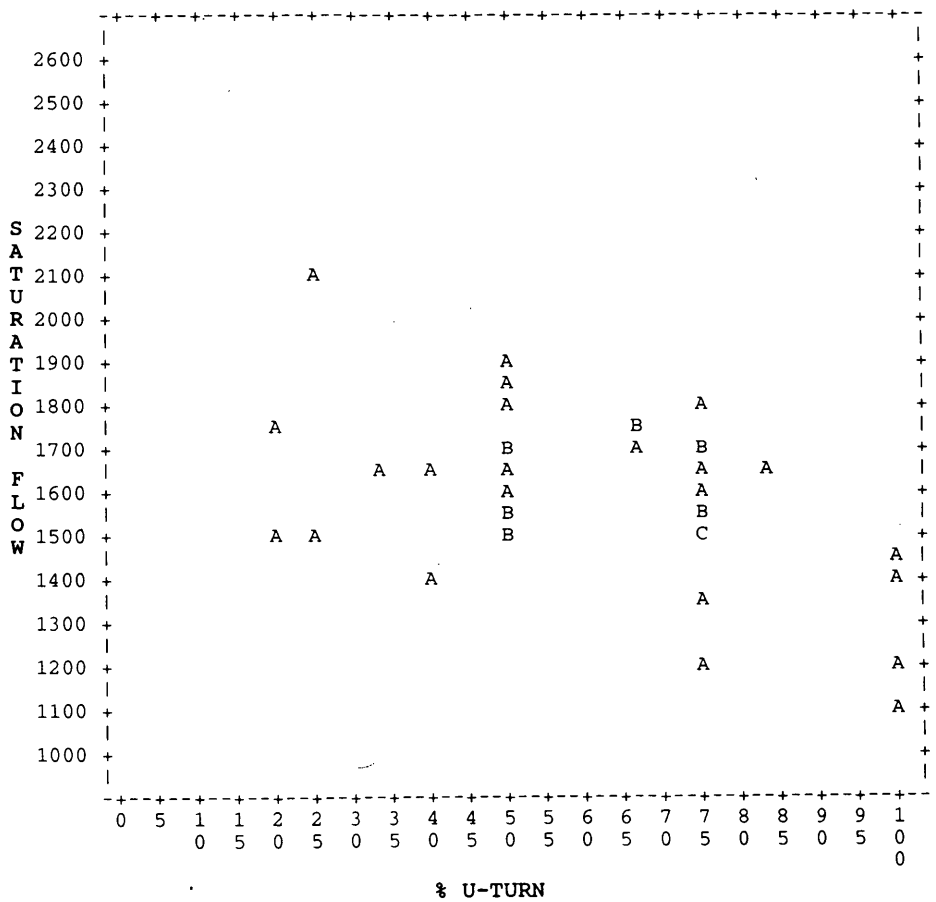
TABLE 2 Summary Statistics for U-Turns Made in First Three Queue Positions

INTERSECTION	# U-TURNS MADE IN FIRST 3 QUEUE POSITIONS	# OBS.	MEAN SATURATION FLOW (pcphgpl)	STANDARD DEVIATION
1	0	2	1764	204
	1	7	1531	156
	2	16	1569	189
	3	13	1617	219
2	0	12	1710	180
	1	18	1861	229
	2	15	1921	191
	3	4	1843	166
3	0	28	1877	176
	1	13	1776	202
	2	5	1819	340
	3	-	-	-
4	0	39	1838	231
	1	15	1835	266
	2	11	1944	261
	3	-	-	-

NOTE: The symbol "-" indicates that data were not available.

TABLE 3 Summary Statistics for Queues in Which Last Vehicle Did and Did Not Make a U-Turn

INTERSECTION	LAST VEHICLE MAKING U-TURN? (yes/no)	# OBS.	MEAN SATURATION FLOW (pcphgpl)	STANDARD DEVIATION
1	no	15	1631	176
	yes	23	1561	204
2	no	30	1865	206
	yes	19	1803	221
3	no	39	1854	214
	yes	7	1773	133
4	no	47	1915	238
	yes	18	1699	185



NOTE: A = 1 obs., B = 2 obs., etc.
Saturation flow measured in pcphgpl.

FIGURE 2 Scatter diagram for Intersection 1.

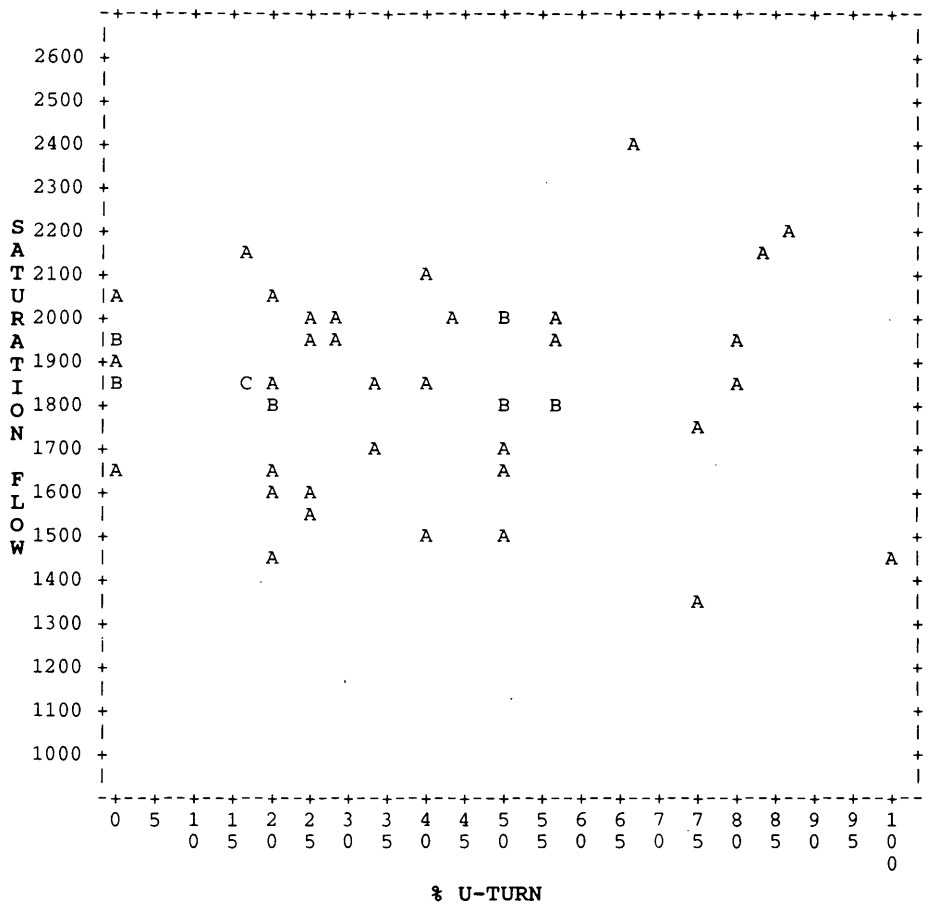


FIGURE 3 Scatter diagram for Intersection 2.

other. Intersections 1 and 4 had wide ranges for the percentage of U-turns, but at Intersection 3, 43 percent was the maximum percentage of U-turns for any queue.

Simple linear regression was used to model the relationship between saturation flow and U-turn percentage. The study team chose simple linear regression over other model forms because of a desire for a simple model consistent with other models in the HCM and a visual inspection of the scatter diagrams. Although there is theory relating saturation flow rates to other nonideal conditions, the study team found no theory on the form of the relationship of saturation flow to the percentage of U-turns. Table 5 gives results of the simple linear regression analysis for each intersection. The simple linear regression models generally showed a poor correlation between saturation flow and the percentage of U-turns. Intersections 1 and 4 had the highest correlation with adjusted R^2 values of 0.20 and 0.14, respectively. In addition, for Intersections 1 and 4 the slopes of the regression lines were negative and the coefficients were significantly different from zero at the 95 percent level according to *t*-tests. These findings suggest that saturation flow tends to decrease as the percentage of U-turns increases. Intersection 3, with a maximum percentage of U-turns less than 50, showed no correlation between saturation flow and the percentage of U-turns (adjusted $R^2 = 0.0141$).

On the basis of visual inspection of Figures 2 through 5 and the fact that no correlation existed for Intersection 3 (for which the maximum percentage of U-turns was less than 50), the study team applied simple linear regression with 50 percent U-turns as a break point. Table 6 gives the results for each intersection for which the percentage of U-turns was less than or equal to 50. Adjusted R^2 values for each intersection indicate that no correlation existed when the percentage of U-turns was less than or equal to 50. Also, the coefficients of the U-turn percentage variable (i.e., the slopes of the regression lines) were not significantly different from zero at the 95 percent level according to *t*-tests. This indicates that there is little change in saturation flow rate as the percentage of U-turns increases from 0 to 50.

Table 7 gives the results of the regression analysis for which the percentages of U-turns were greater than 50 percent. Intersection 3 was not analyzed because only queues with fewer than 50 percent U-turns were observed. Regression analysis on Intersection 2 data indicated no correlation between saturation flow and the percentage of U-turns (adjusted $R^2 = -0.0457$). Analysis of Intersections 1 and 4 showed a stronger

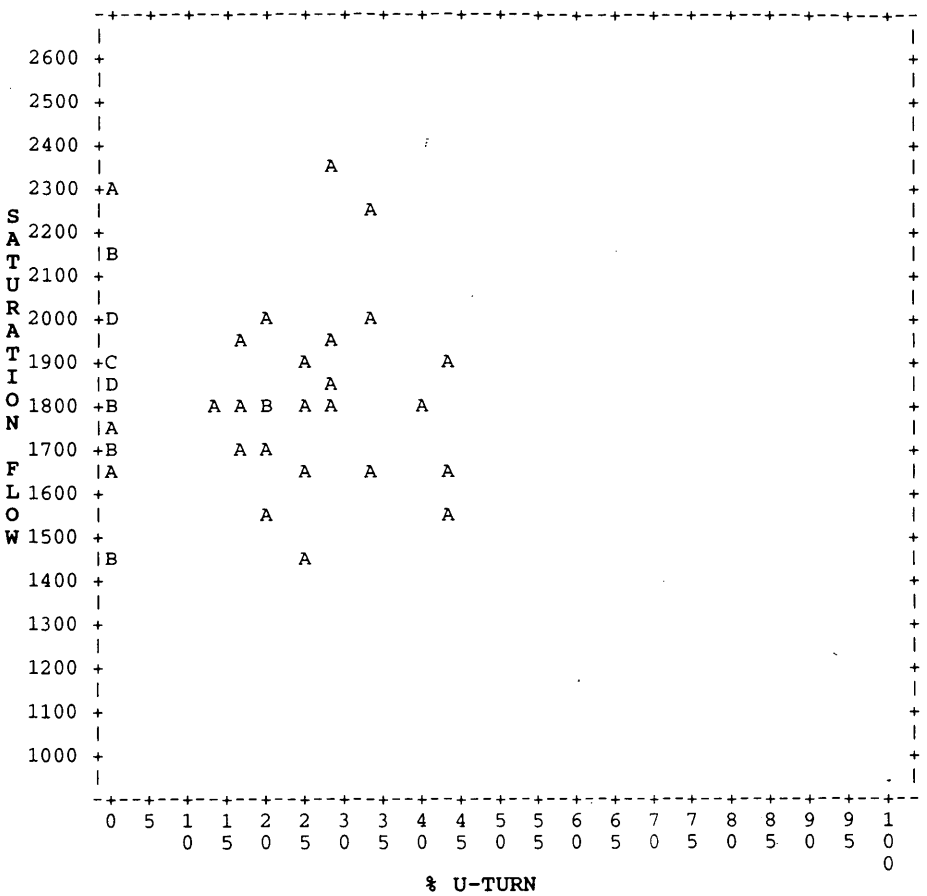


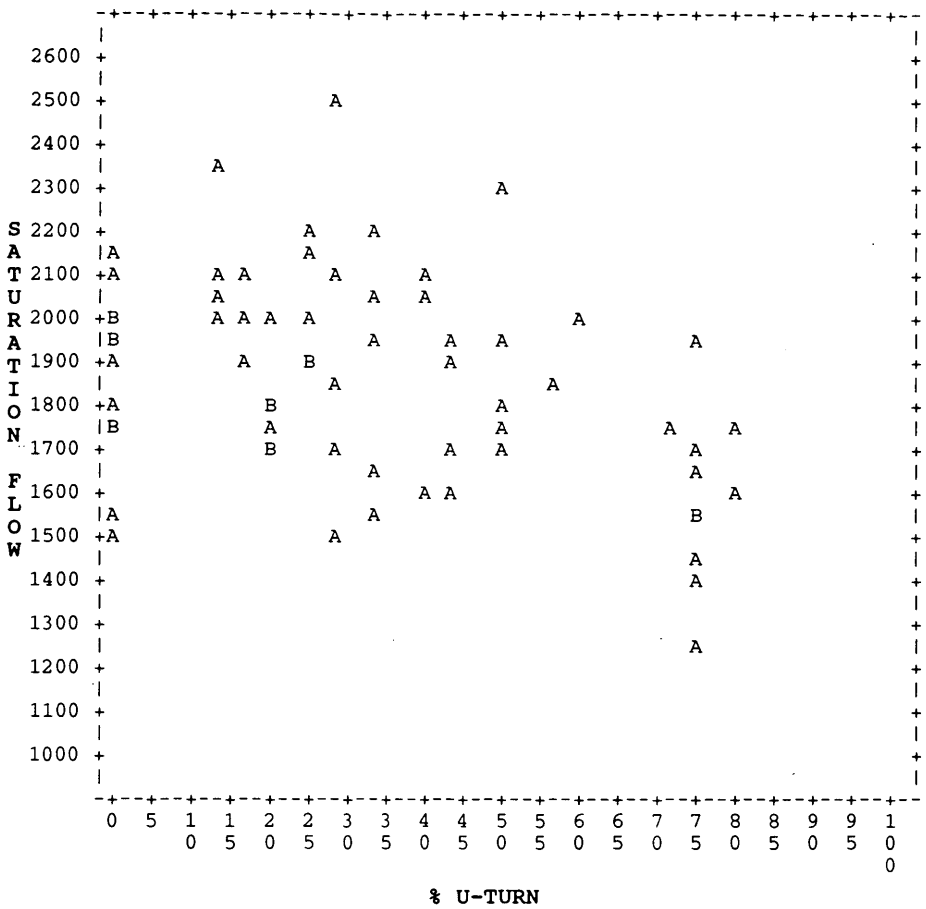
FIGURE 4 Scatter diagram for Intersection 3.

correlation between saturation flow and the percentage of U-turns than in other analyses with adjusted R^2 values of 0.39 and 0.19, respectively. Slopes of -15.4 for intersection 4 and -11.2 for Intersection 1 indicate that saturation flow decreases as U-turns increase beyond 50 percent.

To better understand the data, the study team divided the data set into five separate groups by the percentage of U-turns. The data were grouped on the basis of the results of the linear regression analyses and a visual inspection of the scatter diagrams in Figures 2 through 5, as follows:

1. 0 to 29 percent,
2. 30 to 50 percent,
3. 51 to 65 percent,
4. 66 to 85 percent, and
5. 86 to 100 percent.

Table 8, which is a summary of *t*-tests, gives the mean saturation flow rates of each group within each intersection at the 95 percent confidence level. Intersection 1 showed no significant difference among Groups 1, 2, and 4, although Group 4 had a lower mean than Groups 1 and 2. Group 5, however, was significantly different from the other groups. Group 3 was not analyzed for Intersection 1 because the study team did not record any queues for which 51 to 65 percent U-turns were observed. For Intersection 2 there were no significant differences between any of the groups, although the analysis was hampered because Groups 3, 4, and 5 had small sample sizes. Other reasons why saturation flow did not decrease with the percentage of U-turns could be attributed to the downgrade approach and the width of the cross street that allowed left-turning vehicles to quickly bypass U-turning vehicles. Because the maximum percentage of U-turns was less than 50 for Intersection 3, only Groups 1 and 2 were compared using the *t*-test. The test revealed no significant difference between the two groups. The analysis of Intersection 4 showed no significant difference between Groups 1, 2, or 3. Group 4, however, was significantly different from Groups 1, 2, and 3. Generally for all the intersections, *t*-tests showed no significant differences between Groups 1 and 2. This result confirms the results produced by the regression analysis for 50 percent or lower U-turns. Because of limited sample sizes, *t*-tests for Group 3 were inconclusive. Groups 4 and 5 were shown to be significantly different from other groups at Intersections 4 and 1, respectively, and indicates that saturation flow decreases as U-turns increase above 65 percent.



NOTE: A = 1 obs., B = 2 obs., etc.
Saturation flow measured in pcphgpl.

FIGURE 5 Scatter diagram for Intersection 4.

TABLE 4 Summary of Key Left-Turn Lane Statistics by Intersection

INTERSECTION	# OBS.	MEAN SATURATION FLOW (pcphgpl)	SAT. FLOW STDEV	MIN. SATURATION FLOW (pcphgpl)	MAX. SATURATION FLOW (pcphgpl)	MEAN %U-TURN	%U-TURN STDEV	MIN. %U-TURN	MAX. %U-TURN
1	38	1589	194	1121	2122	62	22	20	100
2	49	1841	212	1367	2381	37	26	0	100
3	46	1842	204	1444	2328	14	15	0	43
4	65	1855	244	1262	2515	33	25	0	80

TABLE 5 Linear Regression Results for Percentage of U-Turns between 0 and 100

INTERSECTION	DEGREES OF FREEDOM	ESTIMATED CONSTANT	CONSTANT STANDARD ERROR	ESTIMATED X COEFF.	X COEFF. STANDARD ERROR	ADJUSTED R-SQUARE	Significant at 95% level using t-test? CONSTANT	X COEFF.
1	37	1844.11	84.57	-4.15	1.30	0.20	yes	yes
2	48	1838.21	53.46	0.07	1.18	-0.02	yes	no
3	45	1859.52	41.80	-1.25	2.04	-0.01	yes	no
4	64	1982.29	46.78	-3.83	1.13	0.14	yes	yes

TABLE 6 Linear Regression Results for Percentage of U-Turns Less Than or Equal to 50

INTERSECTION	DEGREES OF FREEDOM	ESTIMATED CONSTANT	CONSTANT STANDARD ERROR	ESTIMATED X COEFF.	X COEFF. STANDARD ERROR	ADJUSTED R-SQUARE	Significant at 95% level using t-test? CONSTANT	X COEFF.
1	17	1700.99	163.60	-1.07	3.77	-0.06	yes	no
2	36	1862.79	55.2	-1.47	1.81	-0.01	yes	no
3	45	1859.52	41.80	-1.25	2.04	-0.01	yes	no
4	51	1919.79	54.62	-0.48	1.93	-0.02	yes	no

TABLE 7 Linear Regression Results for Percentage of U-Turns Greater than 50

INTERSECTION	DEGREES OF FREEDOM	ESTIMATED CONSTANT	CONSTANT STANDARD ERROR	ESTIMATED X COEFF.	X COEFF. STANDARD ERROR	ADJUSTED R-SQUARE	Significant at 95% level using t-test? CONSTANT	X COEFF.
1	19	2413.44	245.21	-11.17	3.07	0.39	yes	yes
2	11	2220.63	467.88	-4.55	6.32	-0.05	yes	no
3	-	-	-	-	-	-	-	-
4	12	2764.11	577.53	-15.39	7.89	0.19	yes	no *

NOTE: The symbol "-" indicates that data were not available.
* Significant at 90% level.

TABLE 8 Summary of *T*-Tests Comparing Mean Saturation Flow Rates of Each U-Turn Percentage Group

INTERSECTION	GROUP	# OBS.	MEAN SAT. FLOW (pcphgpl)	SAT. FLOW STDEV	GROUPS WITH SIGNIFICANTLY DIFFERENT MEANS (95% LEVEL)
1	1	4	1714	294	5
	2	14	1639	141	5
	3	-	-	-	-
	4	16	1586	155	5
	5	4	1296	155	1, 2, 4
2	1	24	1838	179	none
	2	13	1802	195	none
	3	4	1898	106	none
	4	6	1905	349	none
	5	2	1821	503	none
3	1	39	1844	201	none
	2	7	1830	240	none
	3	-	-	-	-
	4	-	-	-	-
	5	-	-	-	-
4	1	35	1928	227	4
	2	17	1868	218	4
	3	2	1904	105	4
	4	11	1594	185	1, 2, 3
	5	-	-	-	-

NOTE: The symbol "-" indicates that data were not available.

SUMMARY AND CONCLUSION

In this preliminary study, the primary objective was to investigate the impact of U-turns on left-turn-lane saturation flow. The secondary objective was to develop saturation flow adjustment factors on the basis of the percentage of U-turns, should the need for such factors be proven. Saturation flow rates and the percentages of U-turns were recorded for exclusive, single left-turn lanes with protected phasing at four intersections in Raleigh, North Carolina. Data were collected on weekdays between 11:00 a.m. and 2:30 p.m. Queue lengths for saturation flow measurements varied from 7 to 10 vehicles and the queues that were measured contained no buses, trucks, motorcycles, or trailers. The study team calculated the percentage of U-turns on the basis of the fourth vehicle to the last observed vehicle in the queue.

From the analysis, a saturation flow reduction factor appears necessary for left-turn lanes with a large percentage of U-turns. *T*-tests at the 95 percent confidence level between mean saturation flow rates for observations grouped by the percentage of U-turns show significant differences when U-turns are greater than 65 percent. In addition, simple linear regression models had large negative slopes, which supported a decrease in saturation flow with more than 50 percent U-turns. *T*-tests and linear regression analyses provided strong evidence that no correlation between saturation flow and the percentage of U-turns exists for 50 percent or fewer U-turns. The analysis was inconclusive between 50 and 65 percent U-turns because of the small samples in this study. The break point at which the percentage of U-turns starts to be a significant factor may occur in this region.

The study results suggest tentative saturation flow reductions of about 10 percent for U-turn percentages between 65 and 85 and 20 percent for U-turn percentages exceeding 85. Reductions of 10 and 20 percent equate to U-turn adjustment factors of 0.9 and 0.8, respectively. Saturation flow reductions of 10 and 20 percent are in line with the analysis of group means described in this paper and are somewhat conservative compared with the regression line slopes for observations with more than 50 percent U-turns. These reductions are suggestions developed from small sample sizes, and analysts should await further research before applying them.

Although the adjustment factors suggested in this study were derived from observations of single left-turn lanes, the results may also apply to dual left-turn lanes. Currently the 1985 HCM computes the saturation flow for exclusive dual left-turn lanes by multiplying an ideal saturation flow of 1,800 pcphgpl by the number of turn lanes and a dual left-turn lane adjustment factor of 0.92. On the basis of the results presented in this paper, suggested correction factors for dual left-turn lanes are 5 percent for 65/2 percent U-turns and 10 percent for 85/2 percent U-turns.

This preliminary study demonstrated that further investigation into the impact of U-turns on left-turn saturation flow

is warranted. One major and surprising finding from this study was that there was little variation in saturation flow observed when the U-turn percentage was 50 or less. A future follow-up study should focus on intersections that have high percentages of U-turns. Other situations worth attention include intersections with a large percentage of U-turning trucks, U-turns from dual left-turn lanes (to verify that the suggestions given are valid), and U-turns into narrow receiving areas or across narrow medians.

ACKNOWLEDGMENT

The Department of Civil Engineering at NCSU provided support for the study described in this paper.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. *A Policy on Geometric Design of Highways and Streets*. AASHTO, Washington, D.C., 1990.
3. J. E. Hummer. Intersection and Driveway Studies. In *Manual of Transportation Research Studies*. ITE, Washington, D.C., Prentice Hall, Englewood Cliffs, N.J., (in press).
4. *SAS Language and Procedures: Usage, Version 6*, 1st ed., SAS Institute Inc., Cary, N.C., 1989.

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

Probability of Overload at Signalized Intersections

STAN TEPLY

Quality of service at transportation facilities can be determined from a number of objectively measurable criteria. For signalized intersections, delay has been accepted internationally as the prime measure of performance and, in the United States, as the sole basis for the level-of-service measures. In other countries, several other parameters or their combinations are also used, usually as additional measures related to specific design or evaluation objectives. A pilot project at the University of Alberta, which investigated the potential of a supplementary criterion—the probability of discharge overload—is the basis of the discussion. This measure would indicate in how many cycles the number of vehicles left over from the previous cycle plus the number of vehicles arriving exceeds capacity. Such information would assist in the analysis of a specific lane problem. Vehicle demand in successive cycles is treated as a series of dependent events on the basis of the Poisson distribution of arrivals. Surveys and simulations of an overload factor, which somewhat extends the concept of the load factor used in the 1965 *Highway Capacity Manual*, are used as indications of which probability type would provide the most practical representation of measurable parameters. The results of the pilot project suggest that the probability of an overload in one, two, or both of two consecutive cycles is a strong candidate to approximate the percentage of cycles with discharge overloads. It is mathematically simple and, because it resembles the overload factor, easy to measure.

The quality of service offered by a transportation facility manifests itself to the users and to the designer or analyst in a number of ways. The 1985 *Highway Capacity Manual* (HCM) (*1*) employs such criteria as speed, traffic density, percentage time delay, reserve capacity, and so on, depending on the type of facility. The measure used for signalized intersections is the average stopped vehicular delay, which is applied as the sole characteristic of levels of service. In many specific cases, however, the traffic engineer must also analyze other operational features that relate more closely to the problem at hand, such as the capacity, length of the queue in important lanes, probability that an approach or a lane will be oversaturated, delay to passengers, number of stops, emissions, or cost.

The 1965 HCM (*2*) defined *load factor* as the measure underlying the level of service. The 1984 *Canadian Capacity Guide for Signalized Intersections* (CCG) (*3*) does not specify levels of service but relates the ranges of average overall delay to the quality of operation. It also recommends additional criteria, such as the probability of clearance and queueing. The second edition will include a number of other measures of effectiveness (*4*). A 1983 Australian Road Research Board report (*5*) employs overflow queue, delay, number of stops, and queue length.

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

ENGINEERING EVALUATION AND QUALITY OF SERVICE

An examination of the definitions of various criteria used for signal design and analysis indicates that there are two different, but related, reasons for specific measures describing intersection performance:

- Relating the quality of operation to the general public and politicians, and
- Identifying specific features of the operation with the intent of revealing problems, their causes, and contributory factors.

The first reason is clearly related to perceptions. The level-of-service concept represents an attempt to describe the ranges of satisfaction provided by the facility. Traffic engineers apply categories of an objective measure to define different levels of service. The objective measure (e.g., delay in the 1985 HCM) has been selected on the basis of its perceivability by the user of the facility or a decision maker.

The second reason may employ a whole range of operational characteristics. If, for instance, a frequent lane blockage by a spillover queue has been the problem, the analyst should be in the position to determine both the length of the queue and the frequency of this undesirable event. Therefore, it should be possible to both measure and calculate these operational features. Naturally, the computational process may involve analytical formulas, iterative procedures, or simulation.

The focus of this paper is on cycle overload as a specific engineering operational criterion. It discusses its forms and past formulations and presents potential analytical and measurement methods to make its use practical. The intention is not to substitute delay as a measure underlying levels of service, but merely to provide the analyst with an additional tool.

LOAD FACTOR

The load factor was used as an intersection performance criterion to determine the level of service in the 1965 HCM. It was classified as a part of environmental conditions, that is, conditions that cannot be readily changed by alteration of design or control features of the intersection.

This factor was defined very specifically as follows:

> The load factor is a measure of the degree of utilization of an intersection approach roadway during one hour of peak traffic flow. It is the ratio of the number of green phases that are loaded, or fully utilized, by traffic (usually during the peak hour) to the total number of green phases available for that approach during the same period. As such it is also a measure of the level of service on the approach, as discussed later. The load factor for a normal intersection may range from a value of 0.0 to a value of 1.0. (*2*)

The 1965 HCM defined a loaded cycle as a cycle in which there are vehicles ready to enter the intersection when the signal turns green and in which vehicles continue to be available to enter during the entire phase with no unused time or exceedingly long spacings between vehicles at any time because of lack of traffic. It was pointed out that the ending of a loaded green phase will usually (i.e., not necessarily) force some vehicles to stop, and that any stoppages that occur must be caused by conditions at the intersection under study, not by conditions elsewhere. This description therefore also implies that the just-at-capacity cycles were included (i.e., more than just overloaded cycles were considered).

A number of research papers analyzed load factor applications and their relation to delay using simulation (*6*), measurements, or analytical methods (*7–9*).

Experience with the 1965 HCM indicated that the major drawbacks of the load factor application were as follows:

- No computational procedures for design situations were available in the 1965 HCM. Miller (*10*), in 1968, proposed a formula derived from May and Pratt's simulation (*6*) as follows:

$$LF = e^{-1.3\phi} \quad (1)$$

where

$$\phi = (1 - x)\sqrt{sg}/x \quad (2)$$

and

LF = load factor,
x = degree of saturation (i.e. qc/sg),
q = arrival flow,
s = saturation flow,
g = effective green interval, and
c = cycle time

- The load factor determination did not include its variability even under identical traffic conditions. For instance, measurement of the load factor during the same periods of similar days with the same volumes and signal timing will bring about a whole range of load factors. A single measurement is only a part of an unknown distribution and, in itself, may not tell much about the average or the deviations.
- No recognition was given of temporary, very short traffic flow peaks or of the fact that the number of overloaded cycles is time-dependent. The chances that a cycle following an overloaded cycle will also be overloaded are higher than for a cycle following a cycle that left no overflow queue. This is recognized in the inclusion of the evaluation time element in some random overflow delay formulas (*3,5*).
- The loaded cycles are difficult to identify because no visible demonstration of a zero overflow queue occurs.

PROBABILITY OF ARRIVAL OVERLOAD

There were other attempts to overcome the problems of load factor calculations. The most obvious candidate for a similar surrogate measure was the relationship between the arrival distribution and cycle capacity.

This measure is defined as the probability that the number of vehicles arriving during a cycle exceeds the average cycle capacity. It is based on the counting distribution of vehicle arrivals during individual cycles. Under the assumptions of a fixed-time signal operation, a steady mean of the arrival distribution, and the usual conditions at signalized intersections, the arrival fluctuations can be approximated by the Poisson distribution (*3,11–13*).

The Poisson distribution also applies to approach lanes at intersections within coordinated signal systems. Although the internal arrival patterns during individual portions of the cycle depend on the quality of the progression (and may follow different counting distributions), the cycle-based arrival distribution can be expressed by the Poisson distribution in most cases. This distribution is well known to practitioners because it has been widely used as a basis for allocating green intervals or evaluation (*3,11,12*).

The probability of an overload in a cycle (i.e., not a loaded cycle in terms of the 1965 HCM), provided that the preceding cycle was not overloaded, equals the probability that the number of vehicles arriving during that cycle exceeds its capacity:

$$P(x > X) = 1 - \sum e^{-m} m^i / i! \quad \text{for } i = 0 \text{ to } i = X \quad (3)$$

where

x = number of arrivals in a cycle,
X = cycle capacity,
e = basis of natural logarithms,
m = average number of arrivals per cycle, and
i = counter.

The number of arrivals is normally expressed in vehicles (veh) per cycle, although passenger car units per cycle can also usually be used without any loss of representative power.

The formula corresponds to percentage of cycle failures by Drew and Pinnell (*13*), which was also Poisson-based, relating the average number of arrivlas per cycle to the probability of $X + 1$ or more arrivals per cycle.

As shown in Figure 1, this probability is not symmetrical (because of the probability of 0 arrivals). If the average number of arrivals equals cycle capacity, the probability of more vehicles arriving than capacity is still less than 0.5. For example, for $X = 6.0$ veh/cycle, and $m = 6.0$ veh/cycle,

$$P_{\text{arrival overload}} = P(x > X) = P(x > 6) = 0.394$$

In instances for which capacity well exceeds the average number of arrivals, this simple distribution can be used as an approximation of the number of cycles that will be overloaded. Nevertheless, as the average number of arrivals increases (and capacity remains constant) the number of over-

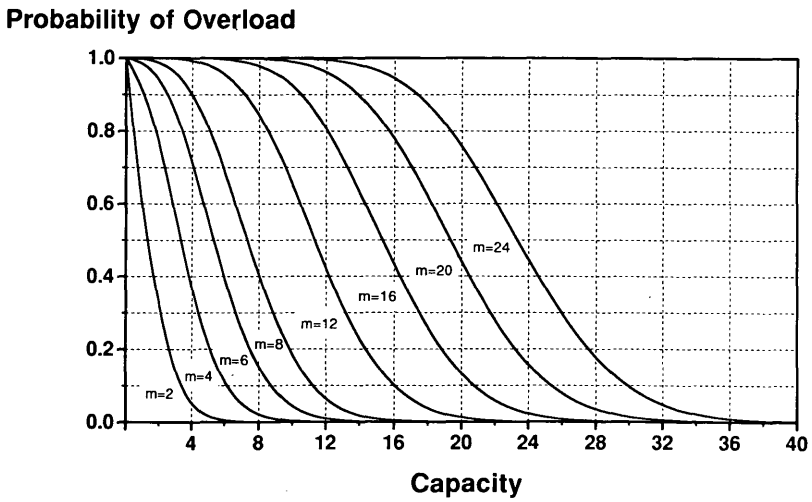


FIGURE 1 Poisson distribution for probability of number of arrivals exceeding capacity of a cycle (probability of arrival overload) for cycle time as time basis for mean of arrival distribution (m).

loaded cycles may exceed this probability measure by a wide margin. For example, in the preceding case of an average number of arrivals equal to six and capacity of six, the queueing system is unstable. Depending on the starting conditions (the presence of an initial queue), the probability that a cycle will be overloaded may reach 100 percent despite the fact that the arrivals at the end of the queue still follow the Poisson distribution. Naturally, if the average number of arrivals is greater than the cycle capacity, the queue will be consistently growing. The computation becomes meaningless because all cycles will be overloaded.

For this reason, it is advisable not to confuse the arrival distribution with the probability that a cycle will be overloaded. It is, however, still a useful simple measure. It has been suggested during the review process of the CCG to call it probability of arrival overload, representing the frequency with which more vehicles than the cycle capacity may arrive. Keeping in mind that since in a random process the number of vehicles arriving during one cycle does not depend on the number of arrivals in the previous cycle, the probability of arrival overload in any cycle is independent of the probability of arrival overload in previous cycles. As a result, probability of arrival overload does not indicate the likelihood that the sum of vehicles waiting and arriving exceeds cycle capacity.

PROBABILITIES OF DISCHARGE OVERLOAD

Miller also suggested a formula for the proportion of phases in which the queue is cleared (*10*). For consistency, a complementary probability term will be used in this paper as follows:

$$P_{\text{cycle overload}} = (1 - p_o) = e^{-1.58\Phi} \tag{4}$$

where

$$\Phi = (1 - x)\sqrt{sg}/x$$

and

x = degree of saturation (i.e., qc/sg),
q = arrival flow,
s = saturation flow,
g = effective green interval, and
c = cycle time.

Messer and Fambro (*14*) have shown a comparison of Poisson probability of arrival overload, Miller's load factor, and Miller's probability of cycle overload.

Obviously, because of the number of cycles considered in the numerator

- Load factor: $(x_{\text{arriving}} + x_{\text{waiting}}) \geq X$ (where X denotes cycle capacity),
- Probability of cycle overload: $(x_{\text{arriving}} + x_{\text{waiting}}) > X$, and
- Probability of arrival overload: $(x_{\text{arriving}}) > X$, the following relationship must exist:

$$\text{LF} > P_{\text{cycle overload}} > P_{\text{arrival overload}}$$

Equations proposed originally for the second edition of the CCG also featured an empirical negative exponential form (*4*) but included a calibration parameter for the series of the means of arrival distributions as follows:

$$P_{\text{cycle overload}} = 100e^{-[X-(m-1)]/a}\ (\%) \tag{5}$$

where

X = cycle capacity,
m = means of arrival distribution, and
a = calibration parameter.

The problem of sequential overloads merits closer inspection. Figure 2 illustrates possible cycle overload situations in a sequence of probabilities for an average of two arrivals per cycle and a cycle capacity also equal to two. The fast growing extent of this somewhat extreme case of a probability tree is the reason that such small numbers were chosen.

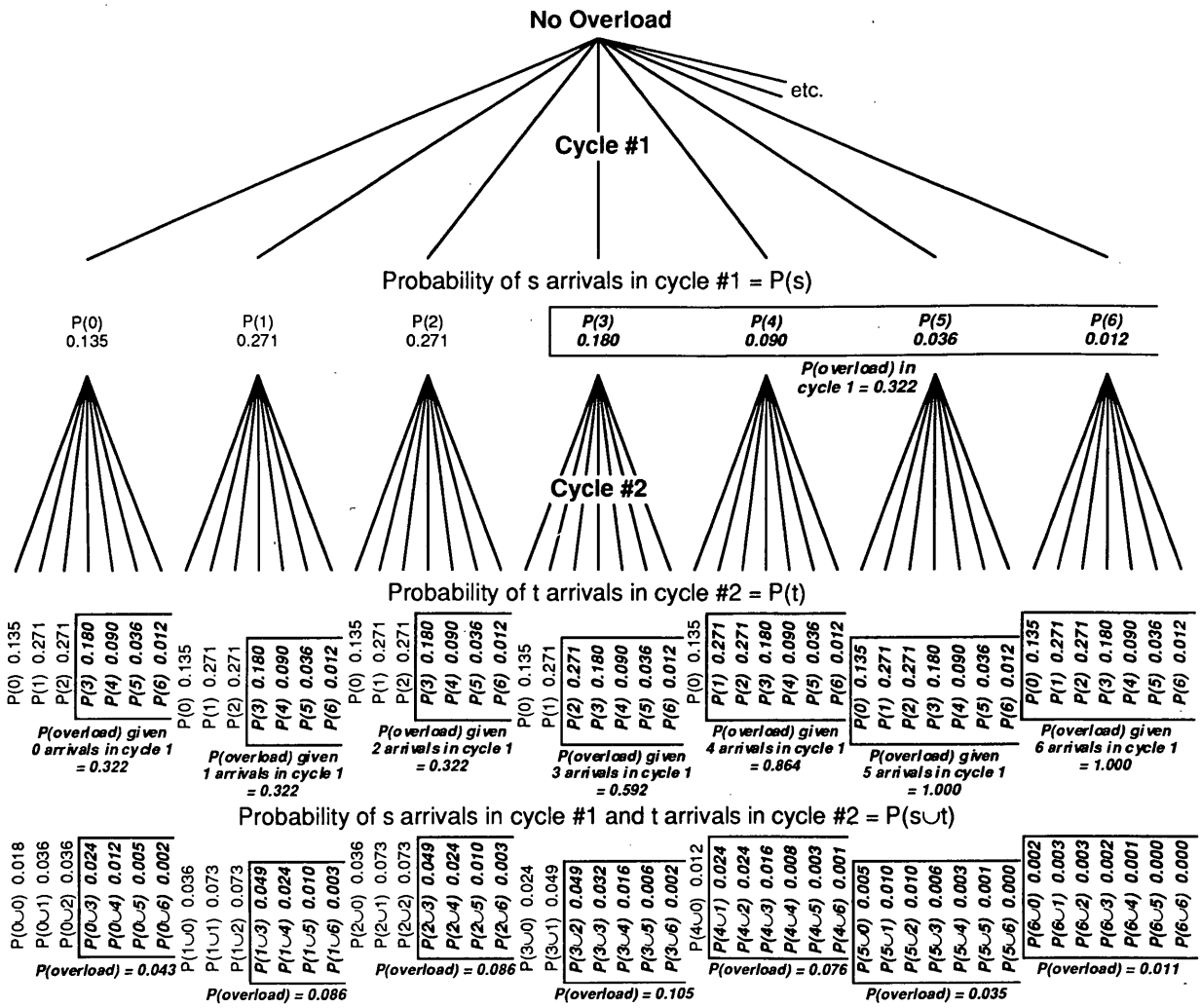


FIGURE 2 Probability tree for sequence of signal cycles (mean arrivals = 2 and cycle capacity = 2—an example of discharge overload in second of two cycles.

Given that there was no leftover queue in the previous cycle, the probability that the first cycle will be overloaded is equal to the probability of arrival overload, that is, to the probability that more vehicles arrived than the cycle can discharge.

Because the number of arrivals in individual cycles can be treated as independent events, the probability that the capacity of the second cycle will be exceeded is the sum of two probabilities as follows:

1. If fewer or equal numbers of vehicles than capacity arrived during the first cycle, the probability that more vehicles than capacity arrive in the second cycle, and
2. If more vehicles than capacity arrived during the first cycle, the sum of the probabilities that the number of leftover vehicles together with the number of vehicles arriving exceeds capacity.

Because the overload in the second cycle depends on the outcome of the first cycle, events must be treated as dependent. The dependence is reflected in the change of the probability assigned to the dependent event.

Figure 2 shows that the resulting probability can be expressed as

$$\begin{aligned} P_{\text{discharge overload in 2nd cycle}} = {} & [P(x \le X)_{1st} * P(x > X)_{2nd}] \\ & + [P(x = X + 1)_{1st} * P(x \ge X - 0)_{2nd}] \\ & + [P(x = X + 2)_{1st} * P(x \ge X - 1)_{2nd}] \\ & + [P(x = X + 3)_{1st} * P(x \ge X - 2)_{2nd}] \\ & + [P(x = X + 4)_{1st} * P(x \ge X - 3)_{2nd}] \\ & + \text{and so on} \end{aligned} \quad (6)$$

where x is the number of arrivals in a cycle and X is the cycle capacity.

The probability of a discharge overload in the first, second, or both cycles (i.e., at least one overload in these two cycles) can be determined as follows:

$$P_{\text{discharge overload in at least 1 of 2 cycles}} = [P(x \le X)_{1st} * P(x > X)_{2nd}] + P(x > X) * 1.0 \quad (7)$$

Or, in an easier way, examining the probability tree, as one minus the probability of no overload in either of the two cycles, that is

$$P_{\text{discharge overload in at least 1 of 2 cycles}} = 1 - [P(x \leq X)]^2 \quad (8)$$

Similarly, the probability of at least one overload in three consecutive cycles can be expressed as

$$P_{\text{discharge overload in at least 1 of 3 cycles}} = 1 - [P(x \leq X)]^3 \quad (9)$$

In an analogous way, one can also use the probability tree to determine the probability that two of two, two of three, or one of two cycles, and so on, will feature a discharge overload.

Figure 3 shows examples of some of such Poisson-based calculations as functions of capacity as an independent variable. Figure 3 shows (a) a simple Poisson distribution for the arrival of overload (equal to the probability of discharge overload in the first cycle); (b) distribution of discharge overloads in the second of two consecutive cycles; (c) distribution of discharge overloads in the first, second, or both (i.e., in at least one of two) consecutive cycles; (d) distribution of discharge overloads in both of two consecutive cycles; and (e) distribution of discharge overloads in at least one of three consecutive cycles. It can be seen that for the greater degrees of saturation reflected here in the greater differences between the number of arrivals per cycle and cycle capacity, the curves are close enough. Any of them could be used for practical considerations (perhaps with the exception of the overload in both cycles). This includes the curves representing arrival overload, shown as probabilities of discharge overload in the first cycle, that is, simple Poisson distributions. As capacity approaches the average number of arrivals (or vice versa in the real world), significant dissimilarities emerge.

A natural question is, What function would be a realistic representation of a true set of events during a longer period of time? Theoretically, many more than two or three consecutive cycles may be overloaded, or many combinations of overloads and number of cycles are possible. The probability of at least one overload in n cycles has been used as an example. Figure 4 shows that this probability converges when many cycles are considered. Note that with the increasing degree of saturation (i.e., m/X) the distributions become flatter and the mean simulated overload factor ($\overline{OF}$) increases.

However, there appear to be two practical limitations to the number of cycles that should be used: (a) within a real set of events, there may be many cycles with no overload (as a result, the probability tree restarts itself); and (b) longer periods of consecutive overloads (say, 10 min or more, i.e., 5 to 10 cycles) usually result from a sudden surge in traffic demand (from a change of the arrival mean). As a consequence of consecutive overload, events are not random any more, and this condition may be hidden from the analyst, even when a peak hour factor (*1*) is considered.

An additional argument lies in the basic notion of the probability of independent events—probability defined as the number of cases favorable divided by the total number of cases. Whereas this principle applies within the different levels of the probability tree, it is not valid for the whole chain of events because the individual probabilities change with the levels (i.e., are dependent). Figure 4 illustrates this point. For the first few cycles the probability that there will be an overloaded cycle is represented by a definite number for any capacity. If, however, capacity equals or is less than the number of arrivals (i.e., $X \leq 6$), the chances that a cycle within the series will be overloaded quickly approach certainty with increasing time (i.e., with the number of cycles considered).

The last deliberation also illuminates the difference between the 1965 HCM load factor and the probability of discharge, defined as the probabilities of overloads in the second or the third cycle, one overload in two cycles, two overloads in three cycles, or in any other way. The load factor was considered more as a sequence of independent events than a chain of dependent events. It was also frequently interpreted as the chance of an individual cycle being overloaded. This way of looking at the problem led quite naturally to its representation by the probability of arrival overload, that is, by a simple Poisson distribution.

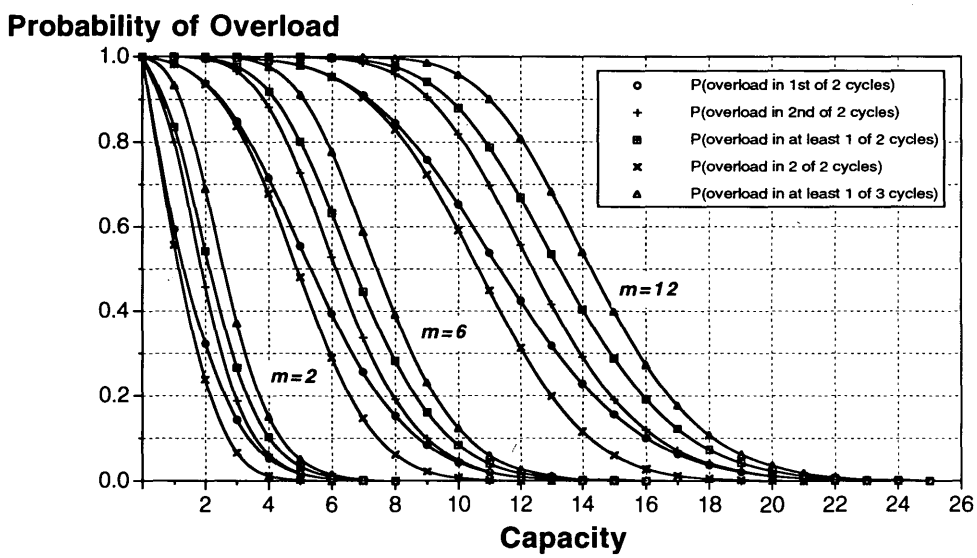


FIGURE 3 Probability distributions of overloads for mean number of arrivals in cycle (m).

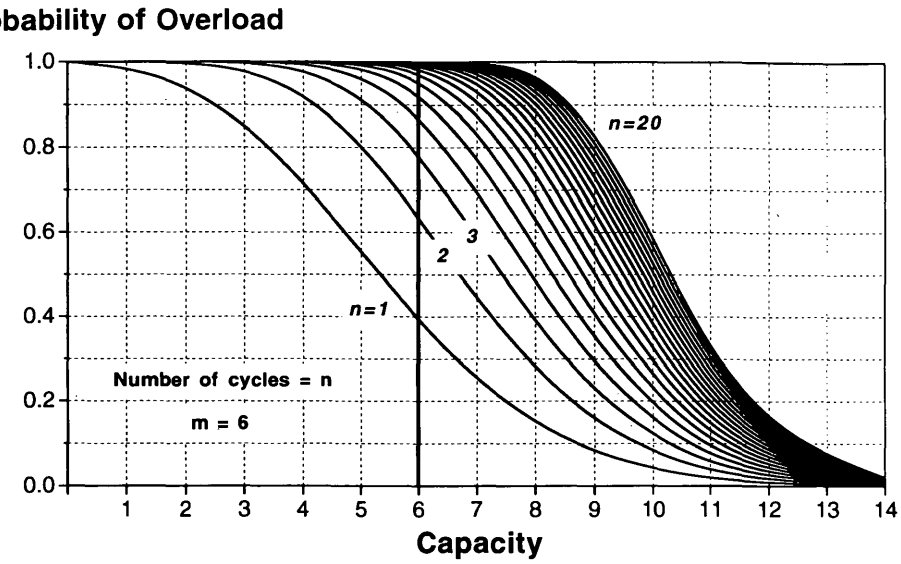


FIGURE 4 Probability of at least one discharge overload in *n* cycles for mean number of arrivals.

SURVEYS AND SIMULATION OF OVERLOAD FACTOR

The importance of knowing how many cycles are or may be overloaded justifies a search for an appropriate measure. To seek guidance as to which of the probabilities of discharge would represent a practical calculable measure, overload factors were studied by both simulation and surveys in a pilot project.

The overload factor was defined as the 1965 HCM load factor with one important exception: only those fully loaded cycles in which at least one vehicle was unable to discharge at the end were included [i.e., similar to Miller's percentage of cycles in which the queue is not cleared (*10*)].

Surveys

In summer 1992 two intersection lanes in Edmonton were surveyed during three types of traffic conditions. The arrival rates for individual surveys were tested on the stability of the mean and were found satisfactory. The cycle capacity was determined directly as the discharge during fully loaded cycles. A total of 10 surveys included 292 signal cycles. Table 1 gives the results.

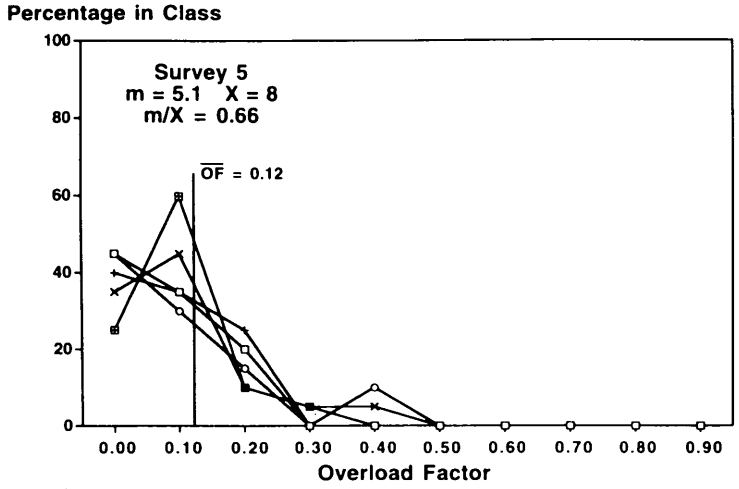
Simulation

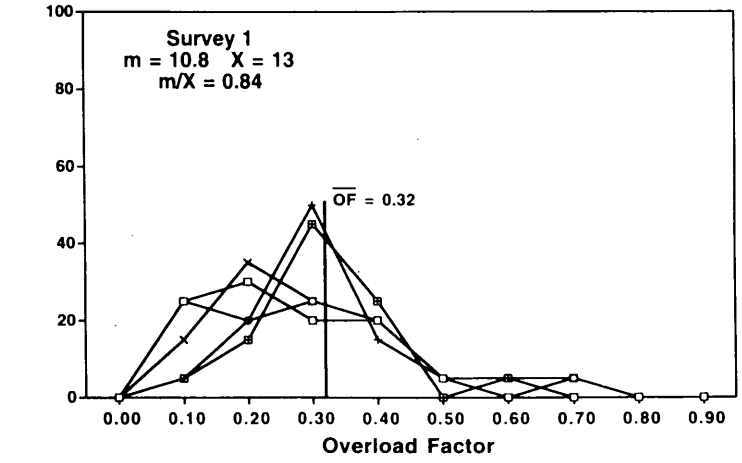
Additional overload factors were obtained from a spreadsheet-based simulation program that used a Poisson distribu-

TABLE 1 Overload Factors Measured at Three Edmonton Situations

Survey Number	Intersection	Period / No. of Cycles	Direction / Lane	Arrival Mean (per cycle)	Measured Capacity (per cycle)	Measured Overload Factor
1	87 Avenue / 109 Street	PM 21	EB 2	10.8	12.9 (13)	0.286
2	87 Avenue / 109 Street	PM 22	EB 2	9.3	12.2 (12)	0.227
3	87 Avenue / 109 Street	PM 26	EB 2	9.2	12.6 (13)	0.154
4	72 Avenue / 114 Street	AM 22	WB 1	5.9	7.6 (8)	0.455
5	72 Avenue / 114 Street	PM 27	WB 1	5.1	7.7 (8)	0.111
6	72 Avenue / 114 Street	PM 26	WB 1	5.3	7.1 (7)	0.192
7	72 Avenue / 114 Street	PM 36	WB 1	5.8	6.8 (7)	0.222
8	72 Avenue / 114 Street	PM 59	WB 1	5.2	7.0 (7)	0.169
9	87 Avenue / 109 Street	PM 27	EB 2	8.7	12.2 (12)	0.074
10	72 Avenue / 114 Street	AM 26	WB 1	6.6	7.0 (7)	0.500

Numbers in ( ) represent rounded values.





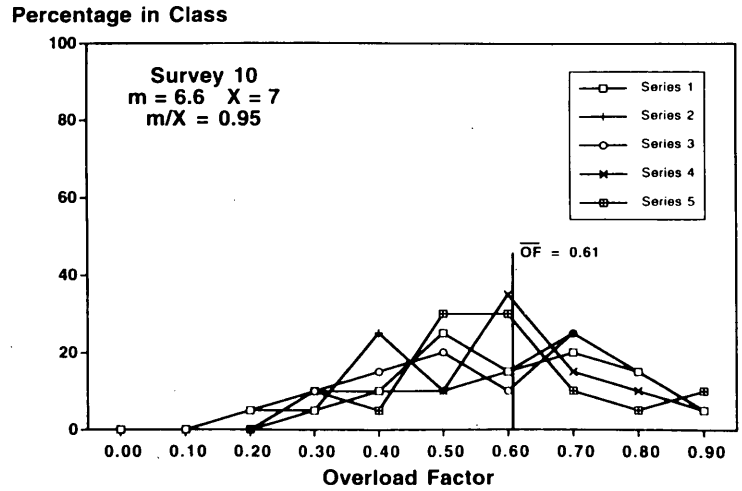


FIGURE 5 Probability distributions of overload factors simulated for arrival and capacity conditions from surveys (m = average number of arrivals per cycle and X = cycle capacity).

tion for vehicle arrivals. All surveyed sets of mean arrivals and capacities were included in the simulation process of this pilot study. Each set was simulated in 10 series of 20 computer runs representing 20 cycles each, thus incorporating 10 × 10 × 20 × 20 = 40,000 cycles and resulting in 10 × 10 × 20 = 2,000 overload factors.

Because of the differences in the surveyed conditions (mean arrivals and capacities) as well as the small number of surveys, it was not considered prudent to investigate the distributions of measured overload factors. On the other hand, simulation provided sufficient data to that end. Probability distributions of the simulated overload factors for three conditions are shown in Figure 5. They exhibit remarkable consistency for each of the three simulated conditions and document the suspected wide range of possible values for higher degrees of saturation. This observation can be interpreted as a greater risk of surprising system failures for closer to capacity conditions. The practical meaning is that on some days, under identical arrival and signal timing conditions as before, many more overloaded cycles may happen.

A related interesting observation was made for the three surveys for which the measured or simulated overload factor exceeded 0.3. 1965 HCM identified load factor over that value as approaching unstable flow. Surveys and many simulation runs showed a number of consecutive overloads, which, indeed, are indicative of unstable conditions.

COMPARISON OF OVERLOAD FACTORS WITH PROBABILITIES OF OVERLOAD

Figure 6 shows the comparison of simulated overload factors with the various forms of probability of discharge overload. The probability of arrival overload is included as the probability of discharge overload in the first cycle. It is encouraging to see the consistency of the patterns. Generally, the simulated overload factors fall between the probabilities of a discharge overload in the second cycle and a discharge overload in at least one of two cycles. Probability of at least one overload in two cycles would be a good approximation for all overload factors and offers a somewhat conservative estimate.

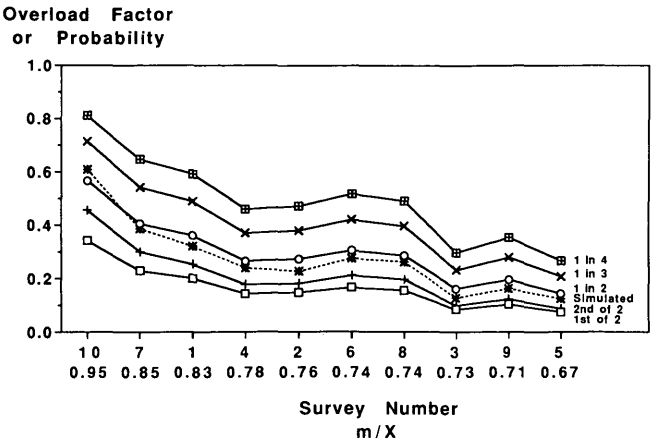


FIGURE 6 Comparison of simulated overload factors with different forms of probability of discharge overload (ranked by their degrees of saturation m/X, where m = average number of arrivals and X = cycle capacity).

Figure 7 depicts the overload factor values generated by individual simulation series and identifies the ranges of ±1 standard deviation. Because the measured overload factors are included, comparisons are possible, keeping in mind the previously discussed problems associated with a one-point measurement in an unknown distribution and the wide range of values for identical input values. Under the assumption of a normal distribution of overload factors, about two-thirds of all values should fall within the indicated range. The measured overload factor falls well within that range for most of the surveys. An examination of the distributions in Figure 5 indicates that even for Surveys 7 and 4, the measured values

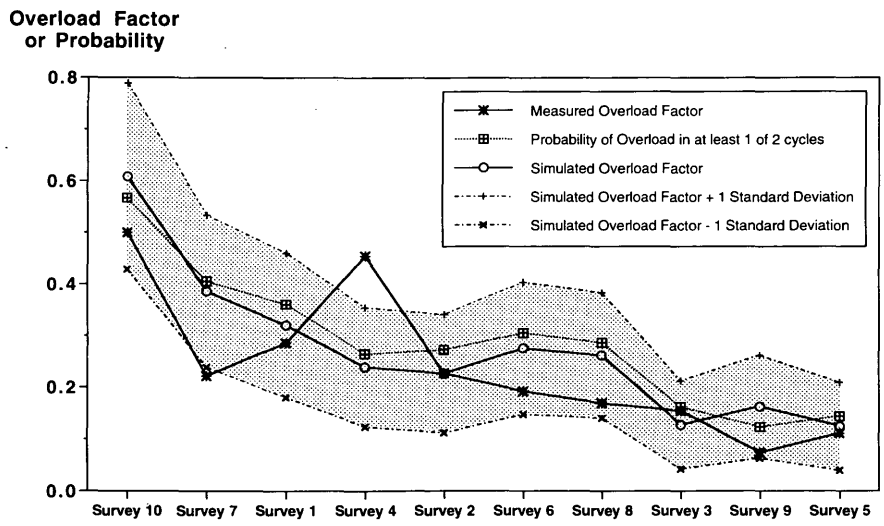


FIGURE 7 Comparison of simulated and measured overload factors with probability of discharge overload in at least one of two consecutive cycles (range represents ± 1 standard deviation from simulation series).

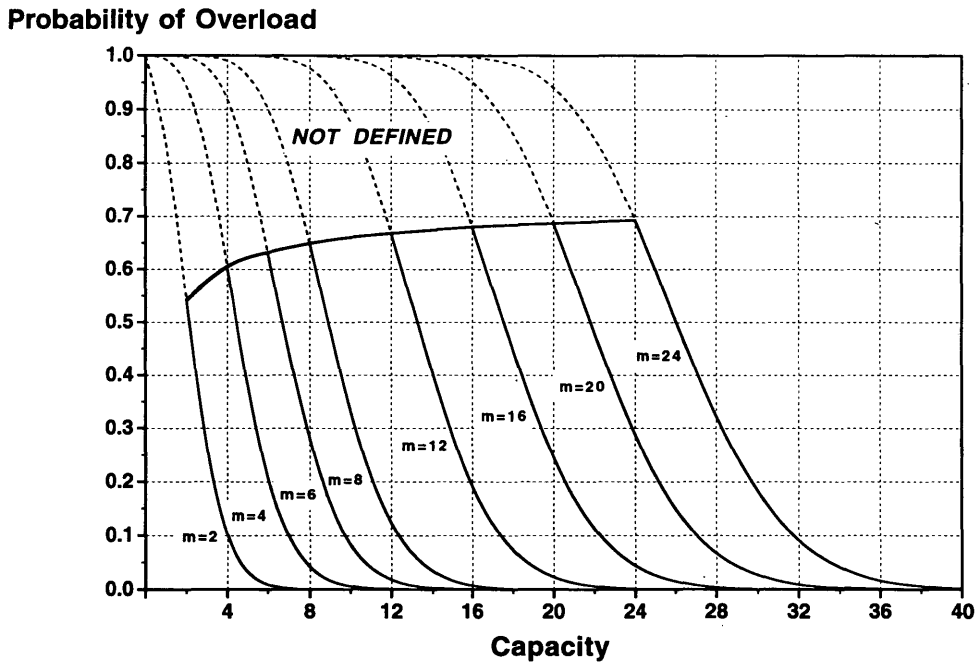


FIGURE 8 Overload factors shown in a tentative system of probability of discharge overload represented by probabilities of at least one overload in two consecutive cycles.

are still quite representative of the main distribution classes. The limited scope of data in this pilot study did not provide sufficient basis for rigorous statistical testing.

FINDINGS AND CONCLUSIONS

This pilot study indicates that a close link exists between the probabilities of discharge overload and the overload factor. The comparison of the measured and simulated overload factors with the probabilities of discharge overload shown in Figure 7 suggest that for the studied situations, the overload factors may be approximated by the formula for the probability of discharge overload in at least one of two consecutive cycles.

Assuming that the conditions described by the arrival mean and capacity have already lasted for some time, the probability values for oversaturated situations are 1.0, and the calculated values do not apply. This constraint is shown in Figure 8, which illustrates the tentatively proposed system to determine the probability of discharge overload.

An extension of this University of Alberta study will examine the problem in detail on the basis of broad field and simulation data. The results so far provide evidence that it is possible to devise a practical measure that would allow the analyst or designer to determine the seriousness of random overloads. Such an evaluation criterion would be instrumental in understanding suitable signal coordination conditions and assessing the impact of the random delay component or of all three traditional delay elements (uniform, random, and continuous). The newly defined overload factor would represent an easy way of verifying the magnitude of the probability of discharge overload by surveys.

ACKNOWLEDGMENTS

The pilot study reported in this paper was funded under an operation grant by the Natural Sciences and Engineering Research Council of Canada. The author would also like to express his appreciation and thanks to two University of Alberta students: Sylvia Dzido assisted in the surveys, and Randall Sonnenberg was responsible for the field work, many calculations, and final figures. The reviewers of this paper are acknowledged with gratitude for their valuable comments and suggestions.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. *Special Report 87: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1965.
3. *Canadian Capacity Guide for Signalized Intersections*, 1st ed. (S. Teply, ed.). Institute of Transportation Engineers, District 7, Canada, 1984.
4. S. Teply. Quality of Service in the New Canadian Signal Capacity Guide. In *Highway Capacity and Level of Service* (U. Branolte, ed.). Balkema, Rotterdam, The Netherlands, 1991.
5. R. Akcelik. *Traffic Signals. Capacity and Timing Analysis.* Research Report 123. Australian Road Research Board, Victoria, 1983.
6. A. D. May and D. Pratt. A Simulation Study of Load Factor at Signalized Intersections. *Traffic Engineering*, ITE, Washington, D.C., Feb. 1968.
7. Y. B. Chang and D. S. Berry. Examination of Consistency in Signalized Intersection Capacity Charts of the Highway Capacity Manual. In *Highway Research Record 289*, HRB, National Research Council, Washington, D.C., 1969.
8. E. F. Reilly and J. Seifert. Capacity of Signalized Intersections. In *Highway Research Record 321*, HRB, National Research Council, Washington, D.C., 1970.

9. J. E. Tidwell and J. B. Humphreys. Relation of Signalized Intersection Level of Service to Failure Rate and Average Individual Delay. In *Highway Research Record 321*, HRB, National Research Council, Washington, D.C., 1970.
10. A. J. Miller. On the Australian Road Capacity Guide. In *Highway Research Record 289*, HRB, National Research Council, Washington, D.C., 1969.
11. B. M. Davidson. Traffic Signal Timing Utilizing Probability Curves. *Traffic Engineering*, ITE, Washington, D.C., Nov. 1961.
12. S. Teply and B. Stephenson. *Micro-Sintral User Manual*. Institute of Transportation Engineers, District 7, Canada, 1990.
13. D. R. Drew and C. Pinnell. A Study of Peaking Characteristics of Signalized Urban Intersections as Related to Capacity and Design. *Bulletin 352*, HRB, National Research Council, Washington, D.C., 1962.
14. C. J. Messer and D. B. Fambro. Critical Lane Analysis for Intersection Design. In *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977.

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

Signal Design for Congested Networks Based on Metering

AYELET GAL-TZUR, DAVID MAHALEL, AND JOSEPH N. PRASHKER

A method is presented for designing signal programs for congested urban networks in which one or more intersections are bound to become oversaturated before the others. The idea on which the method is based is to limit the volumes entering the network, to adjust them to the capacity of the critical intersection, and to avoid intersection blockage. This metering procedure enables the designer to determine the location of the queues and to direct them to long links that can act as buffers. This strategy is called queue location management. The procedure is based on integrating a mathematical programming model with TRANSYT. The model is used to determine the green splits and the level of metering, whereas TRANSYT is used for simulating the dynamic processes within the cycle and for offset optimization. A numerical example demonstrates the procedure.

Congestion in urban networks is one of the most urgent problems with which traffic engineers are confronted. McShane and Roess (*1*) emphasized that oversaturation is not merely another level of the common traffic problem—delays and stops—it is a new kind of problem. The reason has to do with the unique phenomenon typical of such conditions, namely, intersection blockage. Blockage or spillback occurs when a long queue starting at one intersection continues into the one upstream. In addition to causing a deterioration in the capacity of the intersection in which it occurs, blockage eventually leads to a decrease in the capacity of the vicinity. The throughput of the blocked intersections drops rapidly, vehicles accumulate in queues, and routes that regularly face normal conditions become congested. Preventing such situations is the major goal of traffic engineers.

Various researchers have investigated these conditions. Gazis (*2*), one of the pioneers in this field, observed the problem of the critical intersection and tried to optimize its performance. He considered a single intersection with a two-phase control and offered a strategy composed of two programs, one to be implemented during the first part of the rush hour and the other during the second part. Michalopoulos and Stephanopoulos (*3,4*) continued this work and tried to broaden the problem by making modifications in an algorithm. Overflow was handled by limiting queue lengths. They also formulated the interdependencies among intersections in the network and solved the equations that represented all intersections simultaneously. Singh and Tamura (*5*) approached the problem from the network point of view. Assuming that demand exceeds capacity at all links, they described the traffic flow by means of a set of differential equations. Their model tried to minimize queue lengths by adjusting green splits.

Civil Engineering Department, Technion-Israel Institute of Technology, Haifa, Israel.

Lieberman (*6*), McShane and Roess (*1*), and Kirby (*7*) all suggested metering the number of vehicles allowed into the network. The phenomenon that demand levels above capacity exist for short periods during the rush hour was the basis for this remedy. The solution is well known for freeway control (*8*). Lieberman distinguished between the treatment given to a major arterial and that to a critical intersection. He formulated various objective functions for controlling traffic flow along an arterial that consider queues and spillbacks explicitly and that emphasize the goal of efficient use of several parts of the arterial as buffers. For critical intersection (CI) control, Lieberman presented a mathematical model suitable for a network that contains at least one CI. The objective function in this case was the release of the queues in the various approaches to the CI according to predetermined priorities. McShane and Roess listed three types of metering: internal, external, and release. Internal metering controls the discharge rate at internal links of the network in order to prevent too many vehicles from reaching critical intersections. External metering limits the number of vehicles entering the network. Release metering refers to regulating the exit of vehicles from parking areas, garages, and lots. Kirby described research undertaken at the University of Leeds that was concerned with improving and applying metering strategies or, as it is called by the author, gating strategies.

During peak demands, as it is clear that queues are inevitable, there is need to determine their location. Thus, distributing the queues over the network in such a way that the damage that they may cause will be minimal (especially avoiding the damage of spillbacks) should be the explicit goal of the control process. The authors call this planning strategy queue location management (QLM), and it is suited for congested periods, when the capacity of an interior network is lower than the demand at its entry links. The QLM procedure, which integrates a mathematical model with TRANSYT, is described in this paper. The mathematical model is used for metering, whereas TRANSYT serves both as a planning tool and as a simulation tool. TRANSYT is the most widespread tool for planning signal-timing programs (*9*). Its main advantage lies in the overall network approach it takes to determine the cycle time, green splits, and offsets. It does not guarantee a global optimum, but it searches for a good solution while taking many parameters into account (e.g., platoon dispersion characteristics). It also allows the user to input subjective priorities among the various objectives. Despite the intensive use of TRANSYT, it is not suitable for dealing with congested networks.

TRANSYT has two drawbacks in handling oversaturated conditions. The first is that queue lengths are given as an

output of the program but are not considered either in the objective function or as a constraint. The other problem, which is more severe, stems from the requirement to input the volumes on each link before the design process. In uncongested networks, when there is a cycle failure, these volumes are simply the sum over the demand of all movements leading to a certain link. As the network becomes more congested and queues accumulate, the volume of certain links is determined by the green duration. In these cases, the volume of the downstream intersection is not the demand but the throughput of upstream intersections. In other words, during congestion volumes are sometimes determined by the throughput and green durations. Thus a vicious circle is created. TRANSYT requires the volumes in order to calculate signal programs, but volumes are determined by the green durations. This vicious circle can be solved by simultaneously determining the volumes and green durations. This can be done by combining TRANSYT with an external program. Integrating TRANSYT with other procedures has already been implemented for arterial signal design (*10*).

A detailed method of applying the ideas of metering to congested urban networks is proposed in this paper. The different stages of the design process are described in the next section. Then the mathematical model is formulated, and in the final section the procedure is illustrated.

PROPOSED APPROACH

The approach proposed combines two modules into a complete algorithm that implements the concept of metering in a way that reflects reality in a reliable manner. The QLM strategy is applied to the entire network by a mathematical model that considers one cycle as the smallest time unit. The dynamic processes within the cycle, such as platoon dispersion and queue formation, are taken into account in the second module, in which TRANSYT is used. The method was designed for networks in which one intersection is known to have a smaller capacity than its demand, and thus is bound to become a CI. Under congested conditions, the throughput of the CI determines the throughput of the network in the vincinity, and therefore the optimal operation of the CI becomes the major goal.

During certain peak periods, experience indicates that demand is greater than the overall network capacity and thus queues are inevitable. Such peak periods are generally not very long, but their consequences affect traffic-flow conditions long after the demand decreases to below capacity. These lasting effects result from the chain reaction of long queues and spillbacks. Therefore, efforts should be made toward minimizing the damage caused by the temporary imbalance between capacity and demand, and special attention should be given to the prevention of spillbacks.

The following approach is recommended for the short periods when demand is higher than capacity. The goal of the design process is to prevent overflows at critical links by metering the amount of vehicles entering them. The residual amount of vehicles can be accumulated at long links at the entrances of the network, which play the part of buffers. Alternatively, vehicles at these entry links are rerouted to alternative paths either by active means of appropriate signs or by passive means (as drivers, encountering long queues, choose a different route on their own).

A prerequisite for the implementation of the foregoing strategy is to define the subnetwork surrounding the CI, for which signal programs should be redesigned. Such a subnetwork is characterized by entry links that are either long enough to be used as buffers or able to lead to alternative routes. The traffic volumes at the entrances of the network are assumed to be constant during the planning horizon considered. It is also assumed that vehicle paths are predetermined. In those cases in which this assumption is unrealistic, an assignment program should be integrated with the control process.

The terms used in this paper are defined as follows:

- Demand—the number of vehicles per time unit arriving at an entry movement;
- Volume—the number of vehicles per time unit arriving at an internal link movement; and
- Throughput—the number of vehicles per time unit discharged from any movement.

The distinction between demand and volume is necessary because the volume of an internal link is determined either by demand or by the capacity of external links.

The calculation algorithm consists of two types of variables: the volumes within the network and the green durations of all intersections. The algorithm is composed of four stages, which are described in Figure 1 and the following paragraphs.

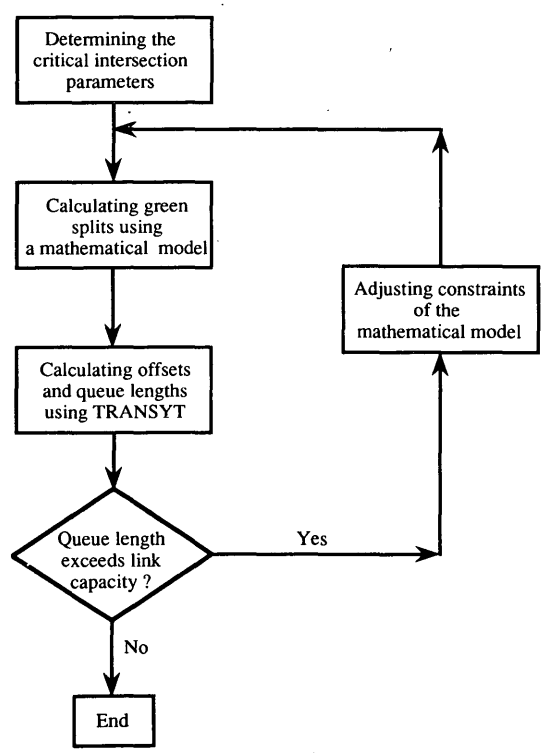


FIGURE 1 Flow chart of the proposed procedure.

Stage 1—Determining Critical Intersection Parameters

The purpose of the first stage is to find the cycle time and green splits of the CI that would achieve maximum throughput. The output at this stage is the required throughput of each movement of the CI and the suitable green durations that would enable obtaining the throughput. To avoid infinite queues, the volumes of these movements should be equal to their throughput. Several ways of calculating these parameters appear in the literature. For example, the means suggested by Mahalel and Gur (*11*) explicitly considers the phenomenon of a decay in saturation flow as the green duration extends beyond a certain value. This phenomenon calls for a short green duration. There is a trade-off between shortening the green durations in order to achieve high saturation flows and prolonging the cycle in order to minimize the fraction of lost time. Maximum throughput is obtained by determining the green durations in a way that the marginal saturation flow is identical for all phases. At this point a maximum average saturation flow is also achieved.

Stage 2—Calculating Green Splits of All Intersections

After the optimal parameters of the CI are determined, the green splits of all other intersections in the network are calculated so that spillbacks are avoided. This objective is achieved by metering the flow of the links leading to the critical movements in such a way that the volumes arriving at each critical movement equal its predetermined throughput. The calculation process is based on the following assumptions:

- The cycle time determined for the CI at the first stage is common to all intersections in the network;
- The demands and saturation flows of the network are deterministic; and
- The smallest time unit considered by the model is one cycle, and therefore the values of the network parameters (demand, queues, etc.) are calculated for the end of each cycle.

As the demand is assumed to be deterministic and constant, the queue length of each movement might either grow at a fixed rate to infinity or be equal to zero. Defining the queue length of an internal link as zero means that the number of vehicles that can be discharged during the green phase must be greater than or equal to its volume. Thus, the throughput of any internal movement equals its volume. As for entry movements, their goal is to ensure that the number of vehicles entering the network should not exceed its capacity. These are the locations at which queues can accumulate if metering is applied. The throughput of these movements is the minimum between their demand and their capacity (Equation (1). In cases in which the green duration is long enough to discharge all vehicles arriving during the cycle (undersaturated conditions), throughput is identical to the demand; otherwise (in oversaturated conditions), it equals the discharge rate per cycle.

$$\text{throughput}\ (m) = \min\ [\text{demand}\ (m); \text{capacity}\ (m)] \qquad (1)$$

where m is an entry movement. The volume of each internal movement is the sum of the throughputs of all movements leading to it.

There are two possible means of applying metering. One is to prolong the green duration of undersaturated movements, if there are such movements at the intersection being considered. If metering is needed for all movements, the "all-red" period should be prolonged beyond the minimum duration required for safety considerations. Such an act might cause difficulties if this period becomes long enough to result in a flow during the all-red interval. A constraint for avoiding such cases can be added to the planning algorithm, although it is not included in the formulation given in this paper. A mixed-integer linear programming (MILP) model was developed for this stage and is presented in detail in the next section.

Stage 3—Calculating Offsets and Queue Lengths of All Intersections

After the green splits and volumes of each phase and movement are determined, they are input into TRANSYT and TRANSYT is used for offset calculation. Additional information derived is the maximum queue length on each link during the cycle. This information cannot be obtained from the mathematical model, as it represents conditions in the network only at the end of the cycle. For every internal link, a comparison is to be performed between maximum queue length and link storage capacity in order to track overflows.

Stage 4—Adjusting Parameters of Mathematical Model

In cases for which overflow is tracked, adjustments in the MILP model are required. The modifications are done by enlarging the difference between the volume and the capacity of these movements. After the necessary adjustments are completed, an iterative process is conducted by repeating Stages 2 through 4. A full example demonstrating the execution of the four stages is given subsequently.

MATHEMATICAL MODEL

The mathematical model is used in the second stage of the solution process and is based on the CI parameters (cycle time and green split), which were calculated in the first stage. As was noted in the previous section, the throughput of the whole network is determined by that of the critical intersection, which thus is an input for the mathematical model. The role of the model is to determine the throughput of each intersection surrounding the CI. A MILP model was developed, in which the integer variables distinguish between oversaturated and undersaturated entry movements (see Equation 1). This distinction is necessary in order to calculate the throughput of these movements.

The parameters of the design process are intended to maximize the throughput of the CI. Any feasible solution of the mathematical model that satisfies this purpose is also optimal.

The parameters of the CI appear in the constraints of the model. The terms used in the model are defined as follows:

- Movement—a group of vehicles having the same destination;
- Link—a set of movements sharing the same approach; and
- Phase—a set of movements receiving the right of way simultaneously.

The goal of the mathematical model is to find a feasible solution that achieves the following constraints. The first set of constraints (Equation 2) simply states that the sum of the green durations of all phases should be less than or equal to the difference between the cycle time and the total lost time. For cases in which metering is determined by the model, the constraint satisfies the "<" sign, and a positive slack variable is added.

$$\sum_j G_{ij} \leq C - F_i \cdot L \qquad \forall i \tag{2}$$

where

C = common cycle time of network (sec)
L = lost time for each phase (sec),
G_{ij} = green duration of phase j at intersection i (sec),
F_i = number of phases at intersection i,
i = intersection index, and
j = approach index.

Note that G_{ij} is a decision variable, whereas C and L are predetermined parameters.

The second set of constraints refers to the movement of the CI only (Equation 3). The volume of any critical movement (a movement of the CI) should be equal to its throughput, as was determined in the first stage.

$$P_{\mathrm{CI}lm} \cdot \sum_{kr \in \Omega_{\mathrm{CI}l}} O_{ksr} = T_m \qquad \forall\, l \in \mathrm{CI} \qquad \forall m \tag{3}$$

where

T_m = maximum throughput of movement m of CI, as calculated in first stage;
Ω_{il} = set of movements leading to link l of intersection i (thus $\Omega_{\mathrm{CI}l}$, in particular, is set of movements leading to link l of intersection CI);
O_{ksr} = throughput per cycle of movement r of link s of intersection k (notations r, s, and k being used to indicate movements leading to link l of CI);
P_{ilm} = proportion of vehicles in movement m of the total volume of link l of intersection i;
l = link index; and
m = movement index.

Thus $\Sigma_{ksr \in \Omega_{il}} O_{ksr}$ is the volume of link l of intersection i, and $P_{ilm} \cdot \Sigma_{ksr \in \Omega_{il}} O_{ksr}$ is the volume of movement m of link l of intersection i. Note that T_m and P_{ilm} are parameters and O_{ilm} is a decision variable.

For all other internal movements in the network, the volume should equal the throughput in order to avoid queues (Equation 4). In contrast to movements of the CI, however, both the volume and the throughput of each internal movement are unknown.

$$P_{ilm} \cdot \sum_{ksr \in \Omega_{il}} O_{ksr} = O_{ilm} \qquad \forall \text{ internal movements} \tag{4}$$

The next set of constraints (Equation 5) ensures that the green duration of the phase to which movement m belongs is long enough to discharge the throughput determined in Equation 4).

$$S_{ilm} \cdot G_{ij} \geq O_{ilm}$$

$$\forall \text{ internal movements (where } m \in \text{phase } j) \tag{5}$$

where S_{ilm} is the saturation flow of movement m of link l of intersection i, and $G_{ij} \cdot S_{ilm}$ is the maximum number of vehicles that can be discharged from movement m of link l of intersection i per cycle.

As for entry links, the number of vehicles entering each link is a predetermined demand D_{il}, and thus the demand per entry movement is $P_{ilm} \cdot D_{il}$. The next set of constraints allow the planner to limit the queue-building rate per cycle at each link (Equation 6).

$$D_{il} - \sum_{m=1}^{M_l} O_{ilm} = q_{il} \qquad \forall \text{ entry links} \tag{6}$$

where D_{il} is the demand per cycle of link l of intersection i, and q_{il} is the queue-building rate per cycle of link l of intersection i.

Equations 7 and 8 define the 0–1 variables (X_{ilm}), which indicate whether the throughput of entry movement m of link l of intersection i is determined by its demand rate ($P_{ilm} \cdot D_{il}$). In such a case, X_{ilm} receives the value of 0; or is determined by its capacity ($S_{ilm} \cdot G_{ij}$) and receives the value of 1.

$$S_{ilm} \cdot G_{ij} - P_{ilm} \cdot D_{il} + M \cdot X_{ilm} \geq 0$$

$$\forall \text{ entry movements} \tag{7}$$

$$P_{ilm} \cdot D_{il} - G_{ij} \cdot S_{ilm} + M \cdot (1 - X_{ilm}) \geq 0$$

$$\forall \text{ entry movements} \tag{8}$$

where

$$X_{ilm} \begin{cases} = 1 & \text{if } P_{ilm} \cdot D_{il} \geq S_{ilm} \cdot G_{ij} \\ = 0 & \text{otherwise} \end{cases}$$

and

M = a very large number.

The throughput per cycle of each movement (O_{ilm}) is the minimum between its demand and its saturation flow per cycle. The minimum operant can be expressed by the variables X_{ilm} as represented in Equation 9.

$$X_{lm} \cdot S_{ilm} \cdot G_{ij} + (1 - X_{ilm}) \cdot P_{ilm} \cdot D_{il} = O_{ilm}$$

$$\forall \text{ entry movements} \tag{9}$$

This equation cannot be part of the MILP model because it is not linear. Thus a new variable is created, according to Equation 10, which is a product of two original variables.

$$SX_{ilm} = X_{ilm} \cdot S_{ilm} \cdot G_{ij} \quad (10)$$

The new variable turns Equation 9 into Equation 11.

$$SX_{ilm} + (1 - X_{ilm}) \cdot P_{ilm} \cdot D_{il} = O_{ilm}$$

$$\forall \text{ entry movements} \quad (11)$$

Finally Equations 12 to 14 define the variable SX_{ilm} by linear equations.

For practical reasons, a minimum green-duration constraint is added to the model (Equation 15).

$$SX_{ilm} - M \cdot X_{ilm} \leq 0 \quad \forall \text{ entry movements} \quad (12)$$

$$SX_{ilm} - S_{ilm} \cdot G_{ij} \leq 0 \quad \forall \text{ entry movements} \quad (13)$$

$$S_{ilm} \cdot G_{ij} - SX_{ilm} \leq M \cdot (1 - X_{ilm})$$

$$\forall \text{ entry movements} \quad (14)$$

$$G_{ij} \geq \text{Mingreen} \quad \forall i \; \forall j \quad (15)$$

The objective function of the mathematical model is to minimize the queue-building rates at the entry links of the network (Equation 16).

$$\text{Min} \sum_{i \in C1} \sum_{j \in C2} W_{il} \cdot q_{il} \quad (16)$$

where

$C1$ = set of entry intersections,
$C2$ = set of entry links, and
W_{il} = weights that can be added to represent relative importance of each link.

The next section will demonstrate the implementation of the above model to a small network.

NUMERICAL EXAMPLE

The network described in the numerical example consists of six intersections, one of which is critical and marked as Intersection 10 (see Figure 2). The network has 8 entry links. At all intersections, a separate phase is devoted to each approach, and thus the link index and the phase index may be combined. Saturation flows of all traffic movements are 1,800 vehicles per hour (vph) (0.5 vehicles per second); the demand on each entry link is 540 vph; and the lost time per phase is 3 sec. Figure 2 presents a schematic description of the network topology and the proportion of vehicles turning at each movement.

The first stage of the design process, in which the CI optimal parameters are calculated, will not be described here in detail, as it is not the main issue of this paper. The cycle time was

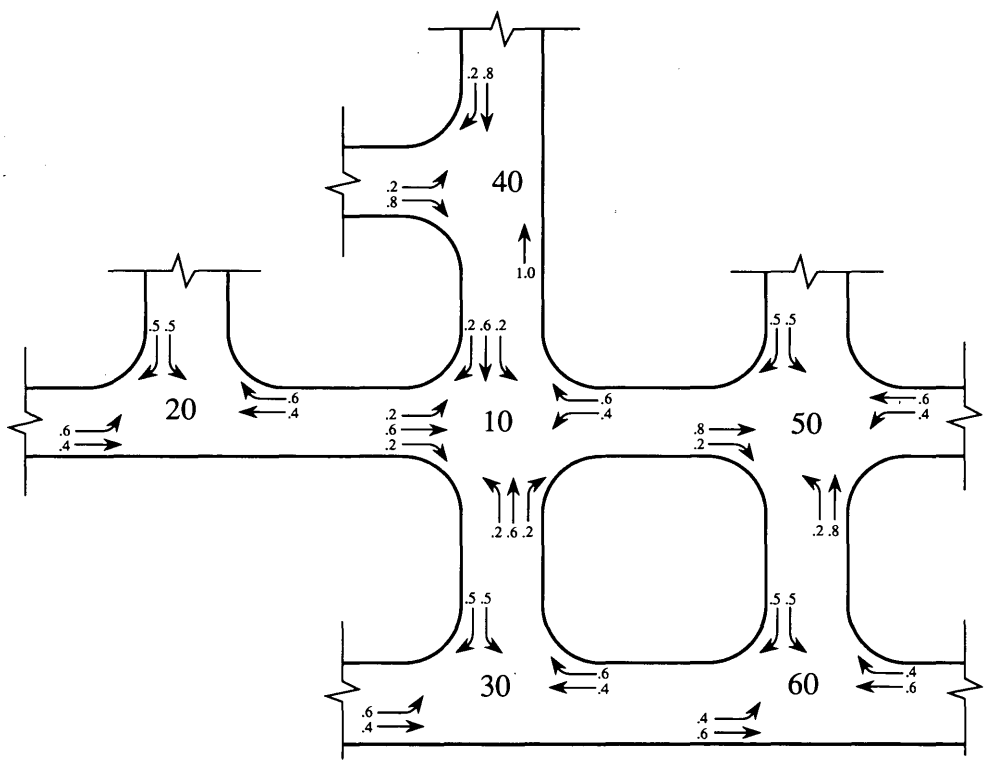


FIGURE 2 Network structure and proportion of turning vehicles of each link.

selected to be 100 sec, and the aggregate throughput of the critical movements is as follows:

- Northbound, 504 vph;
- Westbound, 216 vph;
- Southbound, 504 vph; and
- Eastbound, 360 vph.

In the second stage of the algorithm, the general formulation of the MILP model, as described in the preceding section, turns into a detailed set of equations representing the network. Following are some equations that are part of the full mathematical model formulated for the example given. Each movement is identified by a combination of six digits; for example,

XXYYZZ

where

XX = intersection index (i) (according to the intersection numbering in Figure 2);
YY = link index (l) (10 for northbound links, 20 for westbound links, 30 for southbound links, and 40 for eastbound links); and
ZZ = movement index (m) (10 for a right-turn movement, 20 for a through movement, and 30 for a left-turn movement).

The objective function is as follows:

$$\text{Min}[q_{2030} + q_{2040} + q_{3040} + q_{4030} + q_{4040} + q_{5020} + q_{5030} + q_{6020}] \quad (17)$$

In this equation, equal importance was assigned to all entry links.

Equation 18 is according to Equation 2 and refers to Intersection 30. This intersection has three approaches (phases), each having a lost time of 3 sec, and thus its effective green time should not exceed 91 sec ($100 - 3 \cdot 3 = 91$).

$$G_{3020} + G_{3030} + G_{3040} \leq 91 \quad (18)$$

Equation 19 is a demonstration of Equation 3 for critical movement 101020. Equation 19 sets the volume of Movement 101020, originated from Movements 302010 and 304030, to be equal to the throughput of this critical movement (8.4 vehicles per cycle).

$$0.6 \cdot (O_{302010} + O_{304030}) = 8.4 \quad (19)$$

The next constraints (Equations 20 and 21) refer only to internal movements and are according to Equations 4 and 5. For example, movement 302010 is an internal movement. Equation 20 sets its volume (which is the sum of the throughputs entering it) to equal its throughput, so that no queues build up.

$$0.6 \cdot (O_{602010} + O_{603020}) = O_{302010} \quad (20)$$

Equation 21 states that the capacity of the internal movement is greater than or equal to its volume. The capacity is determined by the green duration multiplied by the saturation flow per second (0.5).

$$0.5 \cdot G_{3020} \geq O_{302010} \quad (21)$$

Movements 304020 and 304030 are entry movements belonging to Link 3040. The link's queue-building rate per cycle is determined by the difference between its demand per cycle (15 vehicles) and the sum throughput of its movements (Equation 22, which is according to Equation 6).

$$15 - (O_{304020} + O_{304030}) = q_{3040} \quad (22)$$

Equations 23 and 24 define the 0–1 variable X_{304030} and are based on the general formulation presented in Equations 7 and 8. This variable distinguishes between the case in which the throughput of Movement 304030 equals its demand ($X_{304030} = 1$) and the case in which it equals its capacity. Note that M was set at 1,000.

$$0.5 \cdot G_{3040} + 1000 \cdot X_{304030} \geq 9 \quad (23)$$

$$0.5 \cdot G_{3040} - 1000 \cdot (1 - X_{304030}) \leq 9 \quad (24)$$

Equation 25 demonstrates the way the throughput of an entry movement (in this case Movement 304030) is determined (according to Equation 11).

$$O_{304030} = SX_{304030} + (1 - X_{304030}) \cdot 0.6 \cdot D_{3040} \quad (25)$$

Equations 26 to 28 define the variables SX_{304030} and OX_{304030} as explained by Equations 12 to 14.

$$SX_{304030} - 1000 \cdot X_{304030} \leq 0 \quad (26)$$

$$SX_{304030} - 0.5 \cdot G_{3040} \leq 0 \quad (27)$$

$$0.5 \cdot G_{3040} - SX_{304030} \leq 1000 \cdot (1 - X_{304030}) \quad (28)$$

For every phase of Intersection 30, a minimum green duration should be obtained (see Equation 15); for example, the constraint for the first phase is Equation 29.

$$G_{3010} \geq 5 \quad (29)$$

The results of the MILP are given in Table 1. In order to summarize these results, link data are presented instead of movements data. These results indicate that queues can be avoided at certain entry links (such as 6020); at others, vehicles are bound to accumulate rapidly (such as at Link 4040). All-red duration was prolonged at Intersection 40 beyond the required minimum, thus the sum of nominal green durations is 56 sec, which is smaller than the effective green time (91 sec).

The third column shows the throughput (in vph) of every movement in the network. Special attention should be given to entry Links 2020, 3040, 4040, 5020, and 5030. Their throughput is lower than the demand rate (540 vph), which reflects the fact that metering is active at those links. In stage three, TRANSYT is used with the output of the MILP model.

Table 2 gives a summary of the main results produced by TRANSYT. It should be noted that queue storage capacity is not defined for entry links as these links are assumed to be very long. As was expected, the maximum queue length of each link, within the cycle, as given by TRANSYT, is longer

TABLE 1 Results of First Iteration of MILP Model

Link No.	Direction	Entry Link	Queue Building up rate (per cycle)	Green duration	Throughput
1010	NB	N	0	28	504
1020	WB	N	0	12	216
1030	SB	N	0	28	504
1040	EB	N	0	20	360
2010	NB	Y	0	86	540
2020	WB	Y	10.83	8	150
3020	WB	N	0	21	384
3030	SB	N	0	6	461
3040	EB	Y	2.32	25	456
4010	NB	N	0	28	384
4030	SB	Y	0	30	540
4040	EB	Y	12.5	5	456
5010	NB	N	0	23	418
5020	WB	Y	12.5	5	90
5030	SB	Y	10.64	8	157
5040	EB	N	0	51	418
6020	WB	Y	0	30	540
6030	SB	N	0	7	120
6040	EB	N	0	54	504

than the queue length at the end of the cycle, as calculated by the MILP model. Special attention should be given to internal links where link capacity tends to be low and overflow must be prevented.

If no queue exceeds the link length, the process ends at this point, otherwise Step 4 of the algorithm is applied. In this case, the equations representing the queue in the MILP model should be modified. The difference between the volume entering the link and its capacity is increased. In mathematical terms, Equation 5 turns into Equation 30, in which Diff is a constant defined by the designer.

$$S_{ilm} \cdot G_{ij} - P_{ilm} \cdot \sum_{rsk \in \Omega_{il}} O_{rsk} \geq \text{Diff} \qquad (30)$$

In the example described in this section, spillbacks were detected on Links 1010 (which belongs to the CI) and 5010. For Link 1010, two options are available. The first option is to change the signal program for the CI and to solve this problem from the beginning. The second is to change the right-hand side of the appropriate constraint (Equation 30). The throughput of Link 1010 was reduced from 504 vph to 486 vph, but its green duration remained the same. The difference between the volume of Link 5010 per cycle and its capacity was set at greater than or equal to 1. Table 3 provides a summary of the results of the second iteration of the MILP model and the results of TRANSYT. The changes can be demonstrated for Link 1010. In order to reduce the volume of this link, the green duration of approach 3040 was decreased from 25 to

TABLE 2 Maximum Queue Lengths Resulting from TRANSYT in First Iteration

Link No.	Direction	Entry Link	Saturation degree	Queue storage capacity	Max. queue length
1010	NB	N	1	10	11
1020	WB	N	1	10	3
1030	SB	N	1	14	12
1040	EB	N	1	11	7
2010	NB	Y	0.35	irrelevant	3
2020	WB	Y	3.75	irrelevant	29
3020	WB	N	1	12	10
3030	SB	N	0.99	12	10
3040	EB	Y	1.2	irrelevant	27
4010	NB	N	1	13	11
4030	SB	Y	1	irrelevant	15
4040	EB	Y	6	irrelevant	30
5010	NB	N	0.97	9	10
5020	WB	Y	6	irrelevant	30
5030	SB	Y	3.33	irrelevant	30
5040	EB	N	0.46	9	4
6020	WB	Y	1	irrelevant	15
6030	SB	N	0.95	9	2
6040	EB	N	0.52	11	6

TABLE 3 Results of Second Iteration of MILP Model and TRANSYT

Link No.	Direction	Entry Link	Queue Building-up Rate (per cycle) (MILP)	Green duration (MILP)	Throughput (MILP)	Saturation Level (TRANSYT)	Max. queue Length (TRANSYT)
1010	NB	N	0	28	468	0.93	10
1020	WB	N	0	12	216	1	3
1030	SB	N	0	28	504	1	12
1040	EB	N	0	20	360	1	8
2010	NB	Y	0	86	540	0.35	3
2020	WB	Y	10.83	8	150	3.75	28
3020	WB	N	0	43	383	0.49	4
3030	SB	N	0	26	461	0.98	10
3040	EB	Y	0	22	397	1.36	28
4010	NB	N	0	27	482	0.99	10
4030	SB	Y	0	30	540	1	15
4040	EB	Y	12.5	5	90	6	28
5010	NB	N	0	24	403	0.93	9
5020	WB	Y	10.24	5	90	6	30
5030	SB	Y	12.5	9	162	3.33	30
5040	EB	N	0	50	410	0.46	3
6020	WB	Y	0	30	540	1	15
6030	SB	N	0	7	118	0.94	2
6040	EB	N	0	54	468	0.48	9

22 sec; the flow from this direction thus changed from 456 vph to 397 vph.

Results of TRANSYT after the second iteration (Table 3) show that queues at Links 1010 and 5010 were actually shortened; however, these are not the only changes. The new splits and offsets affected other queues at entry links, and thus all queues should be checked again. Stages 2 to 4 of the algorithm can be repeated until a satisfactory solution is obtained.

CONCLUDING REMARKS

The contribution to traffic management in congested urban networks of the method presented in this work compared with other metering strategies lies in its integration of a mathematical tool and a simulation tool. This integration achieves a better reflection of reality in the planning process and thus ensures more reliable results. Reliability is further increased by the possibility of exploiting the capabilities of each tool without violating the basic axioms on which it was grounded. The proposed method follows the principle of adjusting the capacity of the entry links to the one of the CI through metering at the entrances of the network. Thus, queues are accumulated in predetermined approaches and QLM is achieved, meaning that the length and location of queues are controlled by the planner. Because of differences in the geometry of the various intersections (number of lanes, parking arrangements, etc.), not all movements are oversaturated, even in congested conditions. Therefore, the design procedure explicitly deals with congested conditions when and where they exist; it does not, however, consider oversaturation on all links as a preliminary assumption [in contrast to other works (*3,5*)].

As a part of the procedure advanced here, an MILP model was formulated through which the green durations and volumes of all movements are determined; it ensures that no queues remain in internal links at the end of each cycle. This model was combined with TRANSYT in a way that uses the advantages of the latter while overcoming its inefficiencies. The MILP model creates undersaturated conditions within the network, which are the only conditions suitable for TRANSYT. TRANSYT is then used for offset optimization in calculating the optimal offsets and maximum queue lengths during the cycle. The two planning tools, the mathematical model and TRANSYT, complement each other to achieve signal-timing programs that are designed explicitly for the congested conditions that exist in the network.

Despite the advantages of the proposed approach presented in this paper, some limitations of the algorithms suggested should be addressed. The main deficiency of the method stems from the assumption that the route choice within the network is constant and is known in advance. Further research should concentrate on integrating the two existing modules of the process presented with an assignment module. Another disadvantage of this method, which unfortunately is common to most planning methods used today, is the separation of the split and the offset optimization processes. This separation, though meant to simplify the design algorithm, might damage the quality of the solution obtained.

The third limitation stems from the usage of TRANSYT as the simulation tool. TRANSYT performs a deterministic simulation, and thus does not reflect the stochastic nature of reality. Substituting TRANSYT with a different simulation tool that can continue serving the functions of optimizing offsets and estimating expected queue lengths, while explicitly considering stochastic phenomena, should be considered in the future.

It is hoped that implementation of the method described in this paper in a real network will be tested soon. The conclusions of such an experiment will serve as guidelines for improving the method.

REFERENCES

1. R. W. McShane and R. P. Roess. *Traffic Engineering*. Prentice-Hall, Englewood Cliffs, N.J., 1990, Chapter 30.
2. D. C. Gazis. Optimum Control of a System of Oversaturated Intersections. *Operations Research*, No. 12, 1964, pp. 815–831.
3. P. G. Michalopoulos and G. Stephanopoulos. Optimal Control of Oversaturated Intersections: Theoretical and Practical Considerations. *Traffic Engineering and Control*, Vol. 19, No. 5, 1978, pp. 216–222.
4. P. G. Michalopoulos and G. Stephanopoulos. An Algorithm for Real-Time Control of Critical Intersections. *Traffic Engineering and Control*, Vol. 20, No. 1, 1979, pp. 9–15.
5. M. G. Singh and H. Tamura. Modelling and Hierarchical Optimization for Oversaturated Urban Road Traffic Network. *International Journal of Control*, Vol. 20, No. 6, 1974, pp. 913–934.
6. E. B. Lieberman. Concepts of Control for Oversaturated Networks. Presented at TRB Traffic Signal Systems Committee Meeting, Minneapolis, Minn., 1990.
7. H. R. Kirby. Transport Research at the University of Leeds: Report for 1991. *Traffic Engineering and Control*, Vol. 33, No. 3, 1992, pp. 171–178.
8. A. J. Lindley and D. G. Cappele. Freeway Surveillance and Control. *Traffic Engineering Handbook*, 4th ed. (J. L. Pline, ed.). Prentice-Hall, Englewood Cliffs, N.J., 1992.
9. D. I. Robertson. *TRANSYT: A Traffic Network Study Tool*. Laboratory Report 253. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1969.
10. M. J. Moskaluk and P. S. Parsonson. Arterial Priority Option for the TRANSYT-7F Traffic-Signal-Timing Program. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1988, pp. 57–60.
11. D. Mahalel and Y. Gur. *Optimal Operation of a Critical Intersection* (in Hebrew). Research Report. Transportation Research Institute, Technion, Haifa, Israel, 1991.

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

Graphical Comparison of Predictions for Speed Given by Catastrophe Theory and Some Classic Models

JORGE A. ACHA-DAZA AND FRED L. HALL

Using a common data set, the speed estimates given by catastrophe theory are compared with those of some classic models. After a new proposed set of transformations to real traffic flow data, which include translation and rotation of axes, catastrophe theory is applied to information from the Queen Elizabeth Way in Ontario, Canada. The catastrophe theory model is used for predicting speeds on the basis of information on volume and occupancy. Several proposed models and a double-linear-regime model are also used for predicting speeds. The different estimates for speed are graphically compared with the observed values. A coefficient of determination (R^2) is given as the measure of reliability for all the models. The results show that the catastrophe theory model performs better than the other models.

Recent works on catastrophe theory applied to traffic flow by Navin (*1*) and Hall and others (*2–4*) have shown that traffic flow data exhibit the necessary properties for applying the cusp catastrophe surface. They have also shown that before making that application a transformation in the data is needed. Such a transformation should include a translation and rotation of axes. Developing those ideas, a way of doing those transformations is suggested by this work. Also presented in this paper are the results of the application of catastrophe theory to the transformed data. Using the same data set, other classic models are applied. Expressions for speed against flow given by Greenshields (*5*), Greenberg (*6*), and Edie (*7*), and a double-linear-regime model are also used for predicting speeds. The predictions for speed given by all the models are compared with the observed values. This comparison is made graphically and using a coefficient of determination (R^2).

The paper is organized into five sections. In the first, the data set used and the manipulation for applying the different models are described. The second section presents the elements of catastrophe theory used in this work. The way of doing the proposed translation and rotation of axes are described. In the third, catastrophe theory is applied to the data. In the fourth, the other models are applied to the data. Finally, the predictions for speed given by catastrophe theory are compared one to one against those given by the other models.

DATA

The original data set used in this work is from Station 12 median lane (left-most lane or Lane 1) of the Freeway Traffic Management System (FTMS) for the eastbound Queen Elizabeth Way (QEW) near Toronto, Canada. It consists of 2,936 thirty-sec intervals from 5 consecutive working days in the week from August 20 to 25, 1990. The location of Station 12, 2 km upstream of a recurrent morning bottleneck, and the time of the collection, from 5:30 to 10:30 a.m., made it possible to include information on both uncongested and congested regimes of traffic flow. The QEW FTMS uses paired induction loops and therefore can provide measurements of volume, occupancy, and speed.

Because regression analysis assumes an equal representation of the different zones of the data points, a random sample was drawn. This sample was taken following the procedure suggested by Drake et al. (*8*). The data were divided into intervals of 3 percent of the occupancy. Sixty points were taken from each of the 0 to 2 percent, 3 to 5 percent, 6 to 8 percent, 9 to 11 percent, 12 to 14 percent, 15 to 17 percent, and 18 to 20 percent ranges of occupancy. All 184 of the points with occupancies greater than or equal to 21 percent were also considered. These 184 points include 106 with occupancies from 21 to 30 percent, 42 from 31 to 39 percent, and 36 with occupancies greater than or equal to 40 percent. In this form, the total number of points included in the sample is 604.

CATASTROPHE THEORY

Catastrophe theory is a mathematical model that uses some special functions to represent some practical problems. The theory, created by Thom in 1972 (*9*), reproduces problems where one of the variables exhibits sudden changes and the others present smooth changes, and smooth changes might be expected in the first variable. Navin (*1*), Hall and others (*2–4*) have shown that catastrophe theory, and specifically the cusp catastrophe surface, can be applied to traffic flow, but before making that application the data have to be transformed using an axes translation and rotation.

The cusp catastrophe surface relates three variables, two known as control variables and one as the state variable. Its general expression, as given by Zeeman (*10*), is one that minimizes the potential function

$$W(X) = aX^4 + bUX^2 + cVX \quad (1)$$

with critical surface defined by

$$4aX^3 + 2bUX + cV = 0 \quad (2)$$

Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4L7, Canada.

where

U, V = control variables,
X = state variable, and
a, b, c = parameters of equation.

Figure 1 presents the shape of a full cusp catastrophe surface. Points in the center part of the folded area for Figure 1 represent maxima of the potential function (Equation 1). Because the potential function is to be minimized, these points are not of interest and the center part of the fold can be removed, giving rise to what is known as the perfect delay convention.

However, there are problems where only one value occurs for the state variable in the folded area. For this case, the Maxwell convention (Figure 2) is applied. The upper and lower surfaces of the cusp catastrophe are separated by a vertical cut, and the changes from one to the other surface occur instantaneously. The potential function (Equation 1) is always at a minimum.

APPLICATIONS TO TRAFFIC FLOW

Navin (*1*) was the first researcher to apply the cusp catastrophe surface to traffic operations. Following the double-regime approach proposed by Edie (*7*), he graphically established that the orthogonal projections of a cusp catastrophe can fit the speed-volume, speed-density, and volume-density curves given by Edie's expressions. Navin selected speed as the control variable because of the sudden changes that it presents when changing from noncongested to congested regimes or from congested to noncongested. Density and volume were selected as the control variables, and the origin was related to zero operations. Navin however did not apply any real data to his formulations.

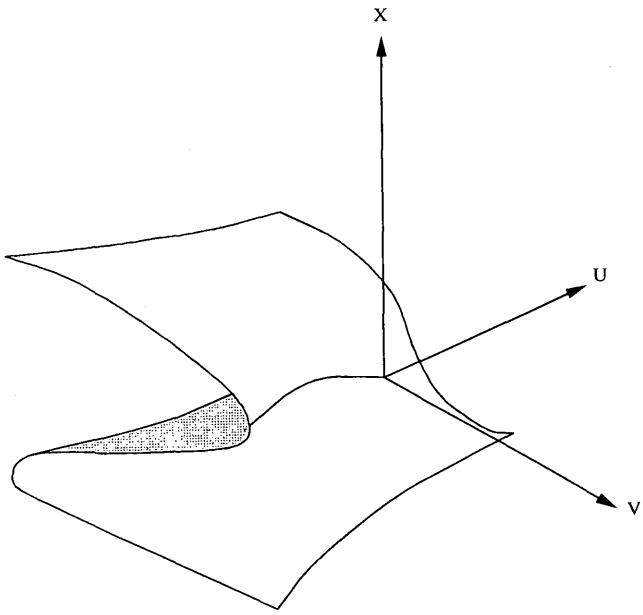


FIGURE 1 Cusp catastrophe.

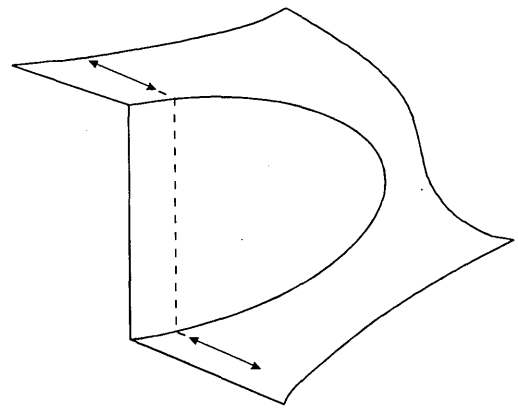


FIGURE 2 Maxwell's convention.

Hall (*2*), using real data from the QEW FTMS, drew representative curves for speed-volume, speed-occupancy, and volume-occupancy and showed graphically, like Navin, that they fit the orthogonal projections of a cusp catastrophe surface. Hall follows the same variable assignment as that of Navin. However, Hall proposed a different orientation for the axes, with the origin being associated with capacity operations. The use of representative curves meant Hall's formulation still needed to be fitted to real data. This was done in a paper by Dillon and Hall (*3*) that used a variables transformation that was in essence an axes translation. The results were not very satisfactory.

Using a visual inspection of real data sets, Forbes and Hall (*4*) compared the properties of the data with those of a cusp catastrophe surface. They found that by using the Maxwell convention, the data exhibited the necessary properties for applying catastrophe theory, but the data still needed an axes rotation before applying this convention because the orientation of the traditional axes for the volume-occupancy plane does not coincide with the position of the Maxwell convention. However, they did not carry through the calculations. Acha-Daza (*11*) followed the ideas of the previous work and applied catastrophe theory to real traffic data. Using about 50 different data sets from five stations on the QEW, he showed that the selection of a new origin with its corresponding axes translation, as proposed by Dillon and Hall (*3*), and an axes rotation for applying the Maxwell convention, as proposed by Forbes and Hall (*4*), gives results that, for some cases, may be considered to be excellent.

APPLICATION OF CATASTROPHE THEORY

The variable assignment for this work follows that of Navin (*1*) and Hall (*2*). Volume (or flow) is related to the control variable U, occupancy to the control variable V, and speed to the state variable X. The position of the origin in Figure 1 is associated for this work, as in that by Hall (*2*), with capacity operations. The origin is selected as the point of maximum volume, maximum occupancy at maximum volume, and minimum speed for maximum volume at maximum occupancy. This leads one to the axes translation given by

$$U_1 = \text{volume} - \text{maximum volume} \quad (3)$$

V_1 = occupancy − maximum occupancy at maximum volume (4)

X_1 = speed − minimum speed at maximum occupancy for maximum volume (5)

It was observed in the data used for this work that in some cases speed presented more than one value for different combinations of occupancy and volume. However, after a careful inspection, it was concluded that these different values were due partly to random fluctuations and partly to the precision used in the FTMS because the information for occupancy is recorded as truncated percentages. Hence the Maxwell convention can be applied.

For this analysis, the axes rotation needed in order to apply the Maxwell convention is done using the expressions

$$U = U_1 \cos\theta - V_1\, Gf \sin\theta \tag{6}$$

$$V = U_1 \sin\theta + V_1\, Gf \cos\theta \tag{7}$$

where Gf is a graphical factor used for solving the problem of scaling in the graphs. The use of the variables without this transformation gave distorted figures for the volume-occupancy plots, and so this parameter was introduced. Gf is given by Gf = maximum volume/maximum occupancy. The value of θ is found iteratively as the angle that minimizes the number of misclassified points, that is, those that are in the congested zone of the traffic flow but are classified to be in the uncongested, and vice versa. The limit between the zones is given by the minimum speed at maximum occupancy for maximum volume.

Once the data have been transformed, an expression of the form

$$X^3 + a_1 UX + b_1 V = 0 \tag{8}$$

is applied. In Equation 8, a_1 and b_1 are parameters that will have to be found by calibration and that are equivalent to

$$a_1 = 2b/4a \tag{9}$$

$$b_1 = c/4a \tag{10}$$

where a, b, and c are the parameters of Equation 1.

Expression 8 is used for the determination of the value of X. From this value for X, the predicted speed is found as S_p = X + minimum speed at maximum occupancy for maximum volume.

The problem of the calibration of Expression 8 was solved using a direct search of the values of a_1 and b_1 that minimize the square of the difference between the observed and predicted values for speed following the Hooke and Jeeves pattern search strategy for two variables (*12*). This direct search was conducted because of the nonlinear or intrinsically linear nature of Expression 8 that does not allow one to use a traditional linear regression analysis.

Looking for a common basis of comparison among the solutions given by catastrophe theory and those of the other models, the plots of observed speed against predicted speed and a coefficient of determination, R^2, between the observed and predicted values for speed were used. The formula for R^2 was used as given by Theil (*13*,p.83).

$$R^2 = \frac{\left[\sum_{i=1}^{n} (S_{io} - \bar{S}_o)(S_{ip} - \bar{S}_p)\right]^2}{\sum_{i=1}^{n} (S_{io} - \bar{S}_o)^2 \sum_{i=1}^{n} (S_{ip} - \bar{S}_p)^2} \tag{11}$$

where

S_{io} = observed values of speed,
S_{ip} = predicted values of speed, and
$\bar{S}_o$, $\bar{S}_p$ = arithmetic means of S_{io}s and S_{ip}s.

In the different figures showing the models, R^2 is shown as R2, since the graphics program could not produce the exponents.

Using the equations just described, catastrophe theory was applied to the reduced data set. The values of the parameters for the variable transformation and for the catastrophe expression used were as follows: maximum volume = 28 veh/30 sec, maximum occupancy at maximum volume = 15 percent, minimum speed at maximum occupancy for maximum volume = 95 km/hr, maximum occupancy = 81 percent, angle of rotation = −7.8 degrees. The application of the calibration process gave value for a_1 = −120 and b_1 = 44,413. The coefficient of determination (R^2) using these values of the parameters is equal to 0.92. The graph of observed against predicted values for speed is shown in Figure 3.

OTHER MODELS CONSIDERED

Because the rest of the models are intrinsically linear or linear in parameters, a least-squares linear regression method was used for the determination of their parameters. The expressions used for the calibration and the values of the parameters are presented in the following.

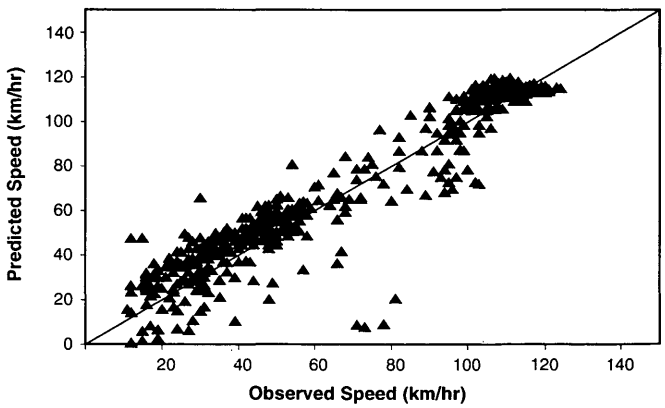


FIGURE 3 Speeds estimated using catastrophe theory plotted against observed speeds.

Greenshields' Model

For the application of the Greenshields' model the expression used was

$$Q = K_j V - (K_j/V_f)V^2 \quad (12)$$

where

Q = flow (veh/hr);
K_j = a constant, jam density (veh/km);
V = speed (km/hr); and
V_f = a constant, free-flow speed (km/hr).

The calibration for the data set considered gave the values of parameters

$$Q = 60.9V - 0.432V^2 \quad (13)$$

The predicted values for speed were calculated using the observed values for the flow (vehicles per hour) and solving the quadratic Equation 13 for V. The calculation of the coefficient of determination (R^2) gave a value of 0.87. The plot of the observed values of speed against those predicted by Greenshields' model is presented in Figure 4. It is important to note that the value of the R^2 corresponds to a reduced sample because values of flow over 2,160 veh/hr make the values of V in Expression 13 imaginary.

Greenberg's Model

Greenberg's model was applied using the expression

$$\ln(Q/V) = \ln(K_j) - V/c \quad (14)$$

where c is the speed at capacity (km/hr). After the determination of the parameters, Expression 14 becomes

$$\ln(Q/V) = 4.41 - 0.0177V \quad (15)$$

The predicted values for speed were calculated using the observed values for the flow and solving for V in Equation 15. The value for the R^2 in this model was 0.611. The plot of the observed values of speed against the calculated speeds is presented in Figure 5.

Edie's Model

Before applying Edie's model and also the two-linear-regime model, the data set was divided in two subsets, one for the uncongested data and the other for the congested points. The critical value used for this division was the minimum speed at maximum occupancy for maximum volume (95 km/hr). The same expression used for Greenberg's model was used for the congested data. For the uncongested data the expression used was

$$Q = K_c \ln(V_f)V - K_c V \ln(V) \quad (16)$$

where K_c is the density at capacity in vehicles per kilometer, a constant. The application of the least-squares regression led to the expressions

$$\ln(Q/V) = 4.11 - 0.0108V \quad (17)$$

for the congested zone, and

$$Q = 515V - 107V \ln(V) \quad (18)$$

for the uncongested zone.

The observed values of flow were used in Equations 17 and 18 to determine the predicted speeds. These predicted speeds were plotted against their observed values (Figure 6). The combined R^2 for both zones is equal to 0.80.

Double-Linear-Regime Model

Greenshields' expression was used twice for this model, one for the uncongested zone and the other for the congested one. For the uncongested zone the expression found is

$$Q = 121V - 0.983V^2 \quad (19)$$

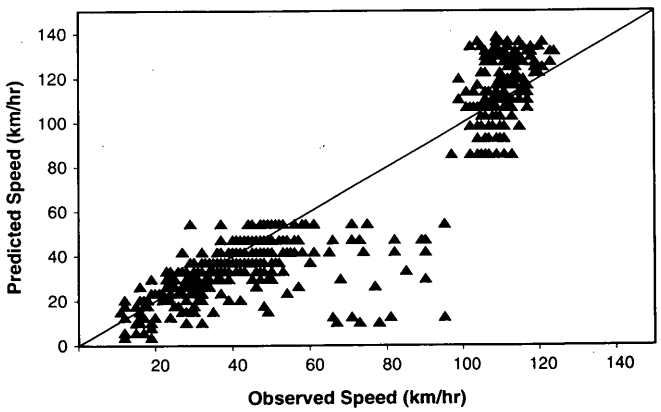


FIGURE 4 Speeds estimated using Greenshield's model plotted against observed speeds.

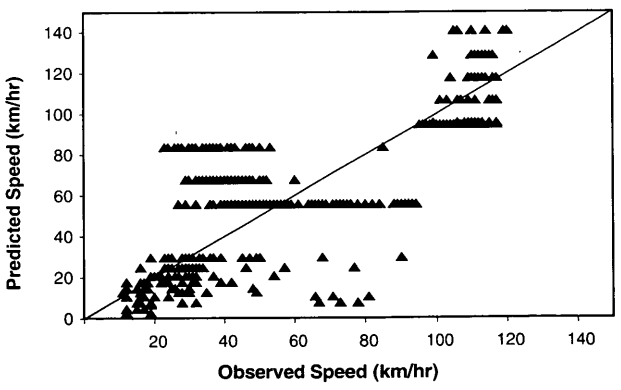


FIGURE 5 Speeds predicted using Greenberg's model plotted against observed speeds.

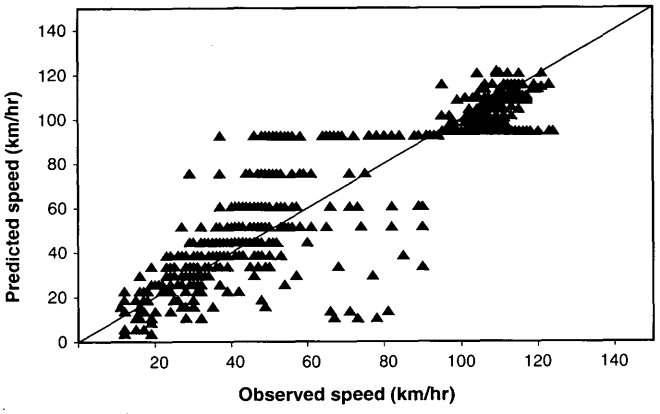


FIGURE 6 Speeds predicted using Edie's model plotted against observed speeds.

and for the congested zone the expression is

$$Q = 54.8V - 0.347V^2 \quad (20)$$

The observed flows were used in Equations 19 and 20 to calculate the predicted speeds. The observed values of speed were plotted against those calculated speeds (Figure 7). The value of the combined R^2 is 0.89. Note that the solutions for V of Equation 19 are imaginary roots for values of the volume over 2,040 veh/hr and for Equation 20 for values of the volume over 1,920 veh/hr. This reduced the sample size for this model to 570 points.

COMPARISON OF CATASTROPHE THEORY AGAINST OTHER MODELS

When observing Figures 3 to 7, several characteristics are found. First, only catastrophe theory can react to changes in concentration. For constant values of the volume, the other models predict constant speeds without considering differences in density. Second, the predictions for speed given by Greenberg's model are well beyond the limit of 150 km/hr. The solutions given by this model are the worst. Third, the double-linear-regime model does not predict intermediate values for speed. It clearly separates the data in two different zones with speeds well above or below the critical value considered (95 km/hr).

Regarding the values of the coefficient of determination (R^2), catastrophe theory gives the best value even when considering the problem of the reduced sample size for Greenshields' and the double-linear models. If only the reduced samples of those two models are considered, the R^2 values for catastrophe theory remain at 0.92 in both cases.

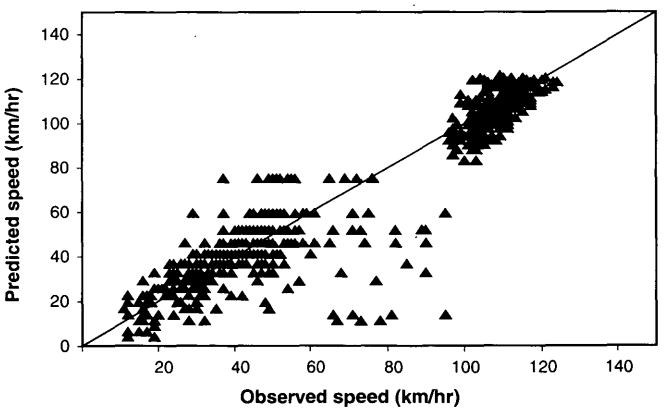


FIGURE 7 Speeds predicted using double-linear-regime model plotted against observed speeds.

CONCLUSION

Solely on the basis of the graphical comparison of Figures 3 to 7 and on the calculated values of R^2, catastrophe theory can be claimed to be the best model. The proposed transformation in the variables and axes works adequately for this case. The application of catastrophe theory to traffic flow has become a reality. It can reproduce the relationship among traffic flow variables better than the other models considered. An additional advantage of catastrophe theory is that it eliminates the necessity for the use of a double-regime model such as Edie's or the double-linear one.

The results presented can lead one to think of catastrophe theory as a model that will give excellent solutions when applied to traffic flow operations. However, this is not always the case. When applying catastrophe theory to some different data sets used in his thesis, Acha-Daza (*11*) found that the solutions given by this model were not necessarily as good as in the example presented in this paper. Data from stations affected by queue discharge flow (i.e., locations where drivers are accelerating, with consequent nonequilibrium values for speed) suggested that the application of catastrophe theory to these cases is not appropriate. The predicted values for speed were nearly constant around two values, one for the congested zone (queue discharge data) and one for the uncongested zone. At those stations, the model did not react adequately to changes in volume or occupancy.

The excellent results given by catastrophe theory can be explained only by the nature of the traffic operations in multilane highways. The correct representation of the operations on these highways is very much closer to a behavioral problem than to a physical one. Those are the kind of situations that catastrophe theory can model and for which its earliest applications were done.

The catastrophe theory model was found to be very susceptible to changes in the values of the parameters used for the axes translation (maximum volume, occupancy at maximum volume, and minimum speed at maximum occupancy for maximum volume). Changes in those parameters can make the solutions for the same data set very different. Probably it is necessary to define those parameters in a different way, trying to identify some constant values rather than defining them only from the observed data. Further work can determine if that will give better solutions.

ACKNOWLEDGMENTS

Most of the research for this paper was performed during studies for Jorge A. Acha-Daza's master's degree in engi-

neering, financial support for which through the McMaster-CIDA Scholarship program is gratefully acknowledged. The authors also wish to acknowledge the financial support of the Natural Sciences and Engineering Research Council of Canada and the assistance of the Ministry of Transportation of Ontario in making the data available.

REFERENCES

1. F. P. D. Navin. Traffic Congestion Catastrophes. *Transportation Planning and Technology*, Vol. 11, 1988, pp. 19–25.
2. F. L. Hall. An Interpretation of Speed-Flow Concentration Relationships Using Catastrophe Theory. *Transportation Research-A*, Vol. 21A, No. 3, 1987, pp. 191–201.
3. D. S. Dillon and F. L. Hall. Freeway Operations and the Cusp Catastrophe: An Empirical Analysis. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1988, pp. 66–76.
4. G. J. Forbes and F. L. Hall. The Applicability of Catastrophe Theory in Modelling Freeway Traffic Operations. *Transportation Research-A*, Vol. 24-A, No. 5, 1990, pp. 335–344.
5. B. D. Greenshields. A Study in Highway Capacity. *Proc.*, Vol. 14, 1934, pp. 448–477.
6. H. Greenberg. An Analysis of Traffic Flow. *Operations Research*, Vol. 7, 1959, pp. 78–85.
7. L. C. Edie. Car Following and Steady-State Theory for Non-Congested Traffic. *Operations Research*, Vol. 9, 1961, pp. 66–76.
8. J. S. Drake, J. L. Schofer, and A. D. May. A Statistical Analysis of Speed Density Hypothesis. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
9. R. Thom. *Structural Stability and Morphogenesis* (D. H. Fowler, translator). Benjamin, Reading, Mass., 1975.
10. E. C. Zeeman. *Catastrophe Theory, Selected Papers, 1972–1977*. Addison-Wesley, London, England, 1977.
11. J. A. Acha-Daza. *The Application of Catastrophe Theory to Traffic Flow Variables*. M. Eng. thesis. Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada, 1992.
12. A. A. Smith, E. Hinton, and R. W. Lewis. *Civil Engineering Systems Analysis and Design*. John Wiley and Sons, New York, N.Y., 1983, pp. 183–185.
13. H. Theil. *Introduction to Econometrics*. Prentice-Hall, Englewood Cliffs, N.J., 1978.

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

Lognormal Distribution for High Traffic Flows

MINJIE MEI AND A. GRAHAM R. BULLEN

The headway probability distributions for very high traffic flows are studied theoretically and empirically. Theoretical analyses show that the lognormal mechanism is applicable to individual headways for drivers in a car-following state. Thus the headway distribution of a traffic stream at high flows should follow the shifted lognormal distribution. Tests on a freeway data set with lane flow rates of 2,500 to 2,900 vehicles per day gave an excellent fit for the shifted lognormal distribution with a 0.3- or 0.4-sec shift. Previously the normal model had been the preferred headway distribution for high traffic flows, but in the present study it did not fit the data very well.

The time headway between vehicles in a traffic stream is an important and well-studied flow characteristic. Many models have been proposed for the probability distribution of headways and tested against highway traffic data. These distributions are widely used in traffic analysis methodologies and in traffic simulations.

Traffic displays distinct characteristics at different flow levels. There have been many studies of the distributions at low and medium traffic flow rates and many distributions have been calibrated. Very few studies, however, have been devoted to headway distributions at high flows for which all of the vehicles are in a car-following state. In this situation it has been most frequently assumed that the headway distribution follows a normal distribution (*1*), but the normal model does not fit the observed traffic data well and lacks a sound explanation of the traffic phenomenona.

The focus of this study is on the mechanism of time headways and their distribution at high flows. The lognormal model is proposed and is tested with field data, and it is compared with other models that have been previously presented.

LOGNORMAL MECHANISM FOR TIME HEADWAYS

The lognormal distribution has been found by many researchers to be the best simple model for headway distributions. Greenberg (*2*) and Daou (*3*) tried to find a theoretical basis for its validity as a headway model. Greenberg (*2*) and Tolle (*4*) found that the model gave good fits to the headway data that they collected. The lognormal model, however, has not been completely established for headway distributions for two reasons. First, no sound theoretical basis has been shown to relate the model to traffic behavior. Second, some statistical analyses with observed data have not been sufficiently rigorous.

The lognormal distribution arises as the result of a multiplicative mechanism acting on a number of factors. For this study, a unique case is of special interest, that is, the law of proportionate effect. This law deals with a variable whose value varies in a step-by-step sequence, such as in a time frame. Suppose that a variable is initially X_0, and after the jth time step it is X_j, reaching its final value X_n after n time steps. At the jth time step the change in the variable is a random proportion of the momentary value X_{j-1} already attained, thus

$$X_j - X_{j-1} = \varepsilon_j * X_{j-1} \tag{1}$$

where the set $\{\varepsilon_j\}$ is mutually independent and also independent of the set $\{X_j\}$. The law of proportionate effect then is,

> A variable subject to a process of change is said to obey the law of proportionate effect if the change in the variable at any step of the process is a random proportion of the previous value of the variable. (*5*)

The importance of the law is its link with the central limit theorem. Expression 1 may be rewritten as

$$\frac{X_j - X_{j-1}}{X_{j-1}} = \varepsilon_j \tag{2}$$

so

$$\sum_{j=1}^{n} \frac{X_j - X_{j-1}}{X_{j-1}} = \sum_{j=1}^{n} \varepsilon_j \tag{3}$$

Supposing the effect at each step to be small, then

$$\sum_{j=1}^{n} \frac{X_j - X_{j-1}}{X_{j-1}} \sim \int_{X_0}^{X_n} \frac{dX}{X} = \ln X_n - \ln X_0$$

giving

$$\ln X_n = \ln X_0 + \varepsilon_1 + \varepsilon_2 + \ldots + \varepsilon_n \tag{4}$$

The central limit theorem may be stated as, "Under very general conditions, as the number of variables in the sum becomes large, the distribution of the sum of random variables will approach the normal distribution" (*6*). The theorem is valid only when each individual variable has a small effect on

M. Mei, New York Department of Transportation, Region 8, Four Burnett Boulevard, Poughkeepsie, N.Y. 12602. A. G. R. Bullen, Department of Civil Engineering, University of Pittsburgh, Pittsburgh, Pa. 15261.

the sum. In Expression 4 the initial value X_0 will be close to 1.0 so ln X_0 will be of the order of the epsilons. The preceding process, therefore, meets all the requirements of the central limit theorem, so ln X_n is normally distributed. The random variable X_n is then lognormally distributed. Therefore, a variable obeying the law of proportionate effect is lognormally distributed provided that the change in each step is small.

Lognormal Distribution of Individual Time Headway

Assume that a driver is in a car-following situation on one lane of a freeway, and the driver either does not attempt to overtake the leading vehicle or does not have the chance to do so.

The headway of the vehicle, denoted by H, is always changing with time because drivers cannot maintain an absolute constant spacing and must adjust their speeds to maintain a safe distance and to keep up with traffic. Therefore H is a random variable changing with time. By specifying a point of time to be the starting point, and the headway at that point as H_0, the headway value after a small interval of time will change a random portion of its original value to become H_1

$$H_1 - H_0 = H_0 * \varepsilon_1 \tag{5}$$

After j time intervals, the change of headway value in the jth interval can be expressed as

$$H_j - H_{j-1} = H_{j-1} * \varepsilon_j \tag{6}$$

where

H_j = time headway at end of jth interval,
H_{j-1} = time headway at end of $j - 1$ interval,
ε_j = random proportion of change of H_{j-1} in jth interval.

With no knowledge about the distribution of ε_j the headway H will be lognormally distributed as long as the ε_j's are small and the initial headway is close to 1.0. The individual headway in a car-following situation, therefore, is lognormally distributed.

Shifted Lognormal Headway Distribution

Headway is the time interval between the arrivals of the corresponding point in a pair of moving vehicles, such as from front bumper to front bumper. The minimum value for a headway is the physical occupancy time of the leading vehicle plus a safety buffer. Then the domain of definition of headways is (d,∞). The random variable following the shifted lognormal distribution will be H', which is the headway, H, subtracted by the minimum spacing, d. The shift, d, can be determined from field data.

The mean, m, and standard deviation, σ, of headways from the observed data are given by

$$m = \frac{1}{n}\sum_{i=1}^{n} H_i \tag{7}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (H_i - m)^2} \tag{8}$$

TABLE 1 Observed Headway Data and Their Parameters

Data Set No.	1	2	3	4	5	6	7	8
Lane No.	2	2	2	2	1	1	1	1
Average Volume(vpm)	46	49	39	42	35	36	35	34
No. of Observations	230	243	155	209	173	178	174	170
Mean (seconds)	1.27	1.24	1.54	1.43	1.70	1.67	1.71	1.75
Standard Deviation	0.60	0.52	0.59	0.63	0.88	0.65	0.85	0.94
Interval (Seconds)	Measured Distribution				Measured Distribution			
0.0 - 0.25	0	0	0	0	0	0	0	0
0.25 - 0.50	1	0	0	0	0	0	1	0
0.50 - 0.75	33	21	6	17	5	2	7	8
0.75 - 1.00	56	72	22	41	21	21	16	20
1.00 - 1.25	43	58	32	39	35	29	27	30
1.25 - 1.50	46	40	26	39	33	35	32	31
1.50 - 1.75	16	21	21	23	21	25	29	20
1.75 - 2.00	10	9	17	17	19	20	23	15
2.00 - 2.25	6	8	11	12	8	17	13	12
2.25 - 2.50	9	5	8	11	4	10	6	7
2.50 - 2.75	2	2	5	2	5	6	4	7
2.75 - 3.00	4	4	4	2	5	6	3	1
3.00 - 3.25	2	0	0	3	7	2	0	4
3.25 - 3.50	0	2	2	1	2	2	2	3
3.50 - 3.75	0	1	1	1	2	1	3	2
3.75 - 4.00	0	0	0	1	0	1	4	2
4.00 - 4.25	2	0	0	0	1	1	1	3
4.25 - 4.50	0	0	0	0	2	0	1	3
4.50 - 4.75	0	0	0	0	2	0	1	1
>4.75	0	0	0	0	1	0	1	1

Denoting the mean and standard deviation of H' as m' and σ' respectively, their values are given by

$$m' = m = d \tag{9}$$

$$\sigma' = \sigma \tag{10}$$

Lognormal Distribution and Time Headway Distribution of a Traffic Stream

To analyze the time headway distribution of a traffic stream, some knowledge about the drivers who make up the traffic stream is needed. Individual drivers will have their own unique desired headways and driving habits. Thus each driver will have a time headway distribution with a unique mean and standard deviation that may vary under different traffic conditions. The traffic stream is made up of individual time headways of different drivers with these different distributions. In a car-following condition, the measured time headway values from a traffic stream are the momentary values of each of the individual lognormal distributions.

The measured time headway distribution of a traffic stream is therefore the combination of the individual time headway distributions. The combined distribution may no longer be lognormal, and its characteristic is rather complicated. Consider the extreme case when the traffic volume is very high and all drivers have to drive at the car following headway. The mean of each individual time headway may still be different but they will converge toward a single value because of the close spacing. If, at very high volumes, the differences among the individual time headway distributions are very small and approach a single distribution, then the measured time headway distribution should converge to the lognormal or to the shifted lognormal distribution.

TESTING LOGNORMAL DISTRIBUTION WITH OBSERVED DATA

Data Collection and Reduction

Time headway data was taken on the two southbound lanes of the four-lane freeway I-279 at the Milroy overpass near

TABLE 2 Results of χ^2 Test of Shifted Lognormal and Normal Distribution with Observed Data

Data Set	#1	#2	#3	#4	#5	#6	#7	#8	
Test Level	Critical Values								
10%	12.0	13.5	12.0	12.0	14.6	12.0	12.0	12.0	
5%	14.2	15.5	14.2	14.2	16.6	14.2	14.2	14.2	
1%	18.2	20.0	18.2	18.2	21.5	18.2	18.2	18.2	
	Quantitative Test Results								
									Accept at 10%
S=0.0	17.67	19.35	1.90	4.61	18.70	1.97	10.72	10.72	12.00
S=0.1	15.81	15.93	2.08	3.52	16.82	2.74	9.87	8.86	12.00
S=0.2	13.21	11.22	2.54	3.71	14.20	2.70	8.61	7.59	12.00
S=0.3	11.68	6.74	4.30	5.31	12.37	3.52	9.54	6.52	12.00
S=0.4	11.87	3.42	4.15	8.65	10.42	5.12	6.84	6.58	12.00
S=0.5	20.69	1.98	6.37	21.94	10.10	8.59	6.21	5.63	12.00
S=0.6		11.26	10.88		10.05	14.55	6.91	7.23	12.00
S=0.7		>20	19.98		12.554		10.69	12.33	12.00
S=0.8					25.559		>25		12.00
Normal	50.27	84.22	26.05	27.26	77.375	33.00	48.12	69.72	12.00
	Acceptable or Not								
S=0.0	-	-	Y	Y	-	Y	Y	Y	
S=0.1	-	-	Y	Y	-	Y	Y	Y	
S=0.2	-	Y	Y	Y	Y	Y	Y	Y	
S=0.3	Y	Y	Y	Y	Y	Y	Y	Y	
S=0.4	Y	Y	Y	Y	Y	Y	Y	Y	
S=0.5	-	Y	Y	-	Y	Y	Y	Y	
S=0.6	-	Y	Y	-	Y	-	Y	Y	
S=0.7	-	-	-	-	Y	-	Y	-	
S=0.8	-	-	-	-	-	-	-	-	
Normal	-	-	-	-	-	-	-	-	

Y ... Acceptable at All Test Levels
- ... Not Acceptable At All Test Levels
S ... Shift (Seconds)

downtown Pittsburgh during morning rush hour. There is an approximately 1 percent downgrade in the section. The average volume was 2,400 vehicles per hour per lane with 5-min flow rates reaching 2,900 vehicles per hour per lane. The average speed was about 75 km/hr, and the traffic flow was smooth and free of shock waves. There were few trucks and little lane changing in the traffic stream.

About 10,000 headways were measured, of which 1,375 were used for model testing in 8 data sets. The right lane is denoted as Lane 1, and the left lane as Lane 2. Data were grouped according to volumes. Data from the literature were used to test the model over other ranges of traffic flows.

Maximum likelihood methods were used to estimate the mean and the standard deviation of the model. The eight data sets and their parameters are given in Table 1.

Testing Models with Observed Data

In the collected data in Table 1, there are only 2 headways of less than 0.5 sec out of 1,375 measured headways. The average occupancy time is about 0.25 sec. Therefore, the shift should be in the neighborhood of 0.2 to 0.5 sec. Tests were carried out with varying shifts from 0.0 to 0.7 sec.

The most common methods for testing goodness-of-fit are the χ^2 test and Kolmogorov-Smirnov test.

Tolle did his analysis on the lognormal model mainly with Kolmogorov-Smirnov test because he found that

> The χ^2 test is not a very forgiving analysis and may be thrown off by only a few "bad" points. In reality, obtainment of actual "good" χ^2 fits from data which are influenced by so many unpredictable variables is not fully expected. (*4*,p.83)

The significance levels used were 1, 5, and 10 percent. The Kolmogorov-Smirnov tests gave acceptable results at all test levels for the eight data sets regardless of the shift. χ^2 tests with all the data sets were acceptable at all the test levels for shifts of 0.3 or 0.4 sec, but for no other shifts. The results are given in Table 2.

The normal distribution was also tested with the same data. None of the tests with the eight data sets gave acceptable results at the test levels.

Another model that has been used for headway distributions is the Pearson Type 3 distribution. With several parameters to determine the location, scale, and shape of the distribution, the model gives acceptable results for the data with χ^2 tests at the test levels. The calibration is more difficult, however, and the model also lacks any explanation of the traffic phenomenon.

CONCLUSION

Theoretical analyses have shown that the lognormal mechanism is applicable to individual headways with traffic in a car-following state. Thus the headway distribution of a traffic stream should converge to the shifted lognormal distribution as flow increases. Statistical tests on a high flow data set gave an excellent fit for the shifted lognormal distribution with a 0.3- or 0.4-sec shift. The lognormal model was superior to the normal model. Although the Pearson Type 3 model also gives good fits in most cases, it lacks a sound conceptual basis. The shifted lognormal model is an excellent and simple model for headway distributions at high flows.

ACKNOWLEDGMENT

This work is based partially on an M.S. thesis submitted by the author in partial fulfillment of the requirements for the degree of Master of Science in Civil Engineering at the University of Pittsburgh.

REFERENCES

1. A. D. May. *Traffic Flow Fundamentals*. Prentice-Hall, Englewood Cliffs, N.J., 1990, pp. 19–22.
2. I. Greenberg. The Log-Normal Distribution of Headways. *Australian Road Research*, Vol. 2, No. 7, March 1966, pp. 14–18.
3. A. Daou. On Flow Within Platoons. *Australian Road Research*, Vol. 2, No. 7, March 1966, pp. 4–13.
4. J. E. Tolle. *The Lognormal Headway Distribution Model*. Ph.D. dissertation. Ohio State University, Columbus, 1969.
5. J. Aitchison and A. C. Brown. *The Lognormal Distribution*. Cambridge University Press, Cambridge, 1963, pp. 22–23.
6. J. R. Benjamin. *Probability, Statistics and Decision for Civil Engineers*. McGraw-Hill, New York, N.Y., 1970.

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

Development and Application of a Methodology Employing Simulation To Evaluate Congestion at School Locations

LILLY ELEFTERIADOU AND ROBERT L. VECELLIO

A methodology was developed to evaluate traffic congestion at school locations during the morning hours, when parents drive their children to school. The methodology is applied at two school sites in the Auburn-Opelika area of Alabama, and involves the development of a simulation model in GPSS/PC that replicates existing traffic conditions. The model simulates the arriving of vehicles, their entering the school driveway, parking, or dropping off students, and leaving the school site. As a result of high interarrival rates and low service rates at the two school sites, queues form on the adjacent street and disrupt traffic. Following model validation, selected components were changed and their effect on queue length was determined. An evaluation of alternatives was conducted to reveal the best recommendations for alleviating the school congestion problems. It was concluded that the methodology developed is a very efficient approach, because conventional methods do not directly apply. The methodology can be used in similar situations in which queueing of vehicles occurs or when the traffic patterns are too complex to describe analytically.

Evaluation of traffic congestion at locations for which flow patterns are complex or unusual can be a very time consuming task if site-specific mathematical models are to be developed. At two school locations in the Auburn-Opelika area of Alabama (Wrights Mill Road Elementary School in Auburn, and Opelika Junior High in Opelika), congestion problems occur during the early morning and mid-afternoon hours when parents drive their children to and from school, even though school buses are available. After entering the circular driveway in front of the school, they stop, unload the children, and leave. The arrival rate, for the approximately 30 min before the classes begin, is very high, and the traffic volumes very much exceed the driveway capacity. The average interarrival time is 11.6 sec at the first school and 16.8 at the second, whereas the average service times are 122.0 and 81.9 sec, respectively. As a result, traffic queues build up on the street. Queue lengths as high as 17 vehicles have been recorded. Waiting times are also high, creating anxiety to the drivers, thus increasing the probability of erratic driving maneuvers and accidents. Because the roadways in front of the schools are two-lane and two-way, through traffic is also much affected. Consequently, there is a need to evaluate traffic movement at the schools, including the loading-unloading process.

L. Elefteriadou, Transportation Training and Research Center, Polytechnic University, Six Metrotech Center, Brooklyn, N.Y. 11201. R. L. Vecellio, Department of Civil Engineering, Harbert Engineering Center, Auburn University, Auburn, Ala. 36849.

The development of a simulation model for each site was considered to be the most appropriate approach to the problem. Computer simulation can model unusual arrival and service patterns, provide detailed information about the process, and evaluate the effect of changes in the system. Thus after formulating and validating a simulation model that replicates a real-world situation, one can experiment with new alternatives economically and study their effects using the model before recommending a final solution. A literature search revealed that even though simulation has been widely used in traffic engineering, the study approach of developing site-specific simulation models to evaluate school zones or similar facilities has not been reported in the literature.

The objectives of this study were to develop a methodology employing simulation that can be used where queueing of vehicles occurs or when traffic patterns are complex or unusual. This methodology is demonstrated at two problem locations, and after examination of traffic and operating conditions, alternative solutions are identified, and the best one to alleviate the congestion problem is recommended. The two models were developed using the simulation language General Purpose Simulation System (GPSS/PC), student version, which runs on an IBM personal computer.

SITE DESCRIPTIONS

Wrights Mill Road Elementary School is located on Wrights Mill Road, a suburban two-lane two-way highway in southeast Auburn. Figure 1 shows the street layout in the immediate vicinity of the school. The width of the school driveway, which is one-way, is 5.2 m at its narrowest point, barely accommodating two passenger cars simultaneously. At all subsequent points, the driveway is wide enough to accommodate two cars.

The main queue builds up on the northbound approach. A smaller queue is created in the southbound lane. Sometimes two vehicles enter the driveway simultaneously, creating two lanes at the entrance. These eventually merge to allow for unloading at the sidewalk in front of the school building. At the driveway exit a police officer directs traffic because vehicles exit at a rate greater than the gaps on the highway can accommodate. In addition, many children cross the road at that point.

School employees account for about a fourth of the peak-hour traffic. Two parking areas totaling 61 spaces are accessible through the driveway. School buses constitute a negli-

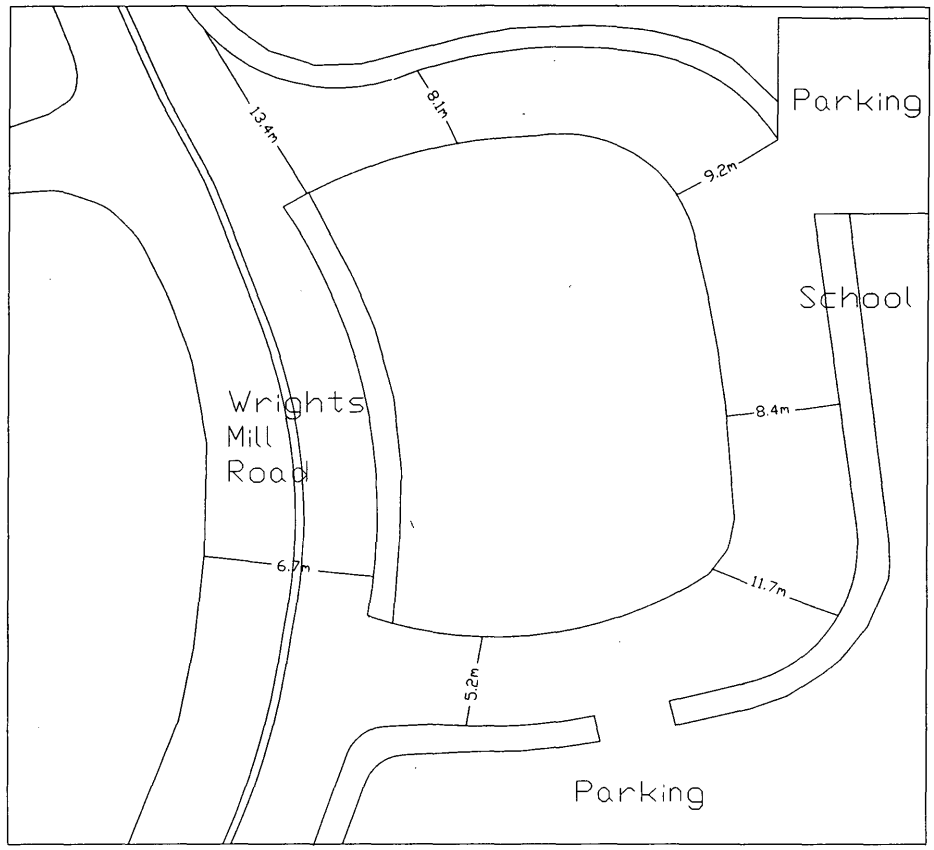


FIGURE 1 Wrights Mill Road Elementary School layout.

gible percentage of vehicular traffic, but they also must wait to enter the driveway. There is a reserved space for buses to unload in front of the school.

Opelika junior high school is located on Denson Drive at McLure Street on the north side of Opelika. Figure 2 depicts the street layout at the school. Although arrival rates during the peak hour are approximately the same for the two approaches, the major queue is formed in the northbound storage lane. Vehicles approaching the driveway entrance from the southbound lane have priority, and waiting times for them are much shorter. A small queue is occasionally created. Unloading is prohibited, except at the sidewalk in front of the school. However, many vehicles unload before reaching that point. As a result, a vehicle may block the entrance to unload while the driveway in front of it has cleared.

There are 60 parking spaces accessible from the driveway for school employees and visitors. Some vehicles pass through the parking area to unload, thus jeopardizing the safety of children who must cross the congested driveway to reach the school sidewalk.

A separate driveway is provided next to the exit of the school driveway for school buses only, which unload at the southern entrance to the school. Thus, they don't interfere with other traffic. Additional parking is also provided near the bus unloading area, which is accessible from a third entrance on Denson Drive.

DATA COLLECTION

In models of waiting line systems, the principal parameters are the interarrival and the service time distributions. From these two independent variables, other system parameters (queue length distribution, transit time distribution, etc.) can be determined. To define the two distributions, the time and license plate number of each vehicle were recorded as it joined the queue and entered and exited the school driveway. Traffic data were collected during the morning peak hour, including the buildup and dissipation of the queue. Weekday traffic operations at the school sites during the peak hours are repetitive; the same people bring their children to school at the same time throughout the school year. Therefore, data collection during one peak period was considered adequate. It is also noted that the data collected at each school site were used for the development of the model for only at that specific school.

MODEL DEVELOPMENT AND VALIDATION

Basic GPSS Modeling Concepts

In GPSS, models are built by selecting certain blocks from the available set and arranging them in a diagram, so that at

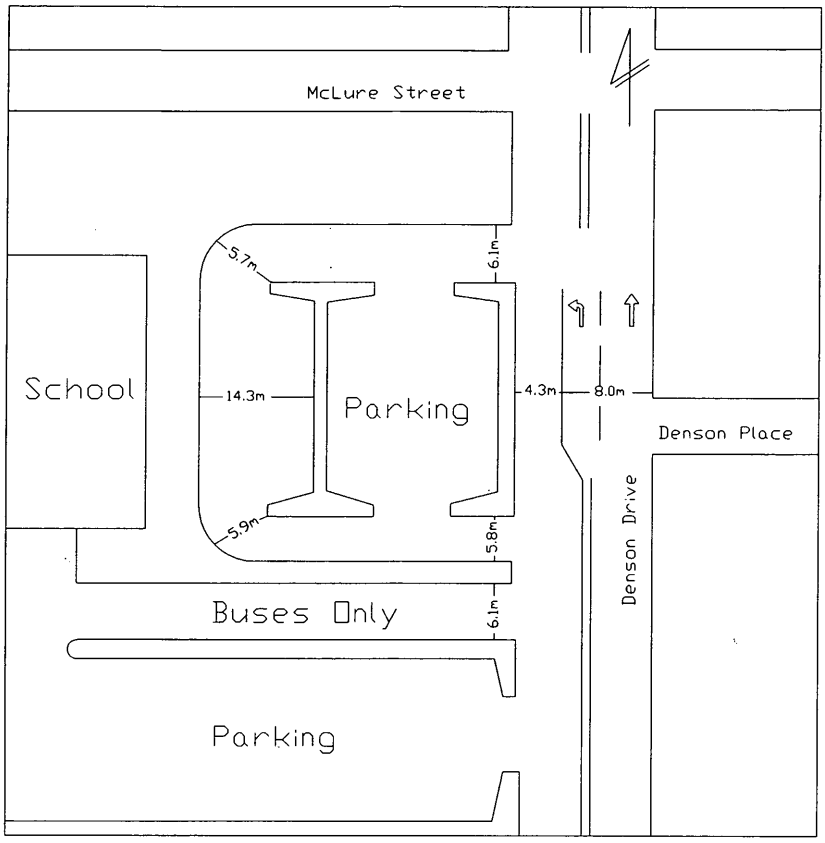


FIGURE 2 Opelika Junior High School layout.

the time of model implementation they (i.e., their images) interact meaningfully with one another (*1*). Blocks, the GPSS statements, represent actions or delays to be encountered by a set of transactions, which enter one block after another. Transactions are the units of traffic, which in the models developed for this study, represent vehicles. The GPSS processor maintains a simulated clock to provide a background for the represented events (*1,2*).

Transactions are brought into a model using the GENERATE block. The arguments of this block define the interarrival distribution. A model can contain many different GENERATE blocks. The TERMINATE block removes from the model the transactions that enter it.

Storages are used to simulate parallel servers. The ENTER and LEAVE blocks are used for the seizure and release of servers, respectively. The capacity of a storage is defined using the statement STORAGE *n*, where *n* is the number of servers. In the two school site models, a storage is used to represent the physical spaces available for service in the driveway. Each time a transaction vehicle enters the storage driveway, one server space is captured. When all spaces are captured, transactions are not allowed to enter the ENTER block. They are required to "wait" in the previous block until a space is available. Thus, waiting time characteristics can be identified. The ADVANCE block provides for the passage of time before a transaction moves to the next block. It is usually used to represent service time.

The QUEUE and DEPART blocks are used to gather statistics describing the involuntary waiting that may occur from time to time at various points in a model. When a transaction moves into a QUEUE block the event "join the queue" is simulated. Similarly, when a transaction moves into a DEPART block the event "depart the queue" is simulated. The two blocks are used as a pair in a model, thus queueing situations can be simulated.

Formulation of Models

Figure 3 shows a flow chart of the two models. For both schools, vehicles arrive and enter the driveway from two directions, which are included in the models. Vehicles enter one at a time, provided there is space available (defined as STORAGE in the programs). A percentage of them, the employees of the school, park (removed from the model). The rest drop off students and exit.

Some assumptions were necessary in building the two models to simplify the actual processes. For both models it is assumed that service time starts when a vehicle enters the school driveway and ends when the vehicle exits the driveway. It would not be feasible to distinguish between the waiting and unloading phases in the driveway, because unloading is not taking place at a specific location. The number of vehicles receiving service simultaneously represents the capacity of the

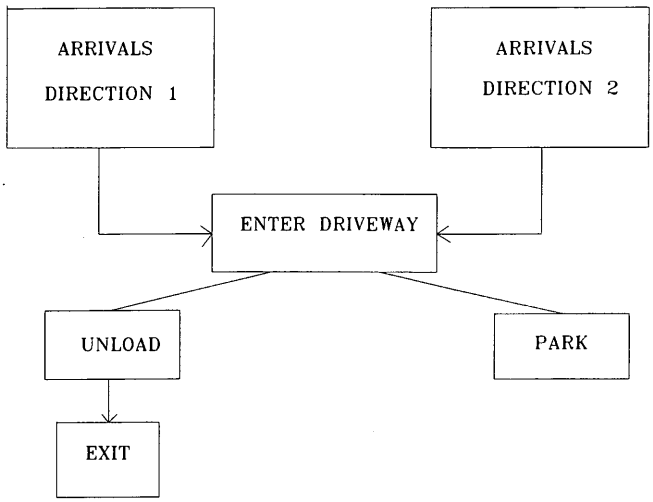


FIGURE 3 Model flow chart.

STORAGE in the models. That capacity is not a fixed number, but it varies over time depending mainly on the drivers, so that it cannot be analytically described. The number of vehicles in the driveway cannot be determined from the data, because, at any time, it is not known how many of the parking vehicles are still in the driveway. Therefore, the determination of STORAGE capacity was a trial-and-error process subject to realistic constraints.

For vehicles entering the driveway to park, it could not be determined how long they were actually occupying space in the driveway. Because they constitute a large percentage of the total number of vehicles, they could not be ignored either. Taking into account that this time is directly related to the service times of the other vehicles, it was assumed to be a fraction of service time. Depending on the location of parking spaces, a factor was determined for each school site, by which service times are divided, to determine the distribution of time spent in the driveway for parking vehicles.

As noted previously, at both school sites there were vehicles approaching from two directions to enter the driveway. Arrival times were recorded only for the direction from which the major queue formed. The minor queues were included in the models with the assumptions that arrival times were the recorded entering times and that the minor queues had priority. It was assumed that these vehicles didn't have to wait, but were entering the driveway as soon as they were arriving. That assumption does not account for the fact that a space may not be available at the time the vehicle arrives.

Random numbers are used in simulation models to create the impression that the value of the next draw from the distribution cannot be guessed (2). In GPSS the eight random number generators (RN1–RN8) each return an identical sequence of random numbers, unless their initial value is changed (using the RMULT statement).

Figure 4 presents the program simulating traffic operations during the morning peak hour at Wrights Mill Road Elementary School. Lines 10, 20, and 30 define the three empirical distributions used in the model: interarrival times for the major queue (Line 10), interarrival times for the minor queue (Line 20), and the service time distribution (Line 30). The second number of each pair is the end of an interval, and the first is the corresponding cumulative frequency. After a random number is drawn (0 to 0.99999), the cumulative frequency interval is determined, and a linear interpolation is performed between the pairs of points at the end of the interval. The intervals were established by sorting the data and arranging them into groups with equal numbers of observations. Different random number generators are used in each function to ensure that sampling is independent for the three distributions. In addition, in Statement 31, initial values are provided for the three random number generator, so that they will not use the same sequence of values.

The capacity of the driveway storage is nine vehicles. Transactions are generated over 23.5 min, or 1,410 sec, using a GENERATE-TERMINATE pair (Lines 190 and 200). Transactions enter the model through two GENERATE blocks (Lines 50 and 160). Line 50 represents the vehicles joining the major queue, and Line 160 represents vehicles arriving from the opposing lane. These two GENERATE blocks make use of the respective functions, ARV and SARV, to bring transactions-vehicles into the model. A maximum number of transactions to be generated is established to ensure that they don't exceed the actual number of vehicles in the system by more than 10 percent. As previously mentioned, transactions in the minor queue have priority in the model, because their arrival times are assumed to be their entering times.

Transactions attempt to enter the storage driveway in Line 70. If no space is available, they are required to wait at the previous block (QUEUE). At the QUEUE block, two queues are identified to distinguish between the major and minor queues, and statistics are gathered for them separately. As soon as a space is available, transactions ENTER the driveway and DEPART the queue. In Line 90, the TRANSFER block is used to remove from the model the 8.3 percent of the total number of transactions, which represents the through vehicles. Even though the through vehicles recorded were only in the northbound lane, in the program the percentage removed is of the total number of vehicles. This has a minor effect on the queue lengths and no effect on service times. It should also be noted that through vehicles in the model are waiting in the queue until a space becomes available in the driveway, whereas in the actual system they drive away as soon as they reach the beginning of the queue.

The TRANSFER block in Line 100 distinguishes between the vehicles that park (24.3 percent of the total) and those that exit the driveway. Those that part are staying in the driveway for one-third of the respective service time (Lines 35 and 110). In randomly selecting transactions to be transferred, the two TRANSFER blocks make use of the RN1 random number generator.

After staying in the ADVANCE block for their service time, transactions LEAVE the driveway (Line 140) and exit the model (Line 150). The blocks in Lines 45 and 210 to 230 are used to tabulate queue lengths every 30 simulated sec and provide the respective distribution.

The structure of the Opelika Junior High School program is similar to the program developed for the Wrights Mill Road Elementary School; therefore the program listing is not shown here.

```
 ;  GPSS/PC Program File A:WRFINAL.GPS    (V 2, #37804)   01-01-1980  01:02:17
10 ARV            FUNCTION    RN2,C14
0,0/.027,2/.062,3/.106,4/.46,5/.487,6/.549,7
.611,9/.752,10/.796,13/.85,15/.894,23/.947,39/1,115
20 SARV           FUNCTION    RN3,C9
0,0/.125,5/.25,8/.375,10/.5,60
.625,85/.75,290/.875,440/1,512
30 SER            FUNCTION    RN4,C17
0,62/.06,69/.119,84/.179,89/.238,93/.298,96
.357,108/.417,115/.488,125/.548,129/.607,133
.69,137/.75,145/.81,153/.869,161/.929,176/1,203
31          RMULT       31,859,565,117
35  PARK    FVARIABLE   FN$SER/3
40  DRV     STORAGE     9
45  INQUE   TABLE       Q5,0,1,25
50          GENERATE    FN$ARV,,,125
55          ASSIGN      1,5
60  WAIT    QUEUE       P1
70          ENTER       DRV
80          DEPART      P1
90          TRANSFER    .917,GONE
100         TRANSFER    .243,UNLD
110         ADVANCE     V$PARK
120         TRANSFER    ,GONE
130 UNLD    ADVANCE     FN$SER
140 GONE    LEAVE       DRV
150         TERMINATE
160         GENERATE    FN$SARV,,,8,1
170         ASSIGN      1,10
180         TRANSFER    ,WAIT
190         GENERATE    1410
200         TERMINATE   1
210         GENERATE    30,,16
220         TABULATE    INQUE
230         TERMINATE

Definitions
ARV:  Interarrival time distribution - Major queue
SARV: Interarrival time distribution - Minor queue
SER:  Service time distribution
PARK: Variable needed to define "service" time for parking vehicles
DRV:  Storage - Driveway
```

FIGURE 4 Wrights Mill Road Elementary School simulation model.

Validation

Validation is the exercise of verifying that the outputs of the model are reasonable, given the inputs and processing steps in the system. That is, validation assures the analyst that the model behaves just as the real system should. Comparisons of sample statistics with model predictions reveal whether differences result from chance or result from inadequacies of the model (*3*).

The property selected for validating the models is queue length. Queue length is the most visible manifestation of traffic congestion and is much easier to measure in the field than delay is to measure. At the study sites, it was the long queue length that the public complained about and, because of safety, that the school administrators worried about. Of course, queue length is positively correlated with delay. Successive observations of queue length depend on their predecessors to some degree (*3*). This dependence of consecutive observations is called autocorrelation, or serial correlation, and may cause an error in the hypothesis testing. Therefore, queue lengths every 30 sec are determined and the respective distributions are compared.

It was found that traffic patterns were stabilized at each site, so care was also taken to ensure consistency in the simulated results. Because vehicles are generated by random sampling from empirical distributions, the simulated results used for validation must reflect stable traffic patterns rather than erratic traffic patterns. To accomplish this, each simulation model was run five times, using five random sampling schemes. Thus five queue length distributions were obtained.

A two-step hypothesis test was then performed. Initially a statistical comparison was made between the most representative simulated distribution and the observed distribution. The most representative distribution was defined by examining the sum of the absolute values of the frequency differences between the observed and each simulated distribution. The most representative simulated distribution exhibited the lowest value using this measure of effectiveness. A hypothesis test was performed between the most representative distribution and the observed distribution. If this test indicated good agreement between the two distributions, then a second hypothesis test was performed between the average of the five simulated distributions and the observed distribution. If the initial hypothesis test revealed significant differences between the two distributions, there would be no reason to perform the second test.

The goodness-of-fit of the simulated queue length distribution to the observed distribution was determined using the

Kolmogorov-Smirnov (K-S) test. The K-S test was selected because it does not lose information because of grouping, as the chi-square test sometimes does. Even though the K-S test is usually used when the variable under consideration has a continuous distribution, and queue length is not a continuous variable, the error that occurs in the resulting probability is in the safe direction: if the null hypothesis is rejected by the test, one can have real confidence in that decision (*4,5*). The K-S test is also statistically more powerful for the smaller sample sizes.

The K-S test was performed by specifying and comparing the cumulative frequency distributions of the observed and the simulated queue lengths. The point at which these two distributions show the greatest divergence was determined and was used to decide whether such a divergence is likely on the basis of chance. A two-tail test procedure was used. It was found that the model data reflect the observed distributions at a 95 percent level of confidence, and it was concluded that the models can be used to evaluate alternative traffic solutions.

SELECTION AND EVALUATION OF ALTERNATIVE SOLUTIONS

The effectiveness of alternative traffic solutions was tested using the models described in the preceding. The process involves the changing of a component of the system and the evaluation of its effect on the operation of the system.

It is required that the conditions under which experiments are performed be controlled as closely as possible. Thus, for the testing of each alternative, the same sequence of random numbers was used, for both the alternative and the original models. This ensures that the observed differences result only from changing the component being tested.

For the evaluation of each alternative, five simulation runs were made and compared with the field data to determine whether they were significantly different. The statistic used to test the effectiveness of the new system is the mean queue length. A hypothesis test for means is performed to determine if the mean of the alternative model queue length is significantly less than the queue length of the existing system. This implies a one-sided test procedure. Since the sample size (n) is large ($n>30$), the test statistic used is the standard normal distribution variable for sample means. The sample mean used in the hypothesis testing is the average of the five computer runs for the alternative system being tested.

Three alternative solutions were evaluated for each school:

1. Reduce the intensity of arriving traffic. This alternative is investigated to determine the impact of a shift in the arrival pattern of traffic. By this it is meant to urge or provide incentives to parents to arrive earlier. If, for example, the school is willing to accommodate children arriving earlier, many parents might be willing to bring their children earlier. Specifically, whether uniformly distributed interarrival times would reduce the average queue length was tested. The uniform distribution is a continuous probability density function used when an event occurs with equal probability between two extreme values, say a and b (*5*). The mean, m, and standard deviation, s, of the uniform distribution are:

$$m = (a + b)/2 \quad (1)$$

$$s = (b - a)/\sqrt{12} \quad (2)$$

Assigning the mean and the standard deviation of the uniform distribution to those of the existing interarrival time distribution, the parameters a and b are determined for the major and the minor queues.

Service times remained the same as in the original model, although if there are fewer vehicles in the driveway, the service times will be somewhat lower. All other components of the system, including the length of the simulated time, remained unchanged as well.

2. Increase driveway capacity. The effect of increasing the available space in the school driveway is tested. The only change to the original program listings is the statement where the capacity of the storage driveway is defined.

3. Remove Parkers. This alternative consists of separating those drivers parking at the school from those drivers unloading passengers.

CONCLUSIONS AND RECOMMENDATIONS

The methodology used in this study involved combining field data with simulation to produce a model that can be used to solve problems of traffic congestion. This method was demonstrated to address traffic problems at two school sites.

At Wrights Mill Road School in Auburn, the traffic congestion can be practically eliminated, if the interarrival times are more uniformly distributed, that is, if parents are urged to avoid the peak congestion time, which occurs from 7:19 to 7:32 a.m. The results of the simulation show that the predicted mean queue length will be 0.89 vehicle, which indicates that there still will be some vehicles waiting to enter the driveway and blocking through vehicles on the street. An additional right-turn bay with a capacity of 3 to 4 vehicles will eliminate that problem. It is recommended that parents be given incentives to avoid the 15-min period of peak congestion. In addition, a right-turn bay with a capacity of 3 to 4 vehicles should be constructed just before the entrance of the school driveway.

At the Opelika School it was determined that removing parking vehicles from the traffic that discharges passengers can improve the traffic conditions at the site. This means that parking spaces need to be relocated. There is additional available parking that can be accessed from a separate driveway, which leads to the south entrance of the school. These additional spaces can be used without causing any parking problems for the school employees. The parking spaces in front of the main entrance should be reserved exclusively for visitors.

It is concluded that development of a simulation model to investigate traffic congestion problems at school sties is a very efficient approach, particularly in cases for which conventional analysis (such as the HCM method) are not directly applicable. The methodology employed can be used in similar situations, such as where queueing of vehicles occurs or when the traffic patterns are too complex to describe analytically.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to Michael Caldwell, Michael Smith, and Leslie Ward for their help in data collection.

REFERENCES

1. T. J. Schriber. *Simulation Using GPSS*. John Wiley & Sons, New York, N.Y., 1974.
2. Minuteman Software. *GPSS/PC Tutorial*. 1988.
3. S. L. Solomon. *Simulation of Waiting-Line Systems*. Prentice-Hall, Englewood Cliffs, N.J., 1983.
4. S. Siegel. *Non-Parametric Statistics for the Behàvioral Sciences*. McGraw Hill, New York, N.Y., 1956.
5. L. Blank. *Statistical Procedures For Engineering, Management and Science*. McGraw Hill, New York, N.Y., 1980.

Publication of this paper sponsored by Committee on Methodology for Evaluating Highway Improvements.