TRANSPORTATION RESEARCH RECORD 596

Flow Theory

TRANSPORTATION RESEARCH BOARD

COMMISSION ON SOCIOTECHNICAL SYSTEMS
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY OF SCIENCES
WASHINGTON, D.C. 1976

Transportation Research Record 596
Price $2.20
Edited for TRB by Marjorie Moore

subject areas
53 traffic control and operations
54 traffic flow

Transportation Research Board publications are available by ordering directly from the board. They may also be obtained on a regular basis through organizational or individual supporting membership in the board; members or library subscribers are eligible for substantial discounts. For further information, write to the Transportation Research Board, National Academy of Sciences, 2101 Constitution Avenue, N.W., Washington, D.C. 20418.

Notice
The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competence and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The views expressed in this report are those of the authors and do not necessarily reflect the view of the committee, the Transportation Research Board, the National Academy of Sciences, or the sponsors of the project.

Library of Congress Cataloging in Publication Data
National Research Council. Transportation Research Board.
Flow theory.

(Transportation research record; 596)
1. Traffic flow—Addresses, essays, lectures. 2. Traffic engineering—Addresses, essays, lectures. I. Title. II. Series.
TE7.H5 no. 596 [HE336.T7] 380.5'08s [388.3'1]
ISBN 0-309-02565-6 76-58916

Sponsorship of the Papers in This Transportation Research Record

GROUP 3—OPERATION AND MAINTENANCE OF TRANSPORTATION FACILITIES
Lloyd G. Byrd, Byrd, Tallamy, MacDonald, and Lewis, chairman

Systems Performance Section

Committee on Highway Capacity and Quality of Service
Robert C. Blumenthal, Alan M. Voorhees and Associates, Inc., chairman
Arthur A. Carter, Jr., Federal Highway Administration, secretary
Donald S. Berry, Joseph W. Hess, Jack A. Hutter, Thomas D. Jordan, James H. Kell, Jerry Kraft, Joel P. Leisch, Edward B. Lieberman, Adolf D. May, Jr., William R. McShane, Louis J. Pignataro, Frederick D. Rooney, John L. Schlaefli, Gerald W. Skiles

Committee on Traffic Flow Theory and Characteristics
Kenneth W. Crowley, Polytechnic Institute of New York, chairman
Robert F. Dawson, University of Vermont, vice-chairman
Edmund A. Hodgkins, Federal Highway Administration, secretary
Patrick J. Athol, John L. Barker, Charles R. Berger, Kenneth C. Berner, Donald E. Cleveland, Leslie C. Edie, John W. Erdman, Antranig V. Gafarian, Nathan Gartner, Denos C. Gazis, Daniel L. Gerlough, John J. Haynes, Richard L. Hollinger, James H. Kell, John B. Kreer, Joseph K. Lam, Tenny N. Lam, Edward B. Lieberman, Carroll J. Messer, Richard Rothery, A. D. St. John, Sidney Weiner, W. W. Wolman

K. B. Johns, Transportation Research Board staff

Sponsorship is indicated by a footnote on the first page of each report. The organizational units and the officers and members are as of December 31, 1975.

Contents

Abridgment

Developing a Predictor for Highly Responsive System-Based Traffic Signal Control

William R. McShane, Department of Transportation Planning and Engineering, Polytechnic Institute of New York
Edward B. Lieberman and Reuben Goldblatt, KLD Associates, Inc., Huntington Station, New York

A prediction methodology must be designed to provide the traffic control policy with accurate and reliable information. The design of the control policy, the precision afforded by the surveillance system, and the formulation of the prediction algorithm must be considered interactively and explicitly. From the onset, this was the approach used in the third generation control (3-GC) in the urban traffic control system (UTCS). This paper reports on the predictor development aspects of that work (1). The prediction algorithm described is intended to be applicable only for undersaturated links. The congested flow control acts on link content rather than anticipated volumes.

Previous approaches to developing a prediction methodology were based heavily on acquiring detailed historical patterns of traffic volume on a link-specific basis in the belief that these patterns were strongly repetitive on a day-to-day basis. At the time this work was undertaken, no large data base was available, particularly none that confirmed the high degree of traffic regularity desired over short time periods of 3 to 6 min. Further, experience in the UTCS test bed argued for the lack of diurnal regularity on such a time scale. Results of work by McShane and Crowley, in a paper in this Record, indicate that considerable regularity does exist, at least in some cities. In view of the past work and the unavailability of the required data base, a totally different approach was explored.

The basic idea was to develop a methodology that did not depend on historical data. This approach implies not that traffic patterns lack any degree of repetition from one day to the next but, rather, that it would be preferable not to depend strongly on such regularity of demand in this project.

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

Three data bases consisting of five test cases were collected for this activity (Table 1).

DEVELOPMENT OF PROCESS MODELS

The process model is a mathematical description of the variation of link-specific traffic volume over time. As such, it forms the basis for development of the prediction model. That is, by knowing the form of the process, one may postulate appropriate prediction models.

The results indicated that little benefit would be derived by correlating volume on the study link to total volume on upstream links. The refinement of identification by phase is also not justified. Thus, link volumes are related to their own past history in the development that follows.

Two techniques were considered for eliminating the nonstationarity of the data over a day: (a) detrending the data and (b) differencing the volume data. A first-order autoregressive model was found to be sufficient on either basis.

PREDICTION MODELS

Development and Analysis

The prediction algorithm is processed during time step (i + 1) by using data acquired during the prior time steps [i, (i + 1), and so on] to predict the required parameter

Table 1. Description of data bases.

Data Base	Description	Test Case
L Street, Washington, D.C.	A gentle rising trend with pronounced fluctuations from one control period to the next	A
E Street, Washington, D.C.	A gentle falling trend with	
	Moderate fluctuations	B
	A sharp reversal	C
Park Avenue, Greenlawn, N.Y.	Pronounced peaking with moderate fluctuations with	
	A sustained peak	D
	A fall-off from the peak	E

Table 2. Values of the variance of the prediction error.

Predictor No.	Predictor Title	Test Case A	B	C	D	E	Sum	Rank
1	Stochastic volume	1.00	1.00	1.00	1.00	1.00	5.00	1
3	Differences	1.17	1.02	1.80	1.07	1.09	6.15	3
4	Average of mean	1.24	1.38	1.50	1.21	1.22	6.55	4
5	Linear extrapolation of mean	1.13	1.00	1.77	1.03	1.03	5.96	2
6	Parameter estimation on volume	1.08	1.13	3.66	1.23	1.18	8.28	7
7	Parameter estimation on variation about mean volume	1.18	1.35	2.56	1.00	1.02	7.11	6
8	Last known volume	1.13	1.09	2.30	1.03	1.03	6.58	5

for time step (i + 2). Hence, because there are two steps to be spanned in the prediction process, from i to (i + 2), we are concerned with developing a two-step predictor.

Evaluation

Consistent with the concept of a first-order process describing the volume sequence, a number of predictor models were formulated (Table 2), and the variance of the prediction errors was noted as follows:

Test Case	Variance	Test Case	Variance
A	288	D	78.5
B	82.1	E	115
C	103		

Predictor 1 is superior to the others over all data bases. Inspection of Table 2 reveals that, for all but one data base, there is little disparity in results among predictors 1, 3, 5, and 8—the latter being the nonpredictor. The exception, test case C, demonstrates an important advantage of predictor 1 over its competitors: its ability to cope with extreme conditions.

The stochastic volume model is a linear filtering model operating on volume measurements taken on the subject link. The computing requirements of this model are very attractive.

Use in Control Policy

The countermeasure in the control policy to predictor errors is the provision of some excess green to cushion against predictor underestimates (i.e., so that congestion is not brought on by inadequate green). If too much cushion must be provided, control flexibility is lost. In test case E, predictor 1 underestimates by three or four vehicles 5.9 percent of the time and by more than four vehicles 8.8 percent of the time. If one provides a cushion of excess green for three vehicles, there is a 0.059 + 0.088 = 0.147 probability that an individual cycle will exceed this cushion. This can be judged acceptable for the purposes of control. Most situations do not require so high a cushion for comparable probability.

The final model recommended was predictor 1: exponentially smoothed trend with a first-order two-step predictor of the variation from the trend having parameter α_j. Work by McShane and Crowley with substantially more data has since indicated that it may be best to continually adjust α_j on-line.

DISCUSSION OF RESULTS

Subsequent to this study, two extended sets of data became available for comparing the efficacy of the selected 2-GC and 3-GC predictors. To afford a comparison with the 2-GC data on a common basis (2) requires that 3-GC predictor data be analyzed for one time step into the future, inasmuch as the only 2-GC predictor data available are for the one time-step prediction.

Basically, the 2-GC predictor provides slightly more accuracy than the 3-GC predictor under conditions in which the traffic flow exhibits a regular behavior. We have not studied the condition in which traffic behavior for the study period departs significantly from the historical trend (e.g., weekends, holidays, inclement weather). As indicated, there is a small practical difference between the two predictor models. However, the 3-GC predictor requires less storage and computational time.

ACKNOWLEDGMENTS

The authors wish to acknowledge the constructive comments of P. Tarnoff and J. MacGowan of the Federal Highway Administration. The work was supported by a Federal Highway Administration contract.

REFERENCES

1. E. B. Lieberman and others. Variable Cycle Signal Timing Program: Volume 4—Production Algorithms, Software and Hardware Requirements, and Logical Flow Diagrams. KLD Associates, Inc., June 1974.
2. Urban Traffic Control System and Bus Priority System Traffic Adaptive Network Signal Timing Period—Software Description. Transportation and Environmental Operations, TRW, Aug. 1973.

Abridgment

Design of Freeway Entrance Ramps and Lane Drops by Using Semi-Markov Processes

Robert J. Abella, Cooper Tire and Rubber Company
David C. Colony and John D. Hansell, University of Toledo

A continuous-time semi-Markov highway model can be used to simulate and design highway situations that require a merging maneuver. With such a model, mathematical manipulations of probability density functions are not required. A general continuous-time semi-Markov model can be used to simulate a number of merging situations by the substitution of appropriate values for the design parameters. The substitutions are made into established general formulas. The continuous-time semi-Markov process can then be analyzed with flow graph and Laplace transform techniques.

The design parameters included in the model are freeway vehicle headway distributions, lane volumes, lane running speeds, and a gap acceptance function that describes a driver's willingness to accept a given headway in an adjacent lane.

DESCRIPTION OF CONTINUOUS-TIME SEMI-MARKOV PROCESS

A continuous-time semi-Markov process is similar to the discrete-parameter Markov process, which has been used by others to model some freeway operations, but the continuous-time semi-Markov process requires two matrices to describe transitions:

1. A transition matrix, similar to that of the discrete Markov process, that defines the probability, p_{ij}, of making a transition from state i to another state j and
2. A holding time matrix that embodies the continuous probability density function, $f_{ij}(x)$, of the time or distance associated with making a transition from state i to state j.

The continuous-time semi-Markov process is used to model a multilane, unidirectional section of highway. In the simplest formulation, the states of the semi-Markov process represent lanes of a freeway. Transitions must always occur from one lane to an adjacent lane. The probabilities associated with a freeway driver making a transition from a given lane or state to any other lane or state are obtained by calculating the probability that the gap that is adjacent to the driver in question is equal to or greater than the driver's acceptance criterion. If the driver accepts this gap, the distance associated with making the transition, or changing lanes, is a random variable and is described by the holding time matrix. If the driver rejects the first gap, he or she must drive at a different speed from that of the vehicles in the adjacent lane and wait for another opportunity to change lanes. The next opportunity occurs when he or she is adjacent to to the next gap in the traffic stream. The waiting time from one merging opportunity to the next can be described in terms of the volume of traffic in the lane being merged into, which a merging vehicle must pass, or by the volume of traffic that must pass the merging vehicle.

An example of the merging process is shown in Figure 1. The figure shows a two-lane, one-way pavement with a vehicle in lane one attempting to merge into lane two. It is assumed in this case that the vehicles in lane two are traveling at a higher rate of speed than those in lane one. If the driver in lane one is constrained by a vehicle in front of him or her, as is normally the case in moderately heavy traffic, he or she must attempt to merge by moving into gap one. If the merging driver rejects gap one, a spacing G_1 separates him or her from gap two. The distance G_1 can be expressed as a function of the traffic that must pass the merging vehicle. It will be referred to as the differential distance and is measured in meters. The differential distance associated with rejecting a gap or making a transition into the same lane is a random variable with a distribution equal to the distribution of rejected gaps. The actual distance on the freeway that the vehicle in lane one must travel while waiting for the next opportunity to merge, a differential distance G_1 away, is dependent on the difference between the mean speeds of the vehicles in lanes one and two. If there is a speed differential between lanes, the differential distance G_1 between gaps can be converted to the downstream travel distance for vehicle one by applying the speed of vehicle one. If, for example, the lane one vehicle in Figure 1 rejects gap one, the time required

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

to move into position to select gap two can be found by using the differential speed between lanes. This time multiplied by the speed, S_1, of the lane one vehicle gives the required downstream travel distance.

A flow graph of the process is shown in Figure 2. This flow graph can be reduced by using Mason's rule to yield the Laplace transform of the probability density function of the first passage time from lane one to lane two in terms of the lane differential distance. The Laplace transform of the differential distance probability density function, f(s), is

$$F(s) = [p_{12}f_{12}(s)]/[1 - p_{11}f_{11}(s)] \quad (1)$$

A variety of forms of gap density functions and gap acceptance functions could be substituted into equation 1 to find the probability density function of the differential distance.

Several methods can be used to represent a driver's willingness to accept a gap between vehicles during a merge maneuver including distributions of critical gaps and gap acceptance functions. The most convenient method of representing a driver's gap acceptance criterion for the continuous-time semi-Markov highway model is a gap acceptance function. The gap acceptance function is assumed to be stationary with respect to the distance from the lane termination point but may vary according to highway conditions and geographical location.

Weiss and Maradudin (1) showed that the probability of accepting a gap can be found by integrating the product of the gap density function and the gap acceptance function over the range of gaps. To obtain meaningful results from traffic delay calculations requires the assumption that vehicles occupy no space. The term gap is in such a case synonymous with spacing, and the former term is used throughout this paper. The probability, p, that any available gap will be accepted is therefore

$$p = \int_0^{\infty} p(x)f(x)dx \quad (2)$$

The probability of staying in lane one is 1 - p.

The distribution of rejected gaps is needed for the continuous-time semi-Markov model. The density function of rejected gaps is a conditional density function that gives the probability that any gap is less than or equal to a certain value, given that the gap is rejected. The probability that the length of a rejected gap is between x and x + Δx is therefore

$$p = [p(\text{gap between } x \text{ and } x + \Delta x) - (\text{gap between } x \text{ and } x + \Delta x \text{ accepted})]/p(\text{gap rejected}) \quad (3)$$

In terms of the gap acceptance function p(x), the gap probability density function f(x), and the probability of accepting a gap, the density function of rejected gaps, g(x), is

$$g(x) = [f(x) - f(x)p(x)]/(1 - p) \quad (4)$$

Equation 4 can be substituted for $f_{11}(x)$ in the semi-Markov process. Several sample problems are discussed elsewhere (2).

COMPARISON WITH ACTUAL FREEWAY CONDITIONS

Data were collected on the 404-m (1325-ft) auxiliary lane of I-475 in Toledo. This auxiliary lane was modeled by

Figure 1. Merging on a two-lane, unidirectional pavement.

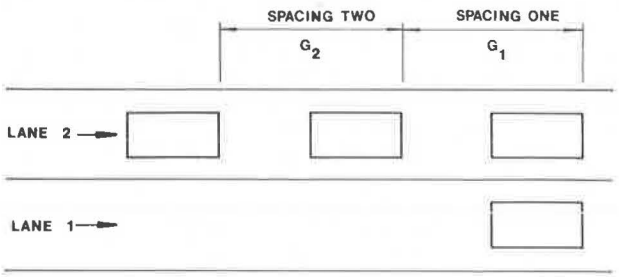


Figure 2. Entrance ramp situation.

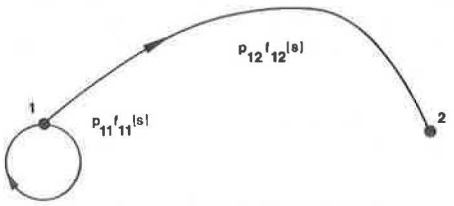


Figure 3. Comparison of observed gap acceptance characteristics and an assumed gap acceptance function.

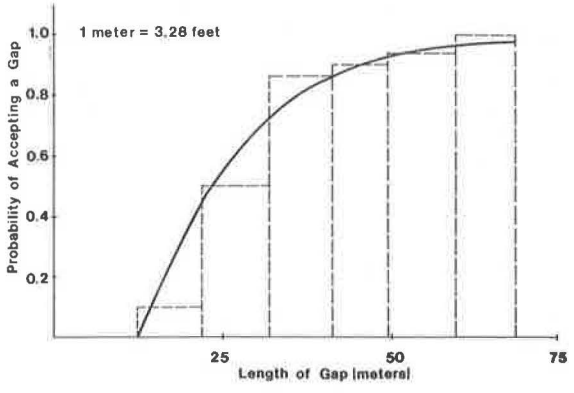
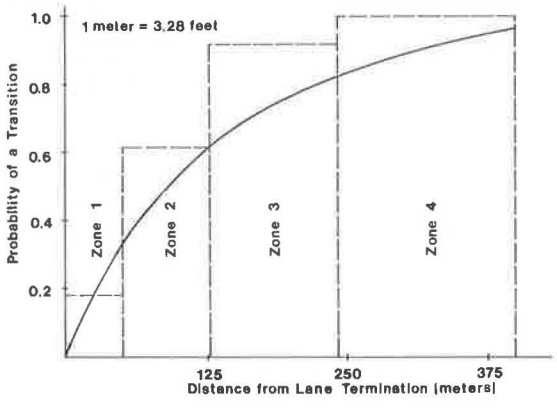


Figure 4. Comparison of results of continuous-time semi-Markov model and data collected at auxiliary lane site.



a continuous-time semi-Markov process.

The distribution of gaps during the study period was found to fit a negative exponential distribution. A mean volume of 1500 vehicles per hour and a minimum gap of 12.2 m (40 ft) were observed. A time mean speed of 82 km/h (51 mph) was measured in lane two, and the entrance ramp design was such that most entering vehicles attained a speed of approximately 74 km/h (46 mph) by the time they reached the auxiliary lane. The volume of entering and exiting traffic was low compared to the lane one volume.

A gap acceptance function was developed from motion picture data collected at the auxiliary lane site. During the study period, 123 accepted gaps and 62 rejected gaps were recorded. The gaps ranged from 12.2 to 64 m (40 to 210 ft). They were grouped in intervals of 9 m (30 ft) so that a gap acceptance function could be established. The probability that a gap in a given interval was accepted was calculated by dividing the total number of gaps in a particular time interval into the number of accepted gaps in that interval. The results are shown in Figure 3.

A gap acceptance function in the form

$$p(x) = 1 - \exp[-\lambda(x - T)] \quad x \geq T$$
$$= 0 \quad x < T \tag{5}$$

is suggested from the figure. A minimum gap, T, of approximately 12.2 m (40 ft) is suggested, with the parameter, λ, equal to 0.025. A gap acceptance function in this form and with the indicated parameters is compared in Figure 3 to the gap acceptance data collected at the site.

The auxiliary lane was modeled by using equation 5 and the negative exponential gap density function in the form

$$f(x) = a\exp[-a(x - N)] \quad N \leq x < \infty$$
$$= 0 \quad x < N \tag{6}$$

The constant 1/a is the mean gap and N is the minimum gap.

The distribution of rejected gaps was found by substituting into equation 4. Because the minimum accepted gap and the minimum observed gap between vehicles are taken to be equal, the result of this substitution is

$$g(x) = (a + \lambda)\exp[-(a + \lambda)(x - N)] \quad N \leq x < \infty$$
$$= 0 \quad x < N \tag{7}$$

The Laplace transform of g(x) is

$$L[g(x)] = (a + \lambda)/(s + a + \lambda)\exp(-Ns) \tag{8}$$

The distance required to merge into a gap was represented by the distribution of rejected gaps. Substitution of the known conditions gives the following result if q is substituted for 1 - p and c for a + λ and F(s) is the Laplace transform of the differential distance probability density function.

$$F(s) = pc/[s + c = cq\exp(-Ns)] \tag{9}$$

The inverse transform of equation 9 can be obtained by partial fraction expansion. It is an infinite series of exponential functions.

$$f(x) = pc\exp(-cx)u(x) + pcqc[(x - N)/1!]\exp[-c(x - N]\,u(x - N)$$
$$+ pc(qc)^2[(x - 2N)^2/2!]\exp[-c(x - 2N)]\,u(x - 2N)$$
$$+ \ldots \quad x \geq 0$$
$$= 0 \quad x < 0 \tag{10}$$

where

c = a + λ,
p = probability of accepting a gap, and
N = size of the minimum gap.

The probability of making a transition for a number of differential distances was calculated by using equation 10 and the measured values of the required parameter. The differential distance was converted to the corresponding freeway distance by using an 8-km/h (5-mph) differential speed. The predicted cumulative transition probabilities were compared to corresponding probabilities calculated from observations of vehicle transitions at the auxiliary lane site. The observed probabilities were found by dividing the auxiliary lane into four zones of from 45.7 to 176.6 m (150 to 550 ft) long and calculating the probability for a vehicle merge into lane one within each zone. The cumulative observed probability of making a transition within a zone and the predicted probability of making transition are shown in Figure 4. Data given in Figure 4 show that the transition probabilities predicted by the semi-Markov model are similar to the observed data.

SUMMARY AND CONCLUSION

The continuous-time semi-Markov model can be used to find the distribution of the time spent by a driver waiting to emerge from an entrance ramp. This distribution can be used to evaluate the freeway entrance ramp designs. The effect of improved visibility on waiting time, for example, could be studied by using the waiting time distribution. Different gap acceptance functions can be used to reflect the effect of improved visibility on a typical driver's merging behavior.

The location of a warning that a freeway lane drop is imminent can also be studied by means of a continuous-time semi-Markov model. Design conditions can then be calculated. This proportion can be related to the level-of-service concept. Providing a warning at a point calculated to allow 95 percent of the vehicles to merge from the lane being terminated may, for example, be associated with level of service A. Further work is required before a definitive relation can be established between the output of a semi-Markov model and levels of service as those levels are currently defined.

REFERENCES

1. G. Weiss and Maradudin. Some Problems in Traffic Delay. Operations Research, Vol. 10, No. 1, Jan.-Feb. 1962, pp. 74-104.
2. R. J. Abella. Analysis and Design of Freeway Lane Drops Using Continuous Time Semi-Markov Processes. Univ. of Toledo, doctoral dissertation, June 1975.

Simultaneous Optimization of Offsets, Splits, and Cycle Time

Nathan H. Gartner, John D. C. Little, and Henry Gabbay, Operations Research Center, Massachusetts Institute of Technology

Setting traffic signals in an urban street network involves the determination of cycle time, splits of green time, and offsets. All existing methods use a sequential procedure for calculating the traffic control variables. A common cycle time is established first, then green splits are calculated for each intersection, and, finally, offsets among the signals are determined. Because these three computation stages are not really independent, the results are often not optimal. This paper describes a new computer method, mixed-integer traffic optimization, designed to optimize simultaneously all the traffic control variables of the network. The method has been programmed in conjunction with the mixed-integer routine of IBM's MPSX optimization system and thus provides a globally optimal procedure. It was applied to several traffic signal networks and is shown to offer certain advantages over existing methods. An example is presented in the paper to illustrate the input requirements, output format, and application of the program.

The traffic control variables in a signalized street network are cycle time, green splits at each intersection, and offsets among the signals. The usual approach for determining these variables consists of three stages.

1. A common cycle time for the network is selected. The signals are then said to be synchronized. This has been shown to improve traffic flow in the network by exploiting the platooned structure of vehicular movement on the signalized links. Because the capacity of a signalized intersection is a function of the cycle length, the common cycle is usually determined by the most heavily loaded intersection, called the critical intersection.
2. Green splits are calculated separately for each signal by apportioning green times for conflicting approaches in proportion to the respective traffic loads.
3. The signals are coordinated by establishing a set of offsets that determine the relative timing among the signals. Only at this stage is a network optimization procedure commonly used.

However, inasmuch as the three stages of signal setting are not independent, this approach cannot guarantee optimality, i.e., the best possible settings for the network. Some of the methods iterate among the different stages to find an improved solution.

Among existing methods, SIGOP (1) scans a predetermined number of cycle times. For each cycle, splits are determined locally at each intersection and offsets are optimized by the OPTIMIZ subroutine by using a random search technique. Network performance is then evaluated in terms of delays and stops, by a coarse simulation of traffic flow on all the links of the network by the VALUAT subroutine. The results obtained from VALUAT dictate the selection of a set of cycle time, splits, and offsets. Three deficiencies of this procedure are apparent. First, the offset optimization procedure determines a local, not necessarily global, optimum; second, the splits are determined independently from the other variables; third, the stochastic nature of traffic flow, which can have a decisive effect on link performance, is not considered. Thus, in recent comparative evaluations of SIGOP and TRANSYT, the lower bound on cycle time was consistently selected as the best value by SIGOP; including stochastic effects would quite possibly have moved the results upward (2, 3, 4).

The combination method and TRANSYT use a critical intersection approach for determining the network cycle time (5). TRANSYT also allows for evaluation of performance at different cycle times, but in view of its rather formidable computing requirements this possibility is seldom used in practice (4). Because capacity of an intersection increases with cycle length, the critical intersection approach is based on analyzing the capacity requirements of the most heavily loaded intersection, i.e., the intersection with the highest sum of representative flow-capacity ratios on all signal phases. The network cycle time is then calculated for this intersection by a method such as Webster's (6) or by specifying the maximal degree of saturation on all approaches to the intersection. This procedure may not be optimal in a network situation because the interaction between the flows and the spatial road network structure of the area is disregarded. Splits and offsets are calculated differently by the two methods. The combination method calculates splits locally at each intersection and uses a

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

series-parallel combination procedure, based on link performance functions, for setting offsets. The procedure can be generalized to non-series-parallel networks by means of dynamic programming (7). TRANSYT (8) uses simulation of traffic flow throughout the network to calculate both splits and offsets. Optimization is performed by a hill-climbing method using a one-at-a-time variable search that requires a complete network simulation pass at each step.

Although these systematic methods have been found to achieve significant improvements in traffic performance (9, 10), they cannot guarantee an optimal solution for all the network control variables in a reasonable amount of computer time. Therefore, further improvements in traffic performance can be made by devising a method that simultaneously optimizes all the network variables.

This paper describes a new computer method and program, mixed-integer traffic optimization (MITROP), designed to achieve this goal. The program was developed as part of the urban traffic control research project sponsored by the Federal Highway Administration (11). The approach taken has been to develop a suitable network traffic model and performance measure and then to apply contemporary optimization techniques to it. The principal technique used is mixed-integer linear programming. A number of computer codes are available for this technique; the one used here was the IBM mathematical programming system (12). This paper discusses the main features of the program and presents a detailed example to illustrate the input data requirements and the output format of the program. A more comprehensive description of the program as well as documentation and application to several test networks can be found elsewhere (13).

TRAFFIC SIGNAL NETWORK OPTIMIZATION PROBLEM

The street network consists of nodes and links. The nodes represent the signalized intersections, and the links represent sections of street that carry traffic in one direction between two adjacent intersections. All parameters and variables of the system are defined below, and the traffic signal setting problem is presented as a simultaneous optimization program.

Definitions

Let S_j denote the traffic signal at node j, and let (i, j) denote the link connecting nodes i and j. We define

- $r_{ij}(g_{ij})$ = effective red (green) time at S_j facing (i, j);
- $\psi_j(AB)$ = intranode offset at S_j, measured as the time from the beginning of green on phase A to the beginning of green on phase B; and
- C = cycle time.

The relations between physical and effective signal timings are shown in Figure 1 for a single two-phase signal; they follow the basic model of traffic signal operation used by Webster and others. The signal phases are denoted by single capital letters A, B, and so on, which are replaced by letter pairs, e.g., (i, j), when they are assigned to links. Vehicle platoons released at the start of green at node i travel to node j on link (i, j). Figure 2 shows the fundamental relationship between travel time and internode offset on a link. Let

- τ_{ij} = travel time of platoon's head from i to j;
- γ_{ij} = arrival time of platoon's head at j, measured relative to start of g_{ij}, so that $-r_{ij} \leq \gamma_{ij} \leq g_{ij}$; and
- ϕ_{ij} = internode offset in extended form.

$$\phi_{ij} = \tau_{ij} - \gamma_{ij} \tag{1}$$

θ_{ij}, the internode offset in reduced form or $[\phi_{ij}]_{\text{mod } C}$ is the time from the start of green at S_i to the start of green at S_j occurring next, so that $0 \leq \theta_{ij} < C$. This time is used to calculate the physical settings for the controller.

Objective Function

The objective of the network optimization procedure is to determine signal settings (offsets, splits, and cycle time) that minimize the disutility encountered by vehicles traveling through the signalized intersections. The particular type of disutility can be set by the traffic engineer and may include delays, stops, acceleration noise, or some combination of these measures. The most widely used measures of performance are delays and stops. Because of the inherent fluctuations in the traffic flow process, which induce random variations about the mean in variables such as total flow and temporal distribution of arrivals within a cycle, it is common to separate the total disutility or performance function into two components. The first component is associated with the mean of the traffic flow process. This component is represented in MITROP by the link performance function (LPF). The second component is associated with the random variations about the mean and is represented by the saturation deterrence function (SDF). Our goal is to minimize the total disutility in the network D. Therefore, the objective function is

$$\text{Min } D = \min \sum_{(i,j)} [(\text{LPF})_{ij} + (\text{SDF})_{ij}] \tag{2}$$

LPF and SDF are aggregated over all links (i, j) in the network, including both internal and input links. In general,

$$(\text{LPF})_{ij} = f_{ij} z_{ij}(\phi_{ij}, r_{ij}, C) \tag{3}$$

$$(\text{SDF})_{ij} = Q_{ij}(r_{ij}, C) \tag{4}$$

where,

- f_{ij} = average flow on link (i, j) in vehicles/h.
- z_{ij} = average loss per vehicle on link (i, j) for traveling through the signal at node j (delay, stops); z_{ij} is a function of all the signal-control variables, i.e., offset, split, and cycle time.
- Q_{ij} = the average overflow queue at the signal stop line on link (i, j); that is, the average number of vehicles that because of the random fluctuations are unable to clear the intersection during the cycle in which they arrive. Q_{ij} turns out to be a function of split and cycle time, but not of offset.

The objective function D represents a loss rate in the network such as vehicle-hours per hour, vehicle-seconds per hour, or vehicle stops per hour. A more detailed description, as well as comparison with field data, is given subsequently.

Constraints

The constraints of the optimization program must represent the street network structure as well as all important relationships among the decision variables of

the program. For each link in the system we have the following relationship:

$$r_{ij} + g_{ij} = C \tag{5}$$

Equation 5 also implies the relationships between settings on opposing links, involving the lost times for each phase (Figure 1). If the signal is to provide sufficient capacity on each link, then

$$s_{ij}g_{ij} \geq f_{ij}C \tag{6}$$

i.e., the flow arriving during one cycle must not exceed the available capacity during that cycle (s_{ij} denotes the saturation flow rate on the link). To facilitate pedestrian crossing we have a minimum red requirement, $r_{ij} \geq (r_{ij})_{min}$. Because there is little gain in capacity with very long cycle times, we should have an upper limit on the cycle length C_{max}. A maximum cycle length also prevents drivers from becoming impatient or believing that the signals are defective. For safety reasons as well as capacity requirements, it is also desirable to have a lower limit on the cycle time C_{min}.

An important physical contraint in any synchronized

Figure 1. Relation between physical and effective signal timings of basic model for traffic signal operation.

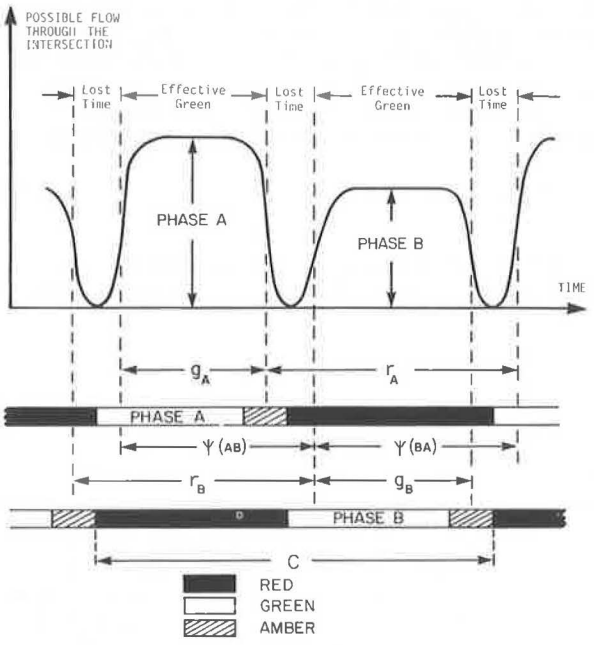


Figure 2. Travel time and offsets on a signal-controlled link.

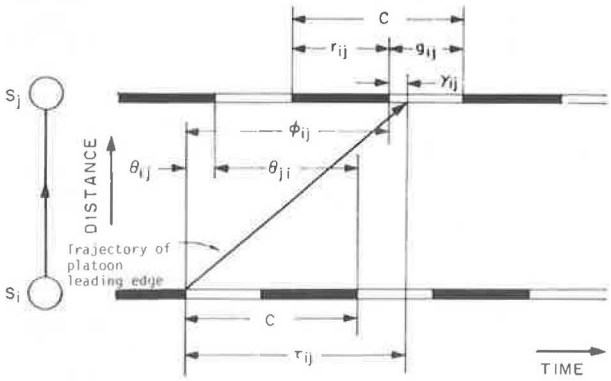


Figure 3. Traffic transition process through a link's exit signal.

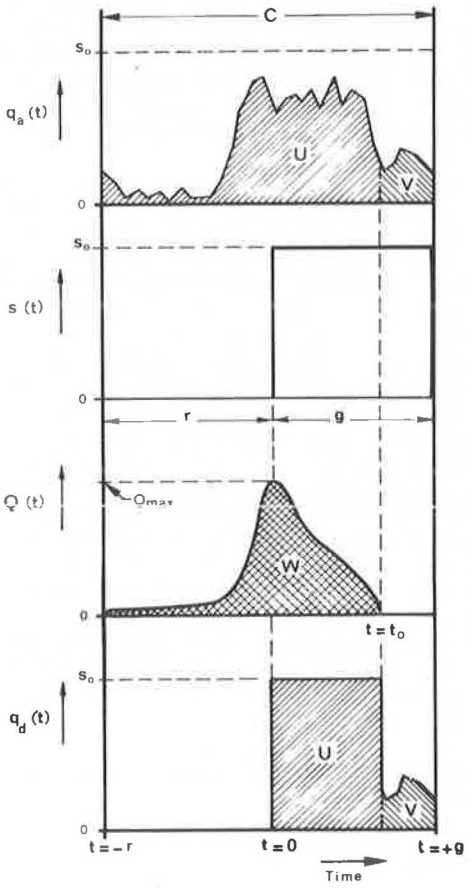
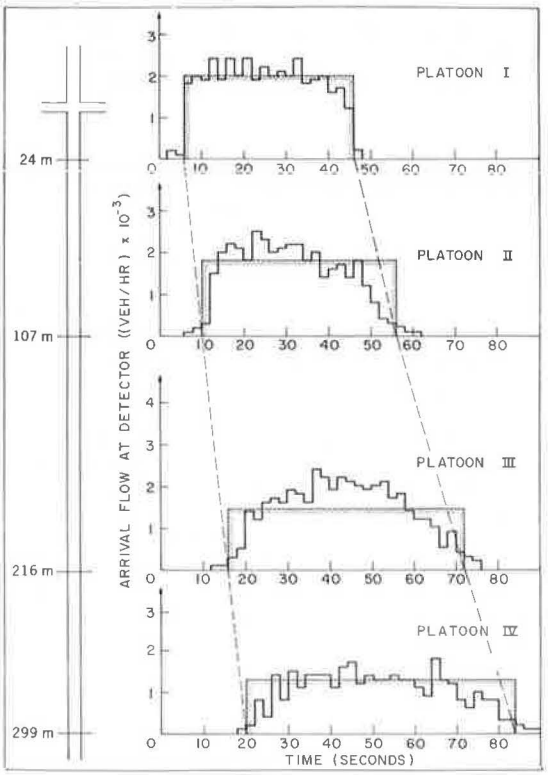


Figure 4. Platoon flow and rectangular approximation at four observation points along a signal-controlled link.



traffic signal network is that the algebraic sum of the offsets around any closed loop of the network equals an integer multiple of the cycle time (14). Both internode and intranode offsets have to be included. In mathematical form, the constraint reads

$$\sum_{(i,j)\in l} \phi_{ij} + \sum_{j\in l} \psi_j(l) = n_l C \quad \text{for each loop } l \tag{7}$$

where n_l = an integer associated with loop l.

Optimization Program

We are now in a position to formulate the traffic signal network optimization problem as the following optimization program: Find values of ϕ_{ij}, r_{ij}, C to

$$\text{Min} \sum_{(i,j)} [f_{ij} z_{ij}(\phi_{ij}, r_{ij}, C) + Q_{ij}(r_{ij}, C)] \tag{8}$$

subject to

$$\sum_{(i,j)\in l} \phi_{ij} + \sum_{j\in l} \psi_j(l) = n_l C \quad \text{for each loop } l \tag{9}$$

$$r_{ij} + g_{ij} = C \quad \text{for each link } (i, j) \tag{10}$$

$$g_{ij} s_{ij} \geq f_{ij} C \quad \text{for each link } (i, j) \tag{11}$$

$$r_{ij} \geq (r_{ij})_{min} \quad \text{for each link } (i, j) \tag{12}$$

$$C_{min} \leq C \leq C_{max} \tag{13}$$

r_{ij}, $g_{ij} \geq 0$; ϕ_{ij}, n_l are unrestricted in sign.

The MITROP processor linearizes in pieces the nonlinear components of the objective function so that the program can be solved by mixed-integer linear programming. The MPSX system then uses branch-and-bound techniques to determine simultaneously the optimal values for the continuous signal-control variables and the associated loop integer values n_l.

LINK PERFORMANCE FUNCTION

This section describes the transition process of traffic through a link and the computational procedure for determining link performance. The beginning of green time at S_j (Figure 2) is established as a reference point. Thus, a cycle period (-r, g) consists of an effective red period (-r, 0) and an effective green period (0, g). We use the following notation:

$q_a(t)$, $q_d(t)$ = arrival, departure rate in vehicles/s,
$A(t)$, $D(t)$ = cumulative number of arrivals, departures at time t during a cycle, and
s = saturation flow rate during the green period in vehicles/s.

From the beginning of any red period at S_j, the following relations exist:

$$A(t) = \int_{-r}^{t} q_a(\tau)d\tau \qquad D(t) = \int_{-r}^{t} q_d(\tau)d\tau \tag{14}$$

If the signal is undersaturated, all vehicles arriving during a cycle in which the red period precedes the green can be accommodated in that cycle. Therefore, all performance calculations for the link can be confined to a single interval (-r, g). The queue length Q(t) is given by the difference between the cumulative number of arrivals and the cumulative number of departures:

$$\begin{aligned} Q(t) = A(t) - D(t) &= A(t) \quad \text{if } -r < t \leq 0 \\ &= A(t) - ts \quad \text{if } 0 < t \leq t_0 \\ &= 0 \quad \text{if } t_0 < t \leq g \end{aligned} \tag{15}$$

t_0 denotes the queue clearance time and is obtained by solving

$$Q(t_0) = A(t_0) - t_0 s = 0 \tag{16}$$

The departure rate is described by

$$\begin{aligned} q_d(t) &= 0 \quad \text{if } -r < t \leq 0 \\ &= s \quad \text{if } 0 < t \leq t_0 \\ &= q_a(t) \quad \text{if } t_0 < t \leq g \end{aligned} \tag{17}$$

The complete transition process is shown in Figure 3. This basic model will now be used to calculate the LPF in terms of the delay encountered by the vehicles. Other performance measures such as number of stops, acceleration noise, or the closely related measure of energy consumption can be similarly calculated. For a general approach to calculating these measures of performance see, for example, Huddart or Chung and Gartner (15, 16).

The delay incurred by Q(t) queuing vehicles during an interval dt is Q(t)dt. Therefore, the total delay time Z incurred by traffic during a full cycle (-r, g) is represented by the area under the queue length curve, i.e.,

$$Z(\gamma, r) = \int_{-r}^{g} Q(t)dt = \int_{-r}^{t_0} Q(t)dt \tag{18}$$

The size of this area depends on the arrival time γ of the platoon of vehicles at the signal and, through equation 1, on the offset ϕ and the split r (and, indirectly, on the cycle time C). The average delay per vehicle z is obtained by dividing by the total number of arrivals during one cycle, A_c, which can be calculated by equation 14.

$$z(\gamma, r) = Z(\gamma, r)/A_C \tag{19}$$

MITROP uses a simple model to calculate the LPF. Traffic flow is represented by a rectangular platoon, which is generally made up of a primary component and a secondary component. Alternative assumptions, such as a tadpoling flow pattern (17), could equally well be made. The platoon must correspond to the average flow, f, on the link. Dispersion effects are taken into account via a platoon dispersion factor, which is a function of the link's length and which may be calibrated for each link. Because this traffic flow model is a simplification of reality, an assessment was made of its quality for calculating delay. Comparisons were made between delay computed from the model and that of actual platoons as reported by Hillier and Rothery (18). In their observations, individual vehicle arrivals were recorded at four locations downstream of a signal-controlled intersection operating on a 90-s cycle. The data were subsequently averaged to give the mean number of arrivals by 2-s intervals at each location. The results are shown in Figure 4. The spreading, or dispersion, that takes place along the approximately 305 m (1000 ft) of roadway downstream of the intersection can readily be seen; e.g., the platoon has a 72-s passage time at the fourth location as compared to the 40-s effective green

time at the intersection.

For the comparison, observed platoons were approximated by rectangular platoons containing the same number of vehicles. Dispersion was represented by a piecewise linear function of distance. This yielded the platoons seen superimposed on the observed data of Figure 4. Equations 14 through 19 were then used to calculate delay as a function of arrival time γ for both observed platoons and the rectangular approximations. Various splits were chosen to cover a wide range of possible degrees of saturation of the signals at the different locations. A sample of results of the comparisons is shown in Figure 5. A more extensive evaluation is given elsewhere (13). In most cases we find a close fit between the two functions. The largest deviations occur at extremes of splits or arrival times (and hence offsets), which are expected only rarely in practice. In some cases, the fit, can be improved through better selection of the parameters for the rectangular platoon approximating the actual field data. To be used in MITROP the LPFs are piecewise linearized as shown by the dotted lines in Figure 5. It is noteworthy that in certain regions the linearized approximation is closer to the function derived from the actual platoon than to the function obtained from the rectangular platoon.

SATURATION DETERRENCE FUNCTION

The LPFs are calculated based on the assumption that the traffic flow patterns are identical during each cycle. In practice there are fluctuations about the mean in variables such as total vehicle flow and the temporal distribution of vehicle arrivals within a cycle because of variations in driving speeds, marginal friction, and turns. These random variations can lead to temporary overflow queues that seriously degrade performance. Although this effect is negligible at low degrees of saturation, its predominance at high values has been established in several studies (18, 19, 20). A representation of this effect is needed to prevent green time from approaching its lower bound too closely, thus leading to saturation. If cycle time is also a variable, it is particularly essential that it be determined by the optimization process. This is because there is a fundamental trade-off between capacity loss at short cycles and the inherently large delays of long cycles.

The value of the expected overflow queue, based on the capacity of the signal's approach and the degree of saturation, has been calculated by Wormleighton (21). Following field studies in Toronto, he developed a model describing traffic behavior along a signalized link as a nonhomogeneous Poisson process with a periodic intensity function. According to our notation, the signal's capacity K, in vehicles per cycle, is

$$K = sg \tag{20}$$

and the degree of saturation x is

$$x = fC/sg \tag{21}$$

The results of the computation are given in Table 1. A typical relationship between expected overflow queue and split time in this model is shown in Figure 6. Rather similar deterrence functions have been given by Webster (6) for setting signals at a single intersection and by Robertson (9) for the TRANSYT signal setting model. The rate of delay incurred by the queued vehicles is simply $Q_{ij}(r_{ij}, C)$. This term is called the link's saturation deterrence function and provides the second component of the network objective function given in equation 2. To represent the SDF in a form amenable to mixed-integer linear programming, MITROP makes it piecewise linear, as was done with the LPF. To ensure a minimum level of service, r is restricted so that the degree of saturation stays below 0.95. Thus we obtain an upper limit on the red split, which appears as a vertical constraining line in the (Q, r) diagram. Two additional lines are determined by the two pairs of points (P_1, P_2) and (P_3, P_4), having degrees of saturation $x_1 = 0.95$, $x_2 = 0.90$ and $x_3 = 0.85$, $x_4 = 0.70$ respectively (Figure 6).

SIGNAL NETWORK OPTIMIZATION EXAMPLE

The following data describe a signalized street network for which optimal settings are determined by MITROP by using the IBM 370/165 computer system. Figure 7 shows a sketch of the test network. There are 9 intersections (nodes) and 24 links of which 16 are internal to the network (i.e., interconnect a pair of nodes) and 8 are input links. Table 2 gives the data for the network. Traffic flow on the input links is assumed to be continuous (though random fluctuations are taken into account), which therefore yields a platoon length of one complete cycle time. Table 3 gives the loops of the network; only one independent set of loops has to be considered. Each loop is specified in terms of its links and nodes. For each link (i, j) there is a corresponding internode offset ϕ_{ij} and for each node j there is a corresponding intranode offset ψ_j. These offsets are arranged into constraints according to equation 7, there being a total of eight. In case the orientations of a loop, as given in Table 3, and a link in that loop do not coincide, a negative sign has to be taken for that offset. Table 3 also gives the set of integers, n_l, that has been determined by the optimal solution for each loop l in the network. Table 4 gives the main output data for the network. The optimal value for the objective function includes both deterministic and stochastic delays, according to equation 2.

ANALYSIS OF RESULTS

Experience of researchers and practitioners in urban traffic control has shown that the cycle time may well be the most important of the signal-control variables. This is suggested by both U.K. and U.S. studies (5, 9). The cycle time provides for the necessary capacity to serve the traffic demand at each intersection, and it is the prime determinant of the possible coordination strategy among the signals in the network. Using Webster's notation, we have at each node j in the network the following relationship:

$$\sum_i g_i = C - L_j \tag{22}$$

i.e., the sum of effective green times on all phases i equals the net green time available for crossing the intersection (cycle time minus lost time). Because L_j, the total lost time at node j, is a fixed quantity, the net capacity increases with cycle length. This exposes a trade-off that is amenable to optimization. On the one hand, an increase in cycle time usually increases red times on individual phases and, consequently, the mean waiting time as expressed by LPF. On the other hand, an increase in cycle time also increases capacity and thus reduces stochastic delay due to the randomness in arrivals of vehicles, as expressed by SDF. The interplay between LPF and SDF as a function of cycle time

Figure 5. Comparisons of link performance function for actual and rectangular platoons.

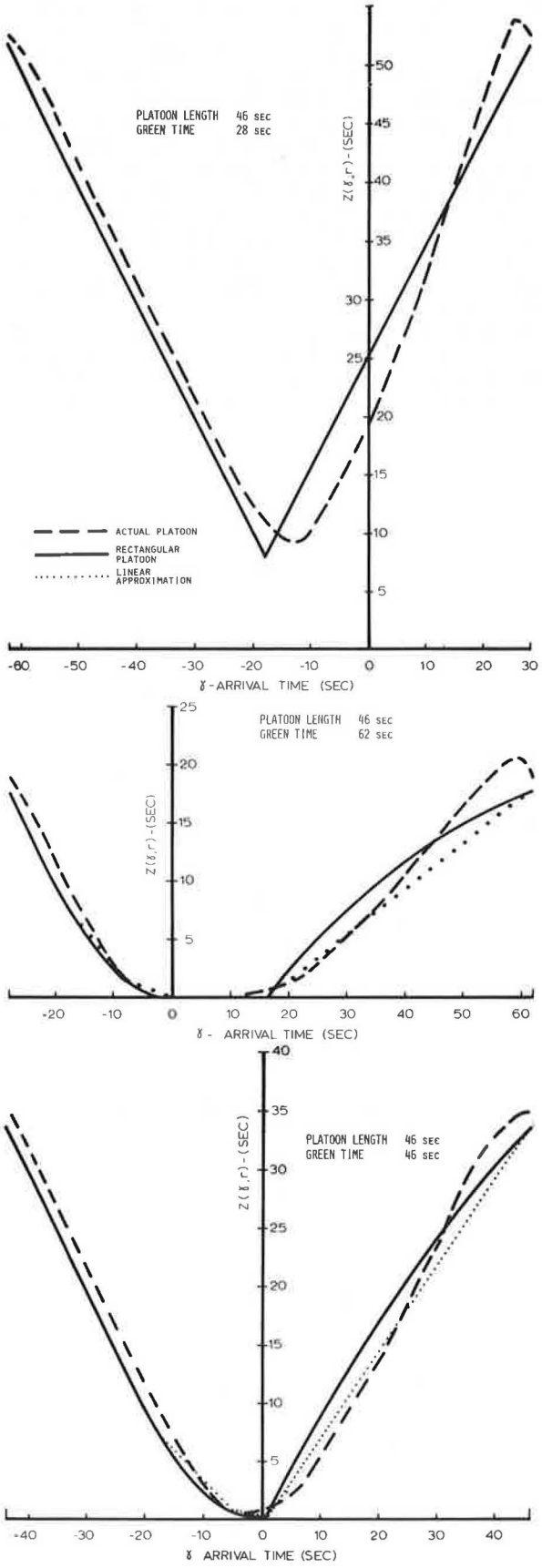


Table 1. Expected overflow queue.

Signal Capacity (vehicles/ cycle)	Degree of Saturation						
	0.20	0.40	0.60	0.80	0.90	0.95	0.975
5	0.00	0.02	0.20	1.15	3.50	8.41	18.36
15	0.00	0.00	0.04	0.70	2.81	7.61	17.50
25		0.00	0.01	0.47	2.41	7.08	16.91
35			0.00	0.34	2.11	6.68	16.45
45				0.23	1.88	6.34	16.05
55					1.68	6.02	15.67

Figure 6. Saturation deterrence function and piecewise linear approximation.

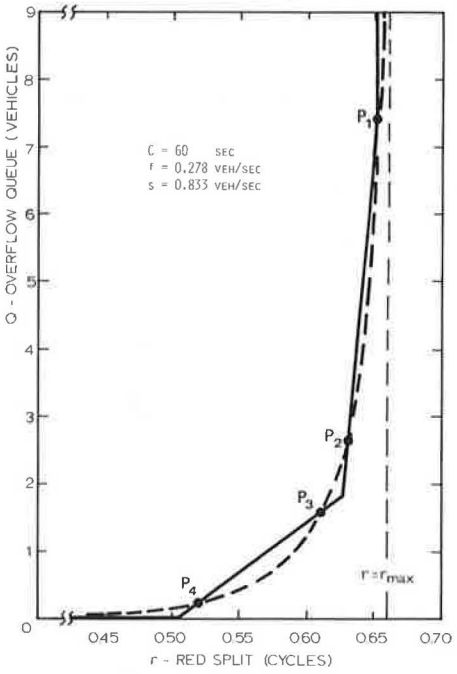
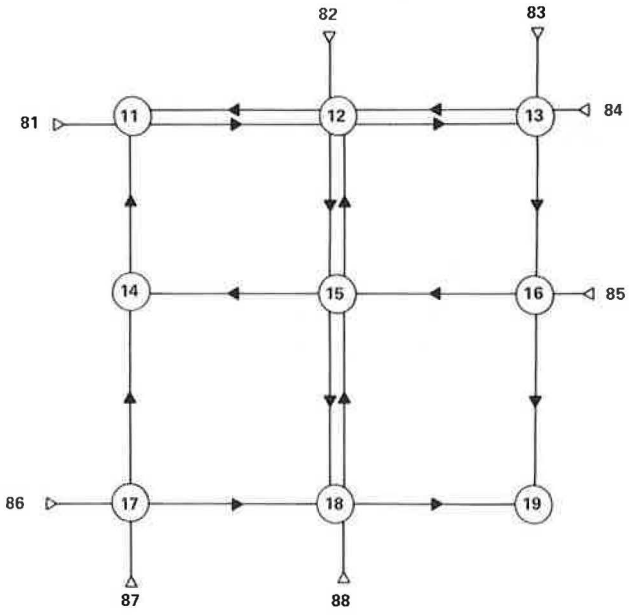


Figure 7. Test network diagram.



has been studied for the signal network example shown in Figure 7. This interplay is shown in Figure 8. It can be seen that the optimal network cycle time is determined as a least cost equilibrium point between delays attributed to the mean traffic flow component and delays attributed to the stochastic component. MITROP determines this point simultaneously with all the other traffic control variables.

The test network was further used to analyze the sensitivity of performance with respect to volume. All volumes in the network were changed by the same percentage, and MITROP was used to optimize offsets and splits for various cycle times. The results are shown in Figure 9. The decisive role played by the cycle time in reaching optimum operating conditions is illustrated here even more emphatically than previously.

CONCLUSIONS

The MITROP program formulates the traffic signal network optimization problem for mixed-integer linear programming. Thus, a global optimal solution to the problem can be achieved via an existing optimization package. Whereas much previous research has been devoted to developing a suitable optimization procedure, the emphasis here is on accurately modeling the traffic flow process and its performance measures. The branch-and-bound procedure used by the MPSX code has proved to be quite efficient in handling the traffic signal network problem because the number of integer variables versus the number of continuous variables in the problem is relatively small. Mixed-integer programming is an active area of research, and further improvements in algorithms can be expected in the future (22). Such developments might make it possible to use MITROP in an on-line mode.

The study also indicates the importance of having all the control variables of the system as simultaneous decision variables. Performance may be significantly degraded when a sequential decision process is used.

Table 2. Input data for test network (Figure 7).

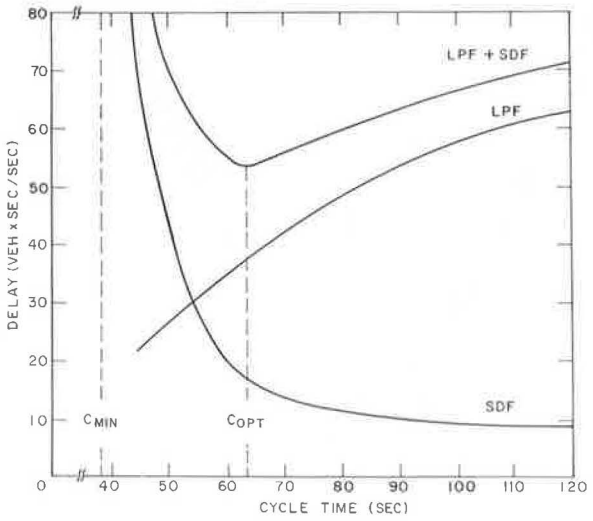
Links	Length (m)	Speed (km/h)	Travel Time (s)	Volume (vehicles/h)	Saturation Flow (vehicles/h)	Platoon Length (cycle)
Internal						
(11, 12)	305	39	28.41	630	1800	0.701
(12, 13)	168	39	15.63	630	1800	0.532
(13, 12)	168	45	13.39	400	1800	0.425
(12, 11)	305	45	24.35	400	1800	0.608
(16, 15)	168	45	13.39	430	1300	0.414
(15, 14)	305	45	24.35	430	1300	0.595
(17, 18)	305	39	28.41	630	1800	0.701
(18, 19)	168	39	15.63	630	1800	0.503
(17, 14)	244	35	24.79	550	3000	0.344
(14, 11)	183	35	18.60	550	3000	0.335
(12, 15)	183	45	14.61	800	3000	0.407
(15, 18)	244	45	19.48	800	3000	0.444
(18, 15)	244	35	24.79	550	3000	0.463
(15, 12)	183	35	18.60	550	3000	0.418
(13, 16)	183	45	14.61	900	2160	0.515
(16, 19)	244	45	19.48	900	2160	0.560
Input						
(81, 11)				630	1800	1.0
(82, 12)				800	3000	1.0
(83, 13)				900	2160	1.0
(84, 13)				400	1800	1.0
(85, 16)				430	1300	1.0
(86, 17)				630	1800	1.0
(87, 17)				550	3000	1.0
(88, 18)				550	3000	1.0

Note: 1 m = 3.3 ft; 1 km/h = 0.6 mph.

Table 3. Loops of the test network and their corresponding integer variables.

Loop	Links and Nodes in the Loop	Optimal Integer
1	(11, 12), (12, 11)	1
2	(15, 12), (12, 15)	1
3	(18, 15), (15, 18)	1
4	(12, 13), (13, 12)	1
5	(14, 11), 11, (11, 12), 12, (12, 15), 15, (15, 14), 14	-1
6	(15, 12), 12, (12, 13), 13, (13, 16), 16, (16, 15), 15	-1
7	(17, 18), 18, (18, 15), 15, (15, 14), 14, (17, 14), 17	0
8	(16, 15), 15, (15, 18), 18, (18, 19), 19, (16, 19), 16	0

Figure 8. Variation of network performance function with cycle time.



Of particular importance is the cycle time, which is strongly affected by the flows in the network. In this context, it is essential to adequately model the stochastic behavior of traffic, which in MITROP is represented by SDF. If, for instance, the cycle time for the test network were determined by Webster's method, the result would be close to 80 s (node 13 in Figure 7 is the most heavily loaded intersection). Inspection of Figure 8 shows that the objective function for this cycle time is roughly 12 percent higher than the optimum value at 63.8 s.

A useful extension of MITROP is that additional performance measures can be assigned to sections of the network, if so desired. MITROP assumes that the LPF is a function of offsets only on that link, similar to assumptions made by SIGOP and the combination method (23). It has been shown that in certain cases, particularly on lightly traveled links, it may be advantageous to follow vehicular movement on two links or more (4, 17). This can be easily done in MITROP by maximizing bandwidth on selected arterial sections, which provides this coupling feature, while simultaneously minimizing delay on other sections of the network. The bandwidth maximization problem was formulated in the past in terms of mixed-integer linear programming by Little (24) and is therefore compatible with the MITROP optimization model.

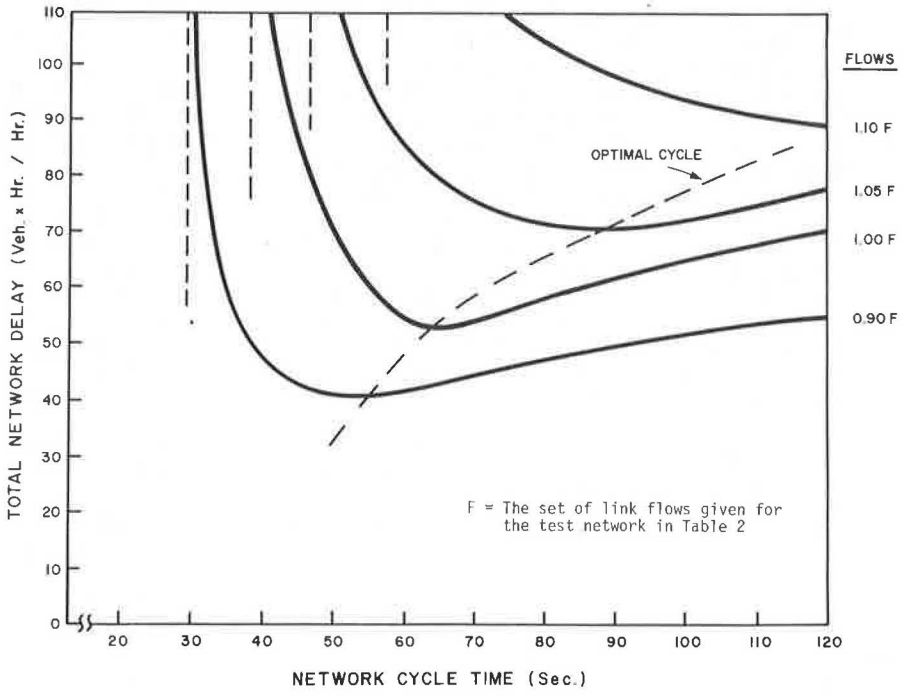
In addition to its use in the network example given in this paper, MITROP was also used for a portion of the urban traffic control system/bus priority system (UTCS/BPS) in Washington, D.C., containing 20 nodes, 63 links, and 21 independent loops, and for an arterial

Table 4. Optimal settings and output data for test network.

Links	Offset		Green Time		LPF (vehicle-h/h)	SDF (vehicles)	Degree of Saturation
	Seconds	Cycle	Seconds	Cycle			
Internal							
(11, 12)	44.6	0.70	28.7	0.45	1.416	1.255	0.778
(12, 13)	33.8	0.53	24.9	0.39	2.175	2.244	0.897
(13, 12)	30.0	0.47	28.7	0.45	0.543	0.0	0.489
(12, 11)	19.1	0.30	37.6	0.59	0.315	0.0	0.373
(16, 15)	30.6	0.48	33.1	0.52	1.752	0.0	0.641
(15, 14)	33.8	0.53	37.6	0.59	0.566	0.0	0.565
(17, 18)	24.9	0.39	30.6	0.48	1.795	0.690	0.729
(18, 19)	24.2	0.38	24.9	0.39	0.724	2.244	0.897
(17, 14)	29.3	0.46	17.2	0.27	0.339	0.0	0.674
(14, 11)	28.0	0.44	17.2	0.27	0.736	0.0	0.674
(12, 15)	25.5	0.40	21.7	0.34	1.318	0.943	0.775
(15, 18)	31.2	0.49	24.2	0.38	1.263	0.0	0.693
(18, 15)	32.5	0.51	21.7	0.34	0.600	0.0	0.535
(15, 12)	38.2	0.60	24.2	0.38	1.889	0.0	0.479
(13, 16)	21.0	0.33	30.6	0.48	0.780	1.660	0.868
(16, 19)	25.5	0.40	30.0	0.47	0.750	1.930	0.887
Input							
(81, 11)			37.6	0.59	1.650	0.0	0.593
(82, 12)			24.2	0.38	3.771	0.0	0.693
(83, 13)			30.0	0.47	3.855	1.930	0.887
(84, 13)			24.9	0.39	1.767	0.0	0.564
(85, 16)			23.6	0.37	2.195	2.714	0.901
(86, 17)			33.1	0.52	2.119	0.0	0.673
(87, 17)			21.7	0.34	2.742	0.0	0.535
(88, 18)			24.2	0.38	2.450	0.0	0.479

Note: Cycle time = 63.8 s; objective function = 53,120 vehicle-h/h.

Figure 9. Sensitivity of network performance function with respect to cycle time and flows.



street with 11 signals in Waltham, Massachusetts. The results are described in detail in a technical report (13). The MITROP computer program is currently operational and available from the Operations Research Center at the Massachusetts Institute of Technology.

ACKNOWLEDGMENT

The research reported in this paper was supported in part by the Federal Highway Administration, U.S. Department of Transportation, and in part by the Army Research Office. The contents of this paper reflect the views of the authors and do not necessarily reflect the official views or policies of the sponsoring agencies.

REFERENCES

1. SIGOP: Traffic Signal Optimization Program, A Computer Program to Calculate Optimum Coordination in a Grid Network of Synchronized Traffic Signals. Traffic Research Corp., New York, 1966.
2. J. A. Kaplan and L. D. Powers. Results of SIGOP-TRANSYT Comparison Studies. Traffic Engineering, Vol. 43, No. 12, 1973.
3. P. D. Whiting. The Glasgow Traffic Control Experiment: Interim Report on SIGOP and TRANSYT. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Rept. LR 430, 1972.
4. C. J. MacGowan and H. S. Lum. SIGOP or TRANSYT? Traffic Engineering, Vol. 45, No. 4, 1975, pp. 46-50.
5. J. Holroyd. The Practical Implementation of Combination Method and TRANSYT Programs. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Rept. LR 518, 1972.
6. F. V. Webster. Traffic Signal Settings. Road Research Technical Paper 30, Her Majesty's Stationery Office, London, 1958.
7. N. Gartner and J. D. C. Little. Generalized Combination Method for Area Traffic Control. TRB, Transportation Research Record 531, 1975, pp. 58-69.
8. D. I. Robertson. TRANSYT: A Traffic Network Study Tool. U. K. Road Research Laboratory, Crowthorne, Berkshire, England, Rept. LR 253, 1969.
9. F. A. Wagner, F. C. Barnes, and D. L. Gerlough. Improved Criteria for Traffic Signal Systems in Urban Networks. NCHRP, Rept. 124, 1971.
10. L. Rach, J. K. Lam, D. C. Kaufman, and D. B. Richardson. Evaluation of Off-Line Traffic Signal Optimization Techniques. TRB, Transportation Research Record 538, 1975, pp. 48-58.
11. P. J. Tarnoff. The Results of FHWA Urban Traffic Control Research: An Interim Report. Traffic Engineering, Vol. 45, No. 4, 1975, pp. 27-35.
12. Mathematical Programming System Extended—Mixed Integer Programming (MIP) Program Description. Technical Publications Department, IBM Corp., White Plains, N.Y., Program 5734-XM4, 1971.
13. N. Gartner, J. D. C. Little, and H. Gabbay. Optimization of Traffic Signal Settings in Networks by Mixed-Integer Linear Programming. Operations Research Center, MIT, Cambridge, Technical Rept. 91, March 1974.
14. N. Gartner. Constraining Relations Among Offsets in Synchronized Signal Networks. Transportation Science, Vol. 6, 1972, pp. 88-93.
15. K. W. Huddart. The Importance of Stops in Traffic Signal Progressions. Transportation Research, Vol. 3, 1969, pp. 143-150.
16. C. C. Chung and N. Gartner. Acceleration Noise as a Measure of Effectiveness in the Operation of Traffic Control Systems. Operations Research Center, MIT, Cambridge, Working Paper OR 015-73, 1973.
17. K. W. Huddart and E. D. Turner. Traffic Signal Progressions—GLC Combination Method. Traffic Engineering and Control, Vol. 11, 1969, pp. 320-327.
18. J. A. Hillier and R. Rothery. The Synchronization of Traffic Signals for Minimum Delay. Transportation Science, Vol. 1, 1967, pp. 81-94.
19. W. Jacobs. A Study of the Averaged Platoon Method for Calculating the Delay Function at Signalized Intersections. Department of Civil Engineering, Univ. of Toronto, BASc thesis, 1972.
20. N. Gartner. Microscopic Analysis of Traffic Flow Patterns for Minimizing Delay on Signal-Controlled Links. HRB, Highway Research Record 445, 1973, pp. 12-23.
21. R. Wormleighton. Queues at a Fixed Time Traffic Signal With Periodic Random Input. Journal of Canadian Operations Research Society, Vol. 3, 1965, pp. 129-141.
22. W. D. Northup and J. F. Shapiro. Design and Computational Experience With an Interactive Mixed Integer Programming System. Paper presented at ORSA-TIMS Joint National Meeting, Chicago, April-May 1975.
23. J. A. Hillier. Appendix to Glasgow's Experiment in Area Traffic Control. Traffic Engineering and Control, Vol. 7, No. 9, 1966, pp. 569-571.
24. J. D. C. Little. The Synchronization of Traffic Signals by Mixed-Integer Linear Programming. Operations Research, Vol. 14, 1966, pp. 568-594.

Discussion

Dale W. Ross, DARO Associates, Inc.

The new computer method MITROP, which the authors have developed for optimizing traffic signal timings, is a significant contribution to the field of traffic control. It is the first comprehensive, mathematically rigorous method to be developed for the simultaneous optimization of cycle time, splits, and offsets in an arbitrary traffic signal network.

Although the name of the method, mixed-integer linear programming, is descriptive of the method used, it does not spotlight the essential contribution of the authors. It is not the use of mixed-integer linear programming that is new; Little (24) applied mixed-integer linear programming 10 years ago. Rather, it is the modeling of the traffic flow and of the signal timing relations that is new. The authors have done a commendable job of (a) realistically modeling traffic flow, (b) accounting for the important real-life constraints on signal timings in their model, (c) leaving all variables—cycle time, splits, and offsets—as optimizable variables in their model, and (d) judiciously approximating the link delay functions by piecewise linear convex functions so that the model is amenable to optimization by using a standard, off-the-shelf mathematical programming package that guarantees a global optimum.

The authors' modeling has some important ramifications. The treatment of random variations in traffic flow through a saturation deterrence function (SDF) for each link probably has the most effect. The authors have

shown how essential it is to include the SDF so that capacity loss at short cycles is prevented. Indeed, it is the interplay of the deterministic and stochastic delay components that makes cycle length optimization important, as the authors have illustrated in Figures 8 and 9. The inclusion of the SDF model also highlights the importance of allowing signal splits to be freely optimized. Previous signal timing programs such as SIGOP and TRANSYT are deficient in freely optimizing cycle time and splits. SIGOP does not model the stochastic delay component at all and only evaluates cycle times that are preselected by the user. As the authors point out, in past applications, SIGOP consistently selected the lower bound on cycle length. TRANSYT uses a critical intersection approach for determining network cycle time and ignores the interaction between traffic flows over the spatial structure of the traffic signal network.

If there is one aspect of the MITROP modeling that could be improved it is the platoon modeling. Currently, platoons are assumed by MITROP to be uniform in flow rate over the green time of an upstream signal. This has a decoupling effect of making platoon flow along one link independent of that along adjoining links. This may affect the ability of the model to develop good progressive movements along major streets within a grid network. Another possible model improvement would be to allow the platoon speed along individual links to be variable within limits; in fact, platoon speed could possibly be yet another optimization variable. Little (24) allowed speed to be variable, and Leuthardt (25) has recently shown that allowing speed to be optimizable can lead to improved signal progressions.

The required computer time for MITROP appears to be less than that for TRANSYT and more than that for SIGOP for the same or similar networks. I draw this tentative conclusion from (a) computer time data given by the authors (13) for an IBM 370/165 computer, (b) computer time formulas given for TRANSYT and SIGOP for an IBM 360/65 (4), and (c) a typical ratio of 3.15 for the computer time of similar scientific programs on an IBM 360/65 versus a comparable IBM 370/165 system (27). Although the authors do not give computer storage requirements for MITROP, its use of the IBM mathematical programming package probably makes it more consumptive of computer memory than either SIGOP or TRANSYT. Its storage requirements probably preclude its use, in present form, for real-time traffic control. Because few people have used MITROP, it remains to be seen how easy it is to use. Specifying input data may be a formidable effort since a large number of constraints must be defined for the mixed-integer linear program. However, it is my understanding that the authors have developed a preprocessor to assist in the data input.

I have some suggestions for improving the mathematical optimization portion of MITROP.

1. One way of reducing the computer memory (particularly main memory, as opposed to bulk memory) requirements of MITROP and thereby making it more amenable to real-time application might be to scan cycle times in an allowable range and to use the Dantzig-Wolfe decomposition principle (27) to solve the resultant (mixed-integer) linear programs. When cycle time is fixed, at one point in the scan, the MITROP constraints become largely uncoupled, and the constraints on the sum of offsets around closed loops provide the main coupling. Each linear program can then be broken into subprograms. In each subprogram, only a small number of the constraint columns need to be examined in main memory of the computer; the rest can be kept in bulk memory.
2. In the branch-and-bound method used to solve the mixed-integer linear program, it might be advantageous to solve the dual linear program to establish bounds. In the dual problem, the integer variables would appear in the linear objective function rather than in the constraints. In this case, when the integers are changed in the branch-and-bound process, feasible solutions of a linear program corresponding to one set of integers are also feasible solutions of a linear program with a different set of integers. This would mean, it is hoped, that fewer linear program iterations would be needed to find a new (dual problem) solution for a change of integers. Because the value of the optimal primal linear program is the same as that of the optimal dual linear program, the dual can be used to establish the bounds.

In summary, the fact that MITROP (a) allows simultaneous optimization of cycle time, splits, and offsets, (b) ensures a global optimum set of signal timings, and (c) has computer requirements comparable to other currently used signal timing methods makes it, in my opinion, a superior method for (off-line) computation of signal timings. Future use of MITROP in a variety of applications will lead to improved ease of use and to an improved understanding of its sensitivity to real data in actual traffic networks.

REFERENCES

25. H. R. Leuthardt. NO-STOP-1 Versus SIGPROG. TRB, Transportation Research Record 531, 1975, pp. 70-81.
26. R. F. Littrell. Economics of Scale in the IBM 360 and 370. Datamation, March 1974, pp. 83-88.
27. G. B. Dantzig and P. Wolfe. Decomposition Principle for Linear Programs. Operations Research, Vol. 8, No. 1, 1960.

Authors' Closure

We thank Ross for his thoughful discussion of our paper. We have further comments on two of the points that he raised.

With respect to platoon modeling, our simple formulation has worked well on the Hillier and Rothery data (18), and so we adopted it. We recognize that the model deemphasizes progression, and we have a number of approaches that would introduce more progression. One of these is particularly straightforward to implement. All the variables required to calculate bandwidth on any street are already in the computer. It would be rather easy to calculate progression bands for key arterials and introduce them into the objective function. The mixed-integer algorithm could then minimize a weighted combination of bandwidths and delays, thereby introducing progression wherever it is not dominated by delay considerations.

With respect to optimization efficiency Ross makes some worthwhile suggestions. We gave our primary research priority to the model, its fidelity, and the method of its translation into an optimization framework. We left the actual mathematical programming algorithm to standard packages. It would be interesting to explore special-purpose algorithms for the problem, although developments in general-purpose algorithms and increases in computing efficiency that resulted from hardware advances have somewhat decreased the pressure for special-purpose programs.

SIGOP II: A New Computer Program for Calculating Optimal Signal Timing Patterns

Edward B. Lieberman and James L. Woo, KLD Associates, Inc., Huntington Station, New York

This paper describes a new signal timing optimization program, SIGOP II. The optimization procedure consists of two major components: a flow model and an optimization methodology. The objective function of system disutility is expressed directly in terms of vehicle delay, stops, and excess queue length (a congestion deterrent). The flow model computes these components of disutility in the course of the optimization procedure. The optimization procedure uses the method of successive approximations within the framework of a dynamic programming methodology. Gradient techniques are applied to explore a response surface representing system disutility and to locate the minimum value. Associated with this minimum value of disutility is the optimal signal setting sought by the optimization procedure. The platoon structure of traffic and its interaction with the control at the intersection are described. Continuity of flow is preserved from one link to the next. Turning movements, lane channelization, and multiphase control are explicitly treated. Practical considerations such as signal split constraints, platoon dispersion, and the effect of short-term fluctuations in volume are included. This program, coded in FORTRAN, is currently being refined and extended. A comparison of SIGOP I and SIGOP II is given.

Within the last decade, considerable effort has been directed toward developing computer programs that provide signal timing patterns superior to those obtained through graphical techniques. Probably the most useful of these, particularly for grid networks, are

1. SIGOP (1),
2. The combination method (2), and
3. TRANSYT (3).

Although graphical techniques still represent the dominant methods used in this country, SIGOP has been used extensively in this country by consultants and by the large municipal agencies. TRANSYT has been widely applied in Europe and has been modified and extended by several agencies abroad. During the last few years, the TRANSYT model has also been used in this country. The combination method has been applied, for the most part, outside the United States. A comparison of these approaches is given by Lieberman (4).

Several evaluations of these and other methodologies have been undertaken by using both simulation (5, 6) and field studies (7). For the most part, the differences in results, expressed in terms of vehicle travel time, have been small. One notable exception is a study conducted on a portion of the urban traffic control system (UTCS) network in Washington, D.C., which indicated a clear superiority of TRANSYT signal timings over those generated by SIGOP, in terms of operational measures of effectiveness. In that study, a third optimization procedure, SOLIS, developed in the form of a crude pilot model, produced signal timings superior to those of SIGOP and generated results, via simulation, not far below those obtained with TRANSYT.

The SOLIS pilot model was developed to test certain concepts incorporated into the UTCS third-generation control policy (8). Because of the performance of SOLIS, we decided to incorporate certain features of the SIGOP and TRANSYT models and otherwise expand the scope of SOLIS. The result of this effort is the SIGOP II model, which was designed to satisfy the following objectives:

1. Develop a new, improved optimization procedure,
2. Minimize the effort and time required to effectively use the model,
3. Eliminate all artifices in the specification of the traffic environment,
4. Include a mechanism to address the problem of congestion arising from long queues that block intersections,
5. Directly treat the effect of turning movements and of turn pockets,
6. Explicitly represent multiphase control so that such signals can be properly coordinated within a control system,
7. Provide an option for evaluating existing signal timing patterns, and
8. Adopt and adapt those features in other models that have been found to be most useful.

UNDERLYING CONCEPTS

A strong interaction exists between the applied signal

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

timing patterns and the dynamic structure of traffic flow. This interaction may be described as shown in Figure 1. As shown, the signal optimization procedure is comprised, in general, of two major components:

1. A flow model to describe the traffic pattern that is compatible with the applied signal timing and
2. An optimization scheme designed to yield signal timings that service the current traffic demand with maximum efficiency.

Clearly, the effectiveness of any optimization procedure will reflect the integrity of its component parts.

FLOW MODEL

The flow model must describe the following processes:

1. Representation of traffic flow as platoons of vehicles that tend to disperse as they traverse each link in the network;
2. The change in disutility (delay, stops), including short-term fluctuations, arising from the interaction of this traffic with the control applied at the intersection;
3. The transformation experienced by the approaching flow due to its interaction with this control, expressed in terms of the volume and platoon structure of traffic departing the intersection; and
4. For multiphase control, representation of each component of flow serviced by each phase; e.g., left-turning traffic must be treated as such and not set equivalent to a somewhat arbitrary number of through vehicles, and left- and right-turn pockets should be specified as such, since their impact on traffic operations is pronounced.

Traffic flow is thus represented in the form of platoons, stratified according to turning movements. The delay and stops experienced by each platoon that encounters either a NO-GO indication at the stop line or a standing queue are computed. As vehicles discharge from a link, the resulting platoon structure is carefully represented. The maximum queue length over a signal cycle is also determined.

OPTIMIZATION PROCEDURE

The relationships between signal control parameters (cycle length, signal split, and offset) and the descriptors of urban traffic flow and geometrics were assessed at the outset of the project. The results (8) indicated that a strong interrelationship (or coupling) existed between signal split and signal offset. This coupling plays an important role in providing minimum delay service to vehicles grouped into well-defined platoons.

Both SIGOP and the combination method ignore the coupling of these two control parameters by assigning fixed values of signal split a priori and then optimizing signal offsets. The TRANSYT model, on the other hand, recognizes the importance of this coupling; it optimizes signal offsets (holding signal splits constant) and subsequently adjusts signal splits at each node. This procedure is then repeated until the performance index settles down.

In a signal timing procedure, this coupling is best represented by incorporating it directly into the formulation. A computational procedure evolved consistent with this approach and based on the method of successive approximations (MSA) used with the dynamic programming technique (9). Essentially, by using the MSA we can transform an N-dimensional problem (where N is the number of network nodes) into a sequence of N, one-dimensional unit problems, which are far simpler to solve computationally.

This unit problem is the determination of the optimal signal setting at the central node of a mininetwork, shown in Figure 2. The unit problem is

$$\text{Min } L_i = \sum_{j=1}^{J} (L_{ij} + L_{ji}) \qquad (1)$$

where

L_i = disutility for mininetwork centered at node i,
L_{ij} = disutility for link (i, j) = $W_{ij} \times M_{ij}$,
M_{ij} = $(\text{DELAY})_{ij} + \alpha(\text{STOPS})_{ij} + (\text{MQ})_{ij}$,
W_{ij} = relative importance weight assigned to link (i, j),
α = networkwide parameter to express that one vehicle stop is equivalent to α s of vehicle delay, and
$(\text{MQ})_{ij}$ = equivalent delay reflecting excessive queue length.

To solve this unit problem, the control settings at the peripheral nodes are frozen, and the disutility experienced by traffic on all links (approaches and departing links) is calculated and summed to yield L_i. Simple algebraic relations (10) yield the estimated values of delay, stops, and queue length for each link, representing a macroscopic simulation of traffic flow.

Conceptually, the disutility over a mininetwork can be computed for every admissible signal setting at the central node. These resulting values of disutilities can be represented graphically by a switching plane as shown in Figure 3. For a two-phase signal, this switching plane is defined by a horizontal axis of T_G and a vertical axis of T_R where

T_G = switching time for the onset of the green indication facing the reference link, say, Main Street (this time is referenced to a networkwide synchronous pulse emitted once per cycle) and
T_R = switching time for the onset of the red indication facing the reference link.

For each direction, values of $G_{req}^{(j)}$ can be calculated. $G_{req}^{(j)}$ is the amount of green time required to service that approach in direction j carrying the higher per-lane volume of traffic. The reference link corresponds to j = 1. When this traffic volume is lower, this value may represent the minimum green time required for pedestrian clearance or some other specified factor. Switching losses are included for ease of notation.

Because all intersections must be undersaturated for this analysis to apply, we may define the slack time S as

$$S = C - G_{req}^{(1)} - G_{req}^{(2)} \qquad (2)$$

where $S \geq 0$ is a necessary condition. All admissible signal settings can thus be represented as falling within two strips in the switching plane, inclined at 45 deg (Figure 3). By calculating L_i at all these admissible settings, it is possible to draw contours defining a disutility response surface. The objective of the unit problem solution, of course, is to locate that point on the response surface that defines minimum disutility.

Such a point is shown in Figure 3. Note that, in this case, the point lies on a boundary of the region, indicating that all the slack time is to be allocated to one direction of travel. If the signal split had been fixed, a priori, the resulting switching plane representation would have been a line (rather than a strip) inclined at 45 deg, some-

where inside the strip. In this case, the minimum point along that line would probably have corresponded to a value of disutility substantially higher than the actual value. Hence, the potential benefits of properly representing the coupling of signal split with signal offset are quite meaningful.

Figure 1. Structure of computational algorithm.

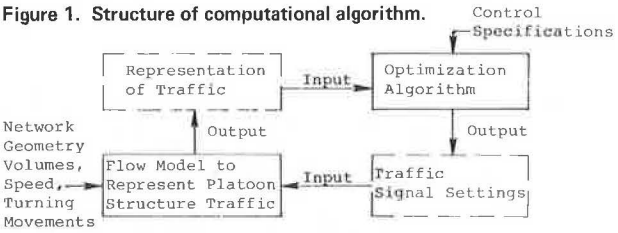


Figure 2. Mininetwork.

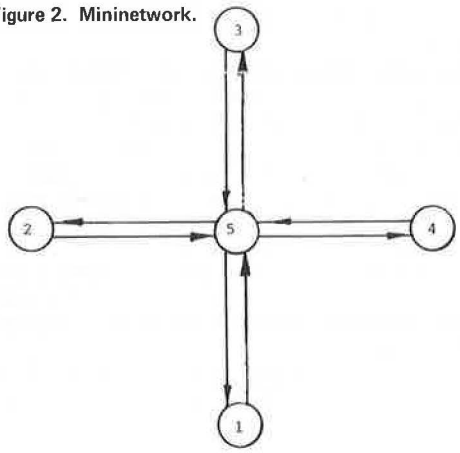
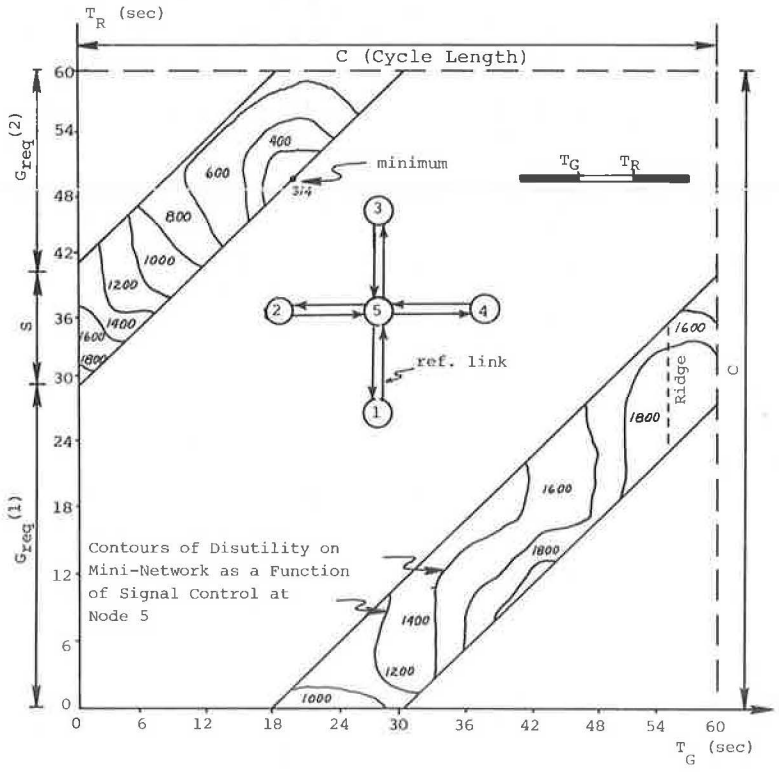


Figure 3. Switching plane.



A gradient methodology is used to locate the minimum on the response surface (10). Using this approach allows the shape of the response surface to take any form; no assumption of linearity or of a quadratic relationship of any type need be asserted. In fact, this response surface will, in general, exhibit several local minima; the gradient method is designed to locate the lowest of these minima.

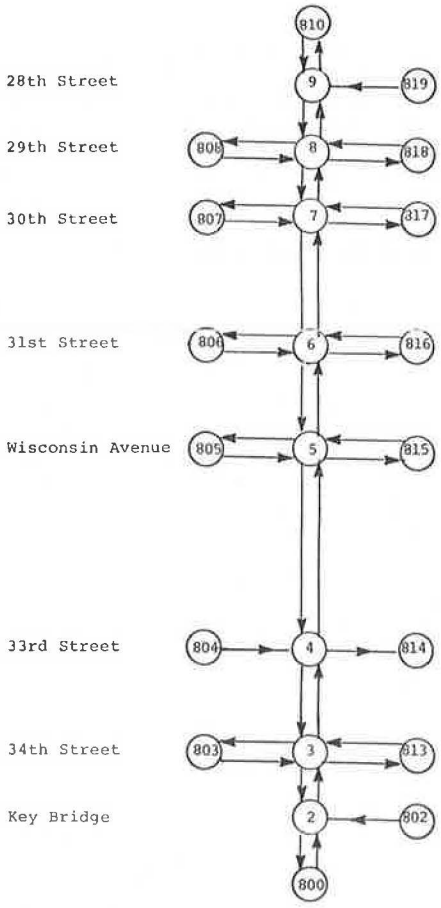
A summary of the optimization procedure follows. The following exogenous data are required:

1. Network topology,
2. Traffic volume on each link, identified as to primary and secondary flows entering the link and turning movements at downstream node,
3. Range of cycle lengths to be investigated,
4. Design speed on each link (average value),
5. Link geometry including length, number of lanes, and turn pockets,
6. Mean queue discharge headways for each network link,
7. Minimum green time at each node for each phase,
8. Several calibration parameters,
9. Signal phasing sequence and traffic movements serviced on each approach by each phase, and
10. Percentage of truck traffic.

The procedure comprises the following steps for each selection of networkwide (common) cycle length:

1. Determine the sequence of nodes along a maximal spanning tree of the network.

Figure 4. M Street arterial network.



2. Prime the control settings at all intersections to optimize traffic operations along this tree.
3. Reverse the sequence of nodes.
4. Optimize the signal control at the central node of each mininetwork by minimizing disutility. Span all network nodes in the indicated sequence by processing the associated mininetworks.
5. Calculate networkwide disutility. If zero or unchanged, the procedure is completed. Otherwise, proceed to step 6.
6. If the specified maximum number of sweeps over the network is satisfied, select best solution and stop. Otherwise, continue by returning to step 3.

Table 1. Comparison of SIGOP I and SIGOP II.

Case	SIGOP I	SIGOP II	Difference	Percent
Mean Speed				
Section 1				
Peak	9.88	20.76	10.88	110
Off peak	13.66	16.92	3.26	23.9
Section 3				
Peak	11.81	17.37	5.56	47.1
Off peak	13.28	18.75	5.47	41.2
Stops per Minute				
Section 1				
Peak	164	91	-73	-44.5
Off peak	109	73	-36	-33.0
Section 3				
Peak	1039	652	-387	-37.2
Off peak	795	527	-268	-33.7
Disutility per Hour $\times 10^{-6}$				
Section 1				
Peak	0.920	0.180	-0.740	-80.4
Off peak	0.359	0.133	-0.226	-63.0
Section 3				
Peak	2.000	0.975	-1.025	-51.3
Off peak	1.349	0.678	-0.671	-49.7

PROGRAM FEATURES

Many features have been incorporated into the program to ease the task of implementing it, to expand its scope of application, and to enhance its precision. These are briefly described below.

Memory Feature

To preserve the continuity of flow from one link to the next and to properly model the platoon structure of traffic on all links require careful modeling of the restructuring of the platoons on a link as a result of their interaction with the control at the downstream node.

Platoon Dispersion

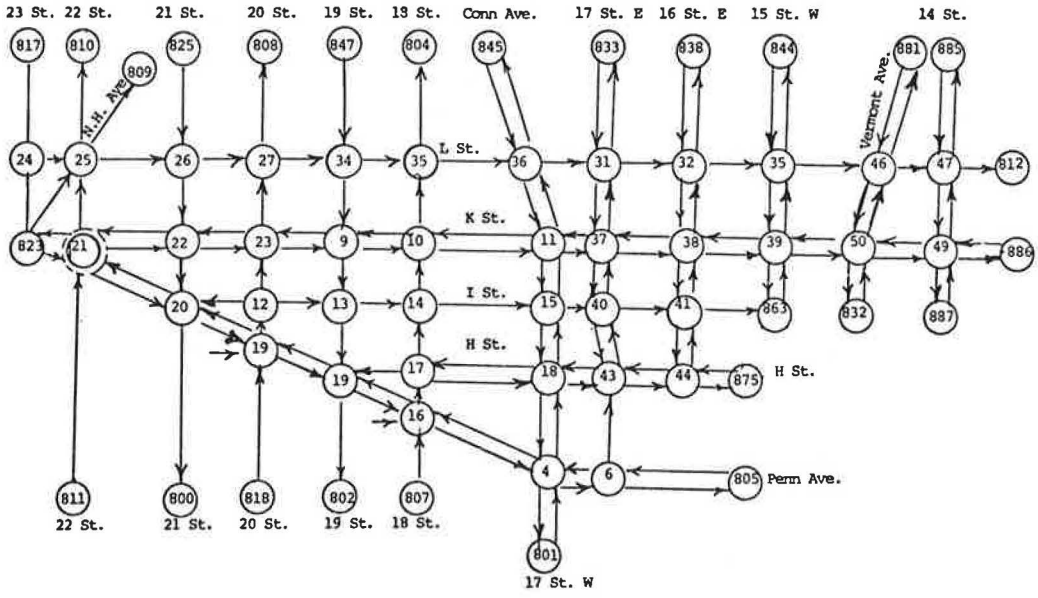
The platoon dispersion model used by TRANSYT has been adapted.

Table 2. Best solution after the indicated number of sweeps.

	5 Sweeps		10 Sweeps		15 Sweeps		20 Sweeps	
Case	Disutility[a]	Deviation (%)	Disutility[a]	Deviation (%)	Disutility[a]	Deviation (%)	Disutility[a]	Deviation (%)
Section 1								
Peak	0.237	32	0.180	0	0.180	0	0.180	0
Off peak	0.132	0	0.132	0	0.132	0	0.132	0
Section 3								
Peak	1.062	9	1.027	5	0.978	0	0.975	0
Off peak	0.809	19	0.678	0	0.678	0	0.678	0

[a]Disutility $\times 10^{-6}$.

Figure 5. Section of UTCS network.



Multiphase Control

Considerable attention has been directed to treat the interaction of multiple phases at a node and the traffic movements on each approach link serviced by each phase.

Short-Term Fluctuations in Volume

The specified link-specific volumes are treated as mean values. In the field, the cycle-by-cycle fluctuations about this mean can affect operational performance. This factor is addressed.

Double Cycling

As an option the user can request the program to halve the cycle length at those lightly loaded intersections that could benefit as a result.

Table 3. Disutility x 10^{-6}.

Case	SIGOP	SIGOP II Sweeps 1	2	3	10
Section 1					
Peak	0.920	0.299	0.259	0.259	0.180
Off peak	0.359	0.242	0.132	0.132	0.132
Section 3					
Peak	2.000	1.182	1.179	1.072	1.027
Off peak	1.349	0.991	0.991	0.911	0.678

On-Line Plotting

The user can request time-distance plots of the signal settings, as are now available in SIGOP.

Source-Sink Flows

Traffic entering or leaving a link via a parking lot or similar facility is treated.

Turn Pockets

The presence of turn pockets and their impact on control settings are included.

Input Requirements

One of the primary reasons that signal optimization programs have not attained wider acceptance in this country is the investment required to learn how to use them. This investment can be reduced if the number of input data items is minimized and their format is simplified to the extent possible. The approach taken in the design of SIGOP II is to eliminate as many input items as possible by replacing them with internal logic in a manner that does not adversely affect user flexibility. Also, the formatting structure of the TRANSYT model, which features uniform field widths, was adapted, and all input data are specified as integers.

Perhaps most important is the fact that no artificial links and nodes are permitted in representing a system

Figure 6. Networkwide signal settings as output by SIGOP II.

ITERATION NUMBER 1

NODE	PHASE	APPROACH LINKS	OFFSET (SEC)	OFFSET (PCT)	DURATION (SEC.)	SPLIT (PCT.)	CYCLE (SEC.)
2	III	(4, 2) (807, 2)	29	(32)	47	52	90
2	I	(800, 2)	76	(84)	28	31	90
2	II	(4, 2)	14	(16)	15	17	90
4	III	(2, 4) (802, 4)	38	(42)	37	41	90
4	I	(6, 4) (801, 4)	86	(96)	34	38	90
4	II	(2, 4) (802, 4) (6, 4) (801, 4)	30	(33)	8	9	90
4	IV	(2, 4) (802, 4) (6, 4) (801, 4)	75	(83)	11	12	90
6	II	(4, 6) (804, 6) (8, 6)	79	(88)	41	46	90
6	IV	(2, 6)	49	(54)	17	19	90
6	I	(4, 6) (804, 6) (8, 6)	66	(73)	13	14	90
6	III	(8, 6)	30	(33)	19	21	90
8	I	(2, 8)	26	(29)	49	54	90
8	II	(806, 8)	75	(83)	41	46	90

THE FIRST TWO PHASES ARE THE MAJOR PHASES SERVICING THE INDICATED APPROACHES

THE REMAINING PHASES (IF ANY) ARE THE MINOR PHASES

Figure 7. Networkwide minima of disutility and link-specific measures of effectiveness as output of SIGOP II.

ITERATION 1 PROVIDES MINIMUM DISUTILITY, 151440, AT A CYCLE OF 90 SECONDS

VEHICLE-MILES/HOUR= 835.5 VEHICLE-HOURS/HOUR= 72.2 MEAN SPEED=11.57 M. P. H. STOPS/MINUTE= 61

LINK FROM, TO	VOLUME (P.C.U./HR.)	EFF. SPEED (M.P.H.)	DELAY (SEC./P.C.U.)	STOPS (PER MIN.)	CAPACITY (P.C.U./HR.)	PERCENT SATURATION	MAXIMUM QUEUE
(2, 4)	1172	9.8	28.2	9.3	1120	81	10
(4, 2)	856	12.4	16.2	8.0	1920	39	3
(4, 6)	1444	10.5	34.7	21.3	2040	66	7
(6, 4)	1284	13.6	22.2	8.7	1680	60	2
(8, 6)	694	9.0	33.1	11.3	640	59	7
(2, 8)	600	21.0	3.1	0.0	1760	28	0
(2, 6)	133	10.8	35.9	2.0	200	23	3

of urban streets. The physical network maps directly into its analytical counterpart, which eliminates much confusion and artistic manipulation of network topology: What the user sees in the field he specifies as input.

Computer Requirements

SIGOP II is coded entirely in FORTRAN and both CDC and IBM program tapes will be available. Its computational speed is somewhat slower than that of SIGOP but much faster than that of TRANSYT for a comparable number of sweeps over the network. Computing time varies linearly with the number of network nodes, rather than exponentially, as is the case for the other models cited. No floating-point operations are executed; only integer data are processed.

Representative Results

The initial evaluation of SIGOP II was conducted on two major networks located in Washington, D.C. One of these is a major arterial, section 1 of the UTCS network (Figure 4), and the other is a grid network, section 3 in the central business district (Figure 5).

Data were furnished by the Federal Highway Administration and included not only the inputs to SIGOP II but also the signal timing patterns generated with the original SIGOP program (1). Two optimization periods were analyzed for each network: p.m. peak period and one off-peak period.

The evaluation option of the SIGOP II program was used to analyze the specified SIGOP signal timing pattern. Because this option uses the same flow model as that embedded within the SIGOP II optimization procedure, the two sets of results may be compared on the same basis.

Table 1 gives the statistical comparison of the two programs for the small sample of four case studies. The data in this table are for the following volume (in vehicle-km/h) and signal cycle lengths (in s):

Case	Volume	Cycle Length
Section 1		
Peak	2 348.8	70
Off peak	1 799.2	90
Section 3		
Peak	15 206.0	80
Off peak	12 423.6	70

These results are unusual in the margin of difference; field studies performed to date have yielded much smaller differences in the performance of traffic operations. Although the original version of SIGOP II produced results that were somewhat optimistic, the difference between the two sets of results is most promising.

Table 2 gives an indication of the number of network sweeps required by SIGOP II to obtain the best solution. A total of 20 sweeps was performed for all cases. After only 10 sweeps, the best solution was obtained for three of the cases and the fourth was only 5 percent removed from the best solution. Figures 6 and 7 show sample outputs of SIGOP II.

The computing time for SIGOP II is comparable to that of SIGOP, when two or three sweeps have been completed. It is instructive to compare the values of disutility on the basis of equal computing time (Table 3).

The results of SIGOP II compare favorably with those of SIGOP even after a few sweeps. This does not imply that only a very small number of sweeps should be executed. A study has been under way to establish the best trade-off between the benefits (in reduced disutility) to be expected and the required incremental investment in computing time for additional sweeps to determine the proper number of sweeps to be performed by SIGOP II.

Other enhancements are currently under development. These are designed to refine the optimization procedure and the estimates of delay and speed and to further ease the task of implementation. In particular, the current internal diagnostic tests are being expanded to provide the user with more information, particularly when a capacity problem exists. Additional features include the capability of specifying fixed offsets for selected links and fixed signal splits at selected nodes.

ACKNOWLEDGMENTS

The authors wish to acknowledge the support and constructive suggestions of H. Lum and J. MacGowan of the Federal Highway Administration. R. Goldblatt of KLD Associates and W. McShane of the Polytechnic Institute of New York also contributed to the development of the model. This work effort was accomplished under a Federal Highway Administration contract.

REFERENCES

1. SIGOP: Traffic Optimization Program, A Computer Program to Calculate Optimum Coordination in a Grid Network of Synchronized Traffic Signals. Traffic Research Corp., 1966.
2. J. A. Hillier. Glasgow's Experiment in Area Traffic Control. Traffic Engineering and Control, Vol. 7, No. 8, 1965-66.
3. D. I. Robertson. TRANSYT: A Traffic Network Study Tool. U.K. Road Research Laboratory, Crowthorne, Berkshire, England, Rept. LR 253, 1969.
4. E. B. Lieberman. Algorithms for Computer Traffic Control—Current State of the Art. Paper presented at Workshop on Computer Traffic Control, Spring Hill, Minn., March 20, 1974.
5. F. A. Wagner, F. C. Barnes, and D. L. Gerlough. Improved Criteria for Traffic Signal Systems in Urban Networks. NCHRP, Rept. 124, 1971.
6. Network Flow Simulation for Urban Traffic Control System—Phase II: UTCS-1 Network Simulation Model, Final Report. Peat, Marwick, Mitchell and Co. and KLD Associates, Inc., June 1973.
7. J. A. Kaplan and L. D. Powers. Results of SIGOP-TRANSYT Comparison Studies. Traffic Engineering, Vol. 43, No. 12, Sept. 1973.
8. Variable Cycle Signal Timing Program, Volume 3. KLD Associates, Inc., Final Rept., May 1974.
9. R. E. Bellman and S. E. Dreyfus. Applied Dynamic Programming. Princeton Univ. Press, 1962.
10. E. B. Lieberman and J. Woo. SIGOP II: Program to Calculate Optimal, Cycle-Based Traffic Signal Timing Patterns. KLD Associates, Inc., Rept. TR-29, Dec. 1974.

Multilane Traffic Flow Process: Evaluation of Queuing and Lane-Changing Patterns

Jens Rørbech, Laboratory of Road Data Processing, Road Department, Denmark Ministry of Public Works

This paper presents a queue-theoretical model for describing traffic flow on a dual two-lane motorway. The model is based on the theory of stationary Markov processes and is closely related to data collection methods made possible by specially developed recording equipment. The model is based on a description of the driver's behavior in traffic. The individual driver alternates between driving under free-flow conditions (state F) at his or her desired speed and without delay and driving in a queue (state K) with consequent delay. To escape from or possibly entirely avoid a queue, a driver must change lanes (alternate between lane 1 and lane 2). The behavior of the driver is thus described by a variable that assumes the discrete values 1F, 1K, 2F, and 2K. This process can be described as a Markov process. The reactions of individual drivers are summed up in a comprehensive description of the average driver's behavior, so that the road traffic model developed is an example of a macrostochastic model. A technique has been developed by which the parameters in the Markov process can be estimated from the collected data, and the paper describes how the model becomes part of a detailed investigation into the vital conditions affecting motorway traffic, e.g., capacity, relationship between speed and traffic volume, and the drivers' use of the motorway.

This paper presents a method for describing the traffic flow on a four-lane motorway and for quantitatively evaluating factors that have a bearing on the traffic flow. A more comprehensive description of the work on the traffic model has been published elsewhere (1). Traffic flow is described as a stationary Markov process.

Breiman (2) discussed some of the road traffic models prepared so far: microscopic, macroscopic, and stochastic models. Microscopic and macroscopic models are deterministic; i.e., they usually compare the traffic with mechanical and physical processes and are in fact unsuitable for describing the traffic situations that are of interest in practice. This is because the traffic situation on a motorway under normal conditions is a result of the correlation among a great number of more or less casual conditions that cannot so easily be represented by means of mechanical or physical analogies.

Stochastic models, however, are of assistance in the construction of traffic engineering models. It is, after all, an intrinsic feature of any traffic flow that it must, through a limited application of service resources, meet manifold demands of individual units with respect to service time, method, speed, and the like—conditions that can often be fairly well described by means of relatively simple stochastic processes. In particular, the theory of queues is, after all, concentrated on the construction of models for queue systems where basic conditions such as the probability of rejection or delay, the number of units in the queue, and the average delay can be calculated provided that the queue system can be broken down into clearly specified functions. The first works concerned with the theory of queues were in fact undertaken to solve purely practical problems in connection with the dimensions to be adopted for telephone installations.

Despite this apparently obvious aid, there are few road traffic models based on stochastic processes, and, with few exceptions, discussion has been confined to descriptions of partial problems in traffic flow. One of the reasons is that road traffic cannot so easily, e.g., as telephone traffic, be subordinated to the specified functions of the conventional theory of queues. For instance, it can rarely be assumed that the arrivals of cars are governed by the Poisson distribution, i.e., that they are independent of each other; as a rule, queues can be observed even when traffic volumes are light, so that all the problems associated with overtaking come into play.

The present road traffic model is therefore based on a description of drivers' behavior in traffic.

THE MODEL

The following description of motorists' behavior on a motorway can be used purely logically to classify a series of conditions affecting the traffic flow process and as a model-technical basis for a data collection technique. It is assumed that all motorists can be divided into two types representing extremes in driving behavior.

The first type is drivers who keep to the right as much as possible. As long as there is room in the right lane, they travel there at the desired speed. If they catch up with a slower driver, they overtake him or her if there is room in the left lane; otherwise, they wait in the queue. As soon as they have completed an overtaking maneuver,

Publication of this paper sponsored by Committee on Highway Capacity and Quality of Service and Committee on Traffic Flow Theory and Characteristics.

they return to the right lane again.

The second type is those drivers who have only one thing in mind: fast driving in the left lane. They force their way by flashing their lights and sounding their horns. Occasionally, they get into a queue, but they stay there only until they have forced the driver in front of them to move to the right. They themselves may also be forced to the right if an even faster driver of their own kind appears in the rear-view mirror; that is what they really respect.

Drivers in both groups alternate between traveling under free conditions at the desired speed and without loss of time (state F) and traveling under queue conditions with loss of time (state K). They have to change lanes (alternation between traveling in lane 1 and traveling in lane 2) to escape from or possibly completely avoid a queue. In other words, the behavior of the motorists on the road is described by a variable that may adopt the discrete values 1F, 1K, 2F, and 2K. This process may reasonably be described as a stationary Markov process.

Figure 1 shows the alternation of the motorists between the four states 1F, 1K, 2F, and 2K. It appears that both models of driving described above (both those who keep to the right and those who keep to the left) lead to the same pattern of change between the four states because of the way in which the queues are formed and dissolved.

The parameters in the Markov process that are registered during the recordings are partly the probability of remaining in the four states and partly the transition probabilities between the different states. The data collection technique and the estimation of the parameters in the Markov process are described below.

THE RECORDINGS

The values of the parameters in the road traffic model were determined with the aid of the recording arrangement shown in Figure 2. One carriageway of the motorway was fitted with eight detectors (tape switches) so that each car was recorded twice. The passage times of all automobiles were recorded separately on a digital magnetic tape recorder. A description of this very special equipment is given elsewhere (1, appendix).

Among the data recorded at each passage were vehicle, position, arrival time, speed, and wheelbase and the queue situation (i.e., whether the car was in a free-flow situation or in a queue situation). Computer programs were used to analyze the recorded passage times. The question of defining the condition of driving in a queue is dealt with elsewhere (1, appendix).

For each passage, the following data are therefore recorded twice (i.e., at cross section A and again at cross section B):

1. Position in the carriageway, lane 1, lane 2, lane change from 2 to 1, and lane change from 1 to 2; and
2. Queue situation, free flow or queue.

With the determination of the wheelbase, the different cars can again be identified at cross section B (the degree of uncertainty with which the wheelbase is determined is much less than the natural spread of wheelbase sizes). In other words, it is possible to record not only the distribution of the cars over the four state variants but also the rate at which they change from one state to another.

An example of the alternation of 800 cars between the different states is shown below, corresponding to the midday traffic in Table 1 described later. The traffic volume is 1081 cars/h and the distance L between cross sections A and B is 100 m (328 ft).

State	1F	1K	2F	2K	Total
1F	244	20	7	2	273
1K	17	50	2	2	71
2F	15	1	294	23	333
2K	1	0	14	108	123
Total	277	71	317	135	800

The Markov property has been tested by recording the traffic at three cross sections A, B, and C and by determining whether a stochastic independence can be assumed between the observations at cross sections A and C while the value at cross section B remains constant. The stationary character has been similarly tested by comparing the observed distributions of the states of the cars at the three cross sections.

Three different traffic volumes are considered in these tests: a situation of light traffic, called morning traffic; a situation of medium-heavy traffic, called midday traffic; and a situation of heavy traffic, called evening traffic. In all three situations, it can reasonably be assumed that the behavior of the drivers on the road can be described by the stationary Markov process shown in Figure 1.

Because of the stationary character of the Markov process, the observations can be confined to a single cross section and can then be generalized to apply to a road section of theoretically indefinite length.

ESTIMATING PARAMETERS IN THE MARKOV MODEL

The recording method described above yields, as a result of the observations, a transition matrix that indicates the states of the individual drivers at the two extreme cross sections of the recording section.

The parameters that determine the actual Markov process are the state probabilities P'_i and the transition intensities Q''_{ij} where P'_i indicates the probability of a car being in state i at a given moment and $Q''_{ij}\,dl$ $(i \neq j)$ indicates the probability of a transition from i to j over the short distance dl. The transition probabilities $P''_{ij}(L)$ are of interest. $P''_{ij}(L)$ is the probability of a transition from i to j over the distance L.

Between the parameters there are the following relations

$$P' = P' P''_{ij} \qquad \sum_i P'_i = 1 \tag{1}$$

from which the state probability P′ can be found when the transition probability P''_{ij} is known. In the same way the state probability P′ can also be found from

$$P' Q'' = O' \qquad \sum_i P'_i = 1 \tag{2}$$

when Q″ is known.

Between the transition intensities and the transition probabilities, the following correlation exists:

$$P'' = \exp(Q'' L) \tag{3}$$

Equations 1, 2, and 3 are valid provided that statistical equilibrium has occurred.

In the estimation of the parameters of the Markov model on the basis of the observations, two considerations in particular become decisive.

1. The sums of lines and columns in the observation matrix will hardly be identical as is demanded if sta-

Table 1. Results of observations on Elsinore Road, 1969.

Traffic	Volume (cars/h)	1F	1K	2F	2K	Number of Lane Changes
Morning	524	30 400 m, ~33.0 stretches of 920 m	1100 m, ~7.4 stretches of 154 m	59 000 m, ~45.2 stretches of 1305 m	9500 m, ~21.3 stretches of 448 m	51.1
Midday	1081	33 500 m, ~50.9 stretches of 658 m	8700 m, ~31.9 stretches of 273 m	41 200 m, ~48.4 stretches of 853 m	16 600 m, ~30.4 stretches of 545 m	37.9
Evening	2736	20 300 m, ~101.9 stretches of 199 m	39 000 m, ~93.7 stretches of 416 m	15 400 m, ~48.2 stretches of 318 m	25 300 m, ~41.3 stretches of 612 m	16.4

Note: 1 m = 3.28 ft; 1 km = 0.62 mile.

Figure 1. The alternation of the motorists between the four states.

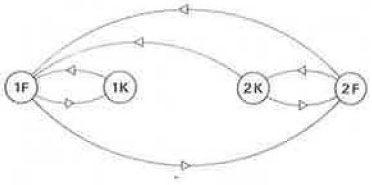


Figure 2. Recording arrangement for determining the values of the parameters in the road traffic mode.

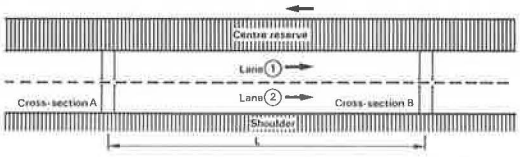


Figure 3. Distribution of drivers over the four state possibilities as a function of the traffic volume and the level-of-service specifications given in the Highway Capacity Manual.

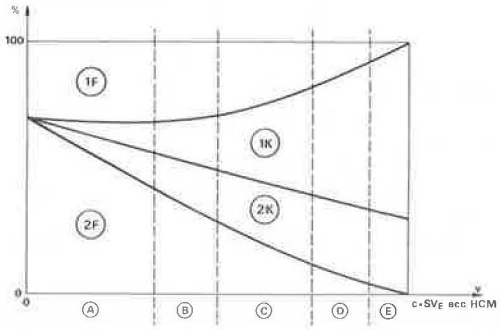
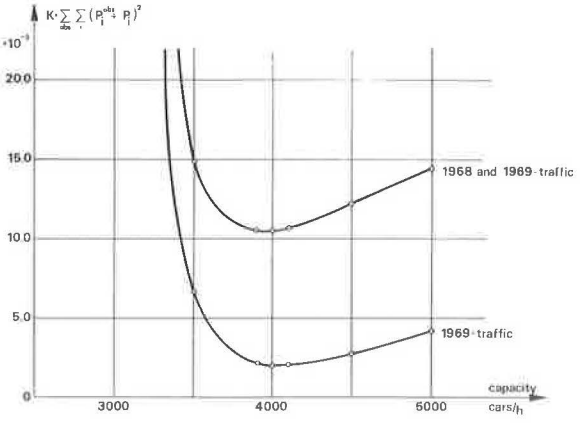


Figure 4. The minimized sum of the squares K as a function of the road capacity c.



tistical equilibrium is assumed.

2. The recording method described earlier merely indicates the states of the individual cars at the extreme points of the recording section whereas the continuous Markov model permits a theoretically infinite number of transitions between the recording points; i.e., the recording method aims more directly at determining the transition probabilities P''_{ij} than the transition intensities Q''_{ij}.

In connection with the work on the road traffic model, an estimation technique has therefore been worked out in which these problems are taken into consideration. The estimates are carried out in a series of stages (1).

1. The transition probabilities that meet the demand for statistical equilibrium are estimated on the basis of a maximum likelihood assumption. This P_{ij} matrix meets the requirement of identical sums of lines and columns (statistical equilibrium has been reached) and forms the basis for the calculation of a set of state probabilities and a set of transition intensities. The latter do not, however, meet the requirement of the transition diagram inasmuch as five of the transitions are not permissible (Figure 1).

2. While the state probabilities generated in stage 1 above are kept constant, the prohibited transitions (defined as the transition intensities generated in stage 1) are shifted in the transition diagram in such a way that the transitions now follow permissible sequences. For instance, the prohibited transition 1K → 2K is now shifted to 1K → 1F → 2F → 2K. Because the process is, after all, a continuous Markov process, it is quite permissible to assume that a motorist changes his or her state repeatedly over the observation section. This gives rise to a new set of transition intensities and therefore also to a new set of transition probabilities. The state probabilities from stage 1 remain unchanged and are kept constant during the entire phase.

3. The estimated parameter values determine a new transition matrix, and a good check is obtained by comparing the observed and the estimated transition matrices. An observation matrix has been represented. The corresponding matrix generated from the estimated parameter values is

State	1F	1K	2F	2K	Total
1F	234.0	19.9	13.4	0.5	267.8
1K	19.9	49.3	0.6	0.0	69.8
2F	12.7	0.6	295.6	21.0	329.9
2K	1.2	0.0	20.3	111.0	132.5
Total	267.8	69.8	329.9	132.5	800.0

The estimated parameters describe a traffic situation inasmuch as the state probabilities determine the average total distance traveled in the individual states, whereas the transition intensities determine the length of time during which the drivers remain in the specific states.

Table 1 gives a summary of the results from record-

ings during situations of light, medium-heavy, and heavy traffic volumes. The recordings date back to 1969 when there were no general speed limits in Denmark and come from an old motorway, Elsinore Road, north of Copenhagen. The tabulations indicate the average conditions encountered by the motorists. During morning traffic, for example, motorists will on the average cover 30 percent of the distance in state 1F, 1 percent of the distance in state 1K, 59 percent in state 2F, and 10 percent in state 2K. A motorist traveling freely in lane 1 will, on the average, be able to travel about 900 m before changing to another state.

CONCEPTS OF CAPACITY AND LEVEL OF SERVICE

Apart from this general description of the traffic situations on Elsinore Road, the road traffic model has been used for discussing in greater detail various vital conditions. The concepts of capacity and level of service and the correlation between speed and traffic volume are discussed below.

If the driver classification in the suggested road traffic model is applied to the approach chosen in the Highway Capacity Manual (3), namely, that the traffic flow on an open section of road for $0 < v/c < 1.0$ will take place at levels of service ranging from A to E, the distribution of the drivers over the four state variants can be illustrated as shown in Figure 3.

It is particularly important to determine the boundary conditions. If the different state probabilities are called P_i (i = 1F, 1K, 2F, and 2K), one arrives at the following results: If the volume-capacity ratio is approximately equal to zero, i.e., level of service A, queue conditions will hardly occur except over short distances; i.e.,

$$P_{1F} + P_{2F} \cong 1.0 \qquad P_{1K} \cong P_{2K} \cong 0 \tag{4}$$

The distribution of the drivers over the two lanes will be governed by, among other things, driving habits, the distribution over the driving models mentioned earlier, and the distribution of the desired speeds. If $v/c \sim 1.0$, i.e., level of service E, all the cars will travel under queue conditions. There will hardly be any lane changing, for it is not possible to escape, in this way, from the queue situation:

$$P_{1F} \cong P_{2F} \cong 0.0 \qquad P_{1K} + P_{2K} \cong 1.0 \tag{5}$$

In this situation, too, the distribution of the drivers over the two lanes will be governed by driving habits. A decisive factor will be the way in which the entire saturation process takes place, again depending on, among other things, the distribution over the driving models, the distribution of the desired speeds, and the risks that the drivers are willing to take to approach their desired speeds as closely as possible.

An estimate of the positions of the boundary lines shown in Figure 3 for the four state variants has been made on the basis of the observations on the Elsinore Road by means of a regression analysis; the structure is apparent from Figure 3. Broadly speaking the drivers use the motorway in one way during light traffic when the right lane (lane 2) handles the greater part of the traffic volume and in another way during heavy traffic when the left lane (lane 1) handles the greater part of the traffic volume. An evaluation of this use of the motorway is given later; it is shown that the distribution is connected with the queue formation pattern and consequently the transition diagram shown in Figure 1.

Fixing of the lines in Figure 3 implies, as the problem is formulated, that the capacity of the road is known beforehand. The regression analyses have been carried out with capacity ranging from 3000 to 5000 cars/h (for two lanes in one direction). In other words, the capacity of the road, c, has been included as a variable in the calculations, which permits an estimate of this vital quantity.

Figure 4 shows the minimized sum K of the squares as a function of the variable capacity c for two series of observations. The best possible agreement between the model and the observed data is obtained when c = 4000 cars/h (for two lanes in one direction), which is the value the Highway Capacity Manual recommends for a motorway like the Elsinore Road.

SPEED-VOLUME RELATIONSHIP

The speed-volume relationship is determined in three stages. All these are time-mean speed relations because of the way in which the observations are arranged. The state probabilities are, after all, measured at one cross section, which means that P_i are time probabilities [the concepts of space-mean speed and time-mean speed as related to the present road traffic model are discussed in greater detail elsewhere (1).]

First, the distribution of the drivers' free speeds is estimated. Thus the mean speed V_f is determined, which represents the left end of the speed-volume curve and the point of intersection with the $v/c = 0$ axis.

The frequency distribution of the free speeds is shown in Figure 5. The speed distribution has been determined as follows: Observations were made during a 2½-h period at a time when the traffic flow on the Elsinore Road was particularly light; 250 cars in lane 1 and 250 cars in lane 2 were selected that traveled at distances greater than 300 m (1000 ft) behind the car in front. The speed distributions corresponding to the two lanes were determined separately and then added up to show, on the basis of the average distribution over the two lanes, the total frequency distribution of the speeds.

In practice it is hardly possible to determine a distribution function of the free speeds on any very reliable basis inasmuch as the selection of the cars traveling under free-flow conditions will always be subject to uncertainty. It is, for example, rather likely that cars of a certain high-speed class will be the first to be caught in a queue and will therefore not frequently be represented as driving under free-flow conditions.

Figure 4 shows the best estimate of the motorists' desired speeds, which have been checked in various ways. It is, for instance, apparent from Figure 5 that there is a logically correct correlation between the desired and observed speeds in the three traffic situations. It can also be shown that those motorists who, in the three traffic situations, travel under free-flow conditions can be isolated as components of that group of drivers that are determined by the distribution of the free speeds.

In the second stage, the speed conditions of the drivers in the capacity situation are evaluated and the mean speed V_m is determined, which represents the right end of the speed-volume curve and the point of intersection with $v/c = 1.0$.

The speed in the capacity situation is determined on the basis of observations of cars traveling under queue conditions in the two lanes. Figure 6 shows the observed time intervals for cars traveling under queue conditions, plotted in the form of speed-volume functions.

As with the free speeds, V_m cannot be determined unequivocally, but a reasonable estimate of V_m can be made on the basis of the observations shown in Figure 6.

Because driving speed in the capacity situation could not be accurately estimated, the final determination of

Figure 5. Distribution of free speeds and those based on observations on the Elsinore Road.

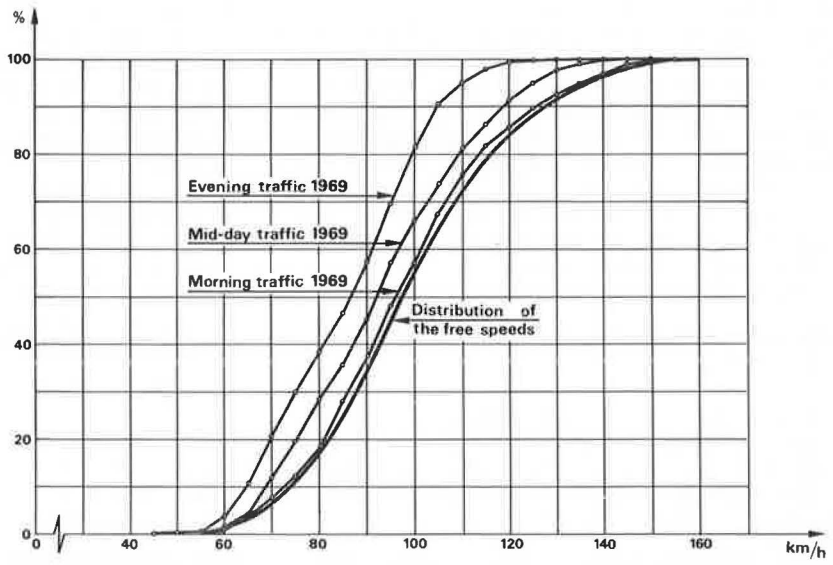
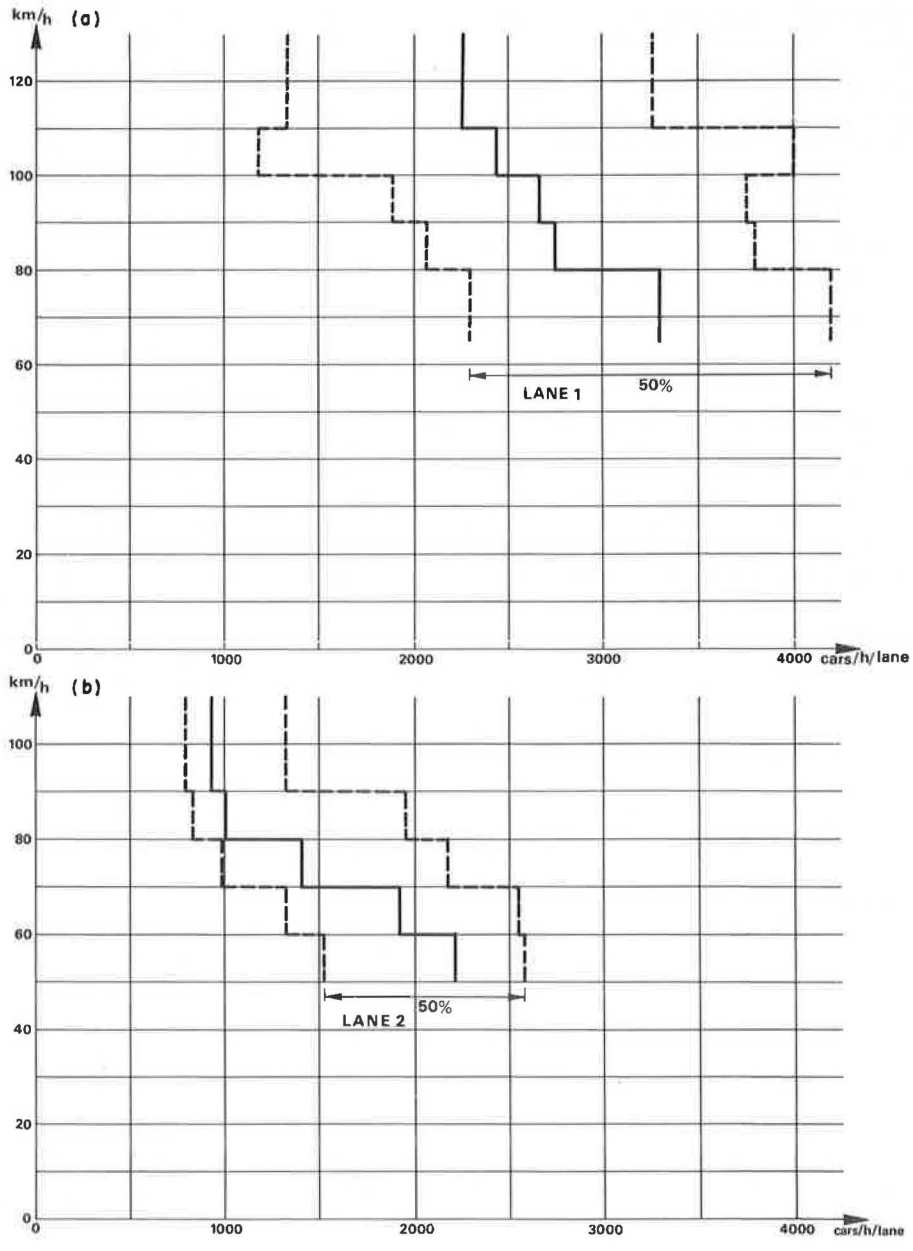


Figure 6. Observed time intervals for cars traveling under queue conditions in (a) lane 1 and (b) lane 2.



the speed-volume curve allows for three alternative values of V_s as shown in Figure 7. The three values were based on the observations of queue conditions in the two lanes (speed distributions for 1K and 2K, Figure 6) and on alternative assumptions concerning the degree to which the drivers will, in the capacity situation, adopt a rational distribution over the two lanes in accordance with their desired speeds, i.e., with high-speed drivers preferring lane 1 and low-speed drivers lane 2.

In the third stage, the shape of the curve in the interval $0 < v/c < 1.0$ is determined. This is done by estimating, on the basis of the recorded probabilities P_i (i = 1F, 1K, 2F, 2K) and the probabilities of transitions between these states, the probability of the drivers being caught in a queue in the two lanes. The result of the estimate is shown in Figure 7.

The shape of the curve in the interval $0 < v/c < 1.0$ is determined by considering the conditions for an average driver who wants to travel at a speed equal to the mean speed of the drivers' desired speeds V_f. As long as the average driver is traveling in the 1F or 2F state, the speed is V_f and no time is lost. In situations 1K and 2K, the driver is traveling at a speed lower than V_f and therefore loses time.

The time-mean speed as calculated on the basis of the behavior of the average driver then becomes

$$V_T = P_{1F} V_f + P_{1K} V'_{1K} + P_{2F} V_f + P_{2K} V'_{2K} \quad (6)$$

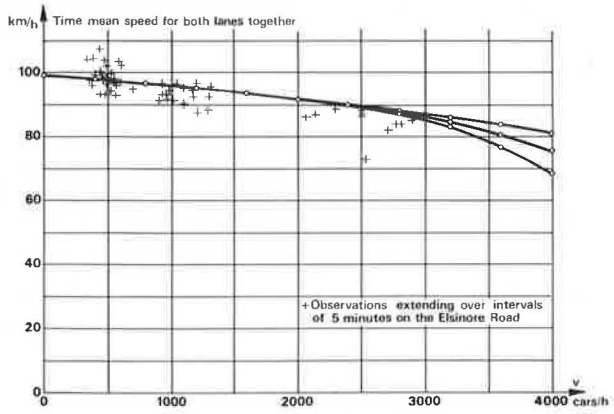
where V'_{1K} and V'_{2K} = speed of the average driver in the two queue situations.

Rørbech (1) describes how the speed in the two queue situations as a function of the traffic volume is estimated from the probability of a driver traveling at free speed being caught in a queue, i.e., on the basis of the observed transition intensities q_{FK}. During light traffic the transition intensities q_{FK} are small inasmuch as the risk of being caught in a queue is low. During the whole saturation process the probability of traveling under queue conditions is increased as $q_{FK} \to \infty$, when $v/c \to 1.0$.

The speed-volume relationship shown in Figure 7 has the form of a steadily decreasing function. The alternative values for the speed in the capacity situation merely affect the last part of the speed curve. Figure 7 shows a number of 5-min observations, which cover a much longer observation period than those that determine the parameters in the Markov process. That is, the data that determine the speed-volume relationship are in principle independent of the plotted speed observations.

The computed curve for the range beyond the very low traffic volumes lies at the upper edge of the observed speeds, but the difference hardly exceeds 5 km/h (3.1 mph) on average.

Figure 7. Speed as a function of the traffic volume.



DRIVERS' USE OF THE MOTORWAY

The transition diagram of Figure 1 can be plotted identically for both of the two ways of driving mentioned earlier (i.e., both the motorists who keep to the right and those who keep to the left) because both ways of driving lead to the same queue formation pattern. A strictly logical consequence of the two behavioral models will therefore take the form of the following queue formation patterns in the two lanes. In the slow lane, an inequitable queue formation will take place in which the driver who arrives first in the queue is the one who will be served last. All the traffic gaps in the left lane will be snatched by the later arrivals in the queue. In the fast lane, an equitable queue formation will occur in which the driver who arrives first in the queue will also be the one who gets out of it first when the slow car in front moves to the right lane.

The difference in queue formation patterns in the two lanes has a decisive influence on the whole use of the road. In light traffic most motorists travel in the right lane. Only about one-quarter of them travel at high speed in the left lane, and they can easily make their way. On the other hand, many drivers, even in light traffic, will travel under queue conditions in the right lane. The reason is that, if they try to make an overtaking maneuver, it will not be long before a fast driver will urge them into the right lane again by flashing his lights; in fact, the road is not used very efficiently in light traffic conditions. At any rate this was the case before the speed restrictions were introduced in Denmark.

In heavy traffic the pattern becomes different. Those who want to drive fast find it more difficult to proceed, and the many drivers who want to travel at medium speeds (i.e., 80 to 100 km/h or 50 to 60 mph) take this opportunity, possibly too eagerly, to move out into the left lane. In heavy traffic almost three-quarters of the motorists travel in the left lane, obviously because of the higher traveling speed. The queue pattern in the right lane is a strong incentive: If one is caught in a queue here, it is almost impossible to get out of it again. Because of the inequitable queue formation pattern, those

Table 2. Effect of a speed restriction on the traffic flow process.

Condition	Observed Mean Speed (km/h)	110-km/h Limit: New Mean Speed (km/h)	110-km/h Limit: Cars Disturbed (%)	120-km/h Limit: New Mean Speed (km/h)	120-km/h Limit: Cars Disturbed (%)	130-km/h Limit: New Mean Speed (km/h)	130-km/h Limit: Cars Disturbed (%)	140-km/h Limit: New Mean Speed (km/h)	140-km/h Limit: Cars Disturbed (%)
Motorists' desired speed	99.4	95.3	28	97.4	16	98.6	8	99.2	3
Morning traffic	97.9	94.1	25	96.0	14	97.1	8	97.7	3
Midday traffic	92.9	90.8	19	92.2	9	92.7	3	92.9	0
Evening traffic	85.6	85.3	5	85.5	0	85.5	0	85.6	0

Note: 1 km/h = 0.62 mph.

cars that join the queue later, i.e., those that travel behind, will snatch all the gaps in the fast lane. This is because the gaps in the fast lane are first available to the cars in the rear of the slow-lane queue.

In themselves these statements are very obvious but, in fact, it has never before been realized that the result was such an uneven, and consequently varying, distribution of the traffic in the two lanes, i.e., most drivers in the right lane during light traffic and most drivers in the left lane during heavy traffic.

One cannot help wondering whether this is an appropriate way of driving. Is the motorway as a technical installation utilized to a sufficient degree? A detailed analysis shows that, in heavy traffic, the uneven distribution leads to a very nearly optimal utilization of the road. It is, in fact, worthwhile for as many as 75 percent of the drivers trying to travel in the fast lane at speeds ranging from 70 to 80 km/h (45 to 50 mph).

It has already been mentioned that, in light traffic, the uneven distribution of the drivers over the two lanes does not represent proper use of the road. The few drivers wanting to travel fast may be able to dominate too much. In this respect, the speed limits recently introduced in Denmark on an experimental basis may have an important bearing. Apart from their generally beneficial influence, these speed limits are also likely to result in an increase in mean speed. This is studied in greater detail below.

Table 2 gives the recorded mean speeds corresponding to the three different situations on the Elsinore Road from Table 1, supplemented by an estimate of the average speed at which the motorists would have wanted to travel if there had been no disturbances from other motorists. As mentioned, all four situations are related to a situation without speed limits.

The table also gives the mean speeds that would result if all the drivers, either voluntarily or under compulsion, obeyed an upper speed limit of 110, 120, 130, or 140 km/h (68, 75, 81, or 88 mph). They were calculated simply by assigning all the speeds above the speed limit to that upper speed limit. As might have been expected, an upper speed limit will have virtually no effect on the mean speed in the case of heavy traffic volumes. More surprising is the likewise almost insignificant impact that the speed limits have on the mean speeds at lower traffic volumes.

A speed limit of 120 km/h (75 mph), possibly marked as 110 km/h (68 mph), i.e., corresponding to the present situation in Denmark, should according to Table 2 cause the mean speed to drop at most by 2 km/h (1.2 mph). Circumstances, however, indicate that the mean speed would not decrease at all but would possibly even increase. In the morning traffic situation it appears that almost all motorists who drive at speeds faster than 120 km/h (75 mph) travel in lane 1, the left lane. A 120-km/h speed limit would have the following effect on the mean speeds of the two lanes:

1. Lane 1, reduction from 116.3 to 110.0 km/h (72.2 to 68.3 mph) and
2. Lane 2, reduction from 89.9 to 89.5 km/h (55.8 to 54.6 mph).

In other words, the effect is confined to the speed conditions in lane 1. However, by removing all speeds higher than 120 km/h in lane 1, it becomes easier for drivers in lane 2 to make use of the traffic gaps in lane 1 for overtaking purposes. If only a few of the drivers in lane 2 would find it easier to make use of lane 1 for overtaking purposes, a speed limit of 120 km/h would hardly cause any decrease in the overall mean speed, all other factors being equal.

A speed limit would thus tend to bring about some leveling out of the speed benefits, with the added advantage of a possible increase in mean speed. The drivers who in a situation without speed limits are particularly prone to get into trouble are those whose desired speeds are around 90 to 100 km/h (56 to 62 mph). These drivers can easily be caught in a queue in lane 2 but may have difficulties in carrying out an overtaking maneuver in lane 1 because of the necessary acceleration to very high speeds. In practical terms, a speed limit would thus assist the drivers wishing to travel at 90 to 100 km/h in changing from the queue speed of 75 to 85 km/h (47 to 53 mph) to the desired speed at the expense of reducing to 120 km/h (75 mph) the speed of drivers wishing to travel at 130 to 140 km/h (81 to 88 mph). If the traffic flow also contains heavy vehicles, travel speed under queue conditions in lane 2 may become lower still, even with low traffic volumes.

This argument for lowering the very high speeds on the Elsinore Road has been based exclusively on considerations for the operational use of the road and has been confined to conditions at low traffic volumes. In a general assessment of the drivers' use of the road, however, it is also necessary to take another aspect into account, namely, the uncomfortably strenuous traffic environment, which can easily be obtained if the traffic is handled at an unduly high speed level, that is, both the high speeds at low traffic volumes and a lack of adaptation of speed and traffic volume at high traffic volumes.

CONCLUSION

The road traffic model presented for describing the traffic flow on multilane roads can be briefly summarized and characterized as follows: On the basis of a description of the action pattern of individual drivers, the average driver's freedom of movement under given external conditions is quantified in a Markov process by using the measurable variables of state probabilities and transition intensities. Inasmuch as there are correlations between these new variables and the conventional road traffic parameters, i.e., traffic volume and speed, factors that have a bearing on the traffic flow can be quantitatively assessed.

A concluding assessment, with suggestions for possible applications of the road traffic model, may naturally be based on a critical evaluation of the model technique that has been used. Fundamentally, the road traffic model represents a method of interpreting the results of observations; it therefore represents a method rather than a model and, as already emphasized, it permits a quantitative description or at least a quantitative comparison of different highway and traffic engineering situations to the extent to which these situations are reflected in the behavior of the drivers.

This paper has mainly concentrated on a traffic engineering application of the model by showing how the traffic flow is affected by changes in the traffic volume. In a broader context, the essential aspect in this connection is the use of the road facility by the drivers since any discussion about the degree of use and possibly about a traffic overload stands and falls after all with the methods designed to indicate whether the drivers' behavior is appropriate. It is thus shown how the distribution of motorists over the two lanes and their eagerness to change lanes and to achieve a more suitable distribution over the two lanes are decisive factors affecting road use.

Among direct applications of the model, mention may also be made of a number of tasks of a highly practical nature. The Highway Capacity Manual does not for example distinguish between calculations for bridges and tunnels and for normal sections of road. Investigations

carried out on an old bridge across the Little Belt in Denmark (4) seem to show that a road across a bridge is governed by quite special conditions. The bridge has only a 5.6-m-wide (18-ft) carriageway and a maximum observed traffic flow of 2718 cars/h. The road traffic model here proposed can be used in an analysis of the behavior of drivers on this type of highway facility with a view to a formulation of capacity and service level data.

Another traffic engineering application of the model might be associated with an evaluation of the effect of heavy vehicles on the traffic flow, e.g., whether it is reasonable to count heavy vehicles in terms of passenger car units, whether the effect of heavy vehicles on the remaining traffic varies with the traffic volume and the service level, the effect of up and down gradients, and so on. These are problems that, apart from the calculations more concerned with traffic engineering aspects, also have a bearing on the geometric design of a road, e.g., criteria for the design of climbing lanes or third running lanes.

The application of the road traffic model to the assessment of aspects more concerned with road geometry has thus been mentioned. By way of example, an evaluation of the traffic flow on the Elsinore Road during hours of darkness has been discussed (1). State probabilities and transition intensities have been recorded after dark on the same day as the morning, midday, and evening traffic situations and have been compared with the values that might be expected under daylight conditions. On the basis of the comparisons it is possible to indicate a quantitative expression of the effect that the new conditions offer to the traffic flow. Such an examination might be used to determine the expediency of, for example, installing road lighting.

REFERENCES

1. J. Rørbech. The Multilane Traffic Flow Process: An Evaluation of Queueing and Lane-Changing Patterns, Based on a Markov Model. Road Department, Denmark Ministry of Public Works, Copenhagen, 1974.
2. L. Breiman. Data and Models in Homogeneous One-Way Traffic Flow. Transportation Research, Vol. 3, 1969, pp. 235-249.
3. Highway Capacity Manual—1965. HRB, Special Rept. 87, 1965.
4. J. Rørbech. Capacity and Level of Service of Danish Highways. Traffic Engineering and Control, Vol. 10, No. 9, Jan. 1969, pp. 462-469.

Discussion

A. G. R. Bullen, University of Pittsburgh

Røhrbech's paper provides a refreshing approach to traffic flow theory by making a conscious and effective effort to bridge the gap between theory and traffic engineering practice. This gap is of long standing. On the one hand, theoreticians have been most reluctant to grasp some of the realities that practice demands while, on the other hand, practitioners have relied too much on makeshift empiricism devoid of any theoretical base. Although the paper does not succeed completely in bridging the gap, the level of achievement is well ahead of many other attempts.

The Markovian model used as the theoretical foundation for the work is logical and provides one of the most effective ways of dealing theoretically with multilane flow. The most doubtful feature of the model is the assumptions about basic driver behavior. Their basis is doubtful, and I do not accept them intuitively at least. Whether the empirical results confirm the claimed behavioral structure depends very much on the rationale for the development of these attributes and on whether any alternative structures were examined. These points are not clear. Although this is a possible weakness, it does appear that it need not be so, since it should be possible to formulate a model that is independent of behavioral assumptions at the micro level and that has a more generalized application to varying roadway geometries.

The empirical base for the model's calibration is limited, and broad-ranging interpretations or conclusions cannot be made at this stage, although the basic opportunity for further empirical work is provided. The model is valid for only two lanes of traffic, and there must be questions when the flow is extrapolated into the boundary of the congested flow region. It is this region that current empirically based models handle least well, and the present model, while pointing to some innovative methods of approaching the problem, needs much more validation. In this latter regard, the applicability of a Markovian model to the congested flow regime is extremely uncertain without considerable care in defining the random variable.

The development of capacity and level-of-service projections is the most interesting part of the applied work. The regression approach developed is innovative and worthy of extension although the limited empirical base limits the immediate strength of the results. Confirmation of the choice of the parabolic relationship would be desirable inasmuch as this is important in making the extrapolations and the empirical foundation is sparse.

In conclusion, I congratulate the author on his innovative look at several important areas of traffic engineering practice. It will, I hope, generate additional efforts to provide more satisfactory theoretical bases to the real world of traffic.

Abridgment

Freeway Lane-Changing Process: A Microscopic Approach

Moshe Levin, Chicago Expressway Surveillance Project, Illinois Department of Transportation

A model analyzing the driver's process of changing from one lane to the adjacent lane is described based on research presented by Levin (1). The model uses gap acceptance concepts to evaluate the probability of accomplishing a lane change within a certain distance under various traffic conditions. Such a model can be used in determining the effectiveness of the location of freeway directional and information signs.

MODEL DESCRIPTION

Lane-Changing Process

A vehicle is traveling in lane i at time mean speed U_i adjacent to a traffic substream traveling in lane i + 1 at speed U_{i+1} (in this case, $U_{i+1} > U_i$). The probability density function of headways in the i th lane may be denoted by $\theta_i(t)$.

At a certain time t a need for a lane change (in this case, to the left adjacent lane) is established by the drivers, and the lane-changing process begins. The driver attempting the lane change looks at the adjacent lane, lane i + 1, and considers some or all of the following:

1. The speed of the approaching (lagging) vehicle in lane i + 1,
2. The relative position of the leading vehicle in lane i + 1,
3. His or her own speed and operational characteristics, and
4. His or her gap acceptance characteristics.

At time t + T, where T is the driver's decision time, the driver either accepts the gap and changes lanes or rejects it and waits for an acceptable gap. If he or she decides to accept the gap, the lane-changing maneuver begins at the earliest possible moment when a safe maneuver can be accomplished. If the gap is rejected, the driver reduces his or her speed, evaluates the next gap, and reaches an appropriate decision concerning that gap.

The lane-changing process may be full or partial. The full process occurs either

1. When the gap encountered immediately after the need for a lane change has been established is rejected because it is smaller than the driver's critical gap, or
2. When the gap encountered is greater than the critical gap but the vehicle is in such a position relative to the lagging vehicle that the lane-changing maneuver is considered hazardous and the gap is rejected.

The full process consists of the following three phases:

1. Waiting for an acceptable gap or lag,
2. Bringing the vehicle to a position relative to the accepted gap so that the lane changing maneuver can be initiated, and
3. The lane changing maneuver.

The partial process occurs when the first encountered gap, after a need for a lane change has been established, is accepted. It takes place under either one of the following forms:

Form A, the position of the attempting vehicle must be adjusted relative to the accepted gap before a safe maneuver can be initiated, and

Form B, the relative position of the attempting vehicle allows the immediate initiation of the lane changing maneuver.

Both of these forms consist of phases 2 and 3 above. Each phase in the various forms of the lane-changing process has its own distribution function with respect to the distance involved. These functions are themselves functions of traffic conditions on the facility. The expression $f_i(x_i, \overline{V}, \overline{U})$ denotes the distribution functions of the distance required to complete phase i under volume condition $\overline{V}$ and speed condition $\overline{U}$. $\overline{V}$ and $\overline{U}$ are vectors representing traffic flow rates and time mean speeds respectively on the lanes of the one-way freeway section

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

during the time of the process.

Any function representing any phase of the process is independent of those characterizing other phases for a given set of speed and volume conditions. The distribution functions of the distance required for completing the various lane-changing processes were considered as the convolutions of the individual distribution functions describing each phase in the appropriate form of the process and can be presented as follows:

1. Full process

$$f_F(x,\bar{V},\bar{U}) = f_{1F}(x_1,\bar{V},\bar{U}) * f_{2F}(x_2,\bar{V},\bar{U}) * f_3(x_3,\bar{V},\bar{U}) \quad (1)$$

2. Partial process, form A

$$f_A(x,\bar{V},\bar{U}) = f_{2_A}(x_2,\bar{V},\bar{U}) * f_3(x_3,\bar{V},\bar{U}) \quad (2)$$

3. Partial process, form B

$$f_B(x,\bar{V},\bar{U}) = f_{2B}(x_2,\bar{V},\bar{U}) * f_3(x_3,\bar{V},\bar{U}) \quad (3)$$

The composite distribution representing the combination of the various lane-changing processes for a given set of volume and speed conditions is

$$f(x,\bar{V},\bar{U}) = a_F f_F(x,\bar{V},\bar{U}) + a_A f_A(x,\bar{V},\bar{U}) + a_B f_B(x,\bar{V},\bar{U}) \quad (4)$$

where a_f, a_A, and a_B are the probabilities of occurrence of the three types of the process.

Structure of the Accepted Gap

The section on the origin lane that parallels the accepted gap may be divided into three zones. One zone is that in which the driver cannot initiate a lane change because such a maneuver might end in a conflict with a leading vehicle on the destination lane. The size of this zone is the safe space lead. If a need for a lane change is established at time t while the attempting vehicle is within the above zone, the driver might find himself at time t + T either within or outside this zone. This, of course, depends on

1. The driver decision time T,
2. The relative speed $U_{i+1} - U_i$, and
3. Where within the zone the need for a lane change has been established.

If at time t + T the driver is outside the above zone, he or she is subject to form B of the partial process since he or she may start the lane-changing maneuver at time t + T. If the driver is still within this zone, he or she is subject to form A of the partial process, but the driver must be outside this zone before initiating the maneuver.

Generally, if the driver's traversed distance relative to the leading vehicle during decision time T is greater than the size of the above zone, the lane change process will be either full or form B of the partial. Conversely, if the size of the zone is greater than the decision relative distance, the driver may be subject to either one of the three types of the process.

The second zone is that in which the driver decides to reject the acceptable gap because a conflict with the lagging vehicle might result if the maneuver is executed. The driver is subject, in this case, to the full process. This zone is a combination of two subzones: the decision relative distance and the safe space lag.

The size of the third zone is, of course, a function of the size of the two previous zones and is the zone in which form B of the process takes place.

Process Phases

Phase 1 occurs when the driver attempting a lane change rejects the first gap he or she encounters in the adjacent lane after establishing a need for a lane change. The driver rejects the lane change because either

1. The gap encountered is smaller than the driver's current critical gap T_1, which is established according to an angular velocity model (2), or
2. The gap is greater than the driver's critical gap but the position of his or her vehicle relative to this gap is such that a lane change would be hazardous.

After rejecting the initial gap, the driver is assumed to change his or her speed and establish, according to the angular velocity model, a new critical gap, T_2. A distance distribution function may be derived from probability considerations of rejecting any sequence of gaps, size of average rejected gap, and speeds in the origin and destination lanes.

Phase 2 involves the movement of the attempting vehicle within the accepted gap before the lane-changing maneuver is initiated. In the case that the safe space lead is equal to or greater than the decision relative distance, the distance function in the full process is a single point function related to U_i, U_{i+1}, and the size of the safe space lead. In form A, the distance function is of a uniform nature related to U_i, U_{i+1}, T, and the safe space lead. In form B, the distance function is a single point function related to U_i and T. In the case in which the decision relative distance is greater than the safe space lead, the distance function in the full process can be described as a single point related to U_i, U_{i+1}, T, safe space lead, and the probability of the time spent in this phase. In form B, the distance function is a single point related to U_i and T.

Phase 3 is the lane-changing maneuver defined as the distance required for a vehicle to move from a straight-ahead path in the origin lane to a straight-ahead path in the destination lane.

Analysis of data collected by Worrall and Bullen (3) revealed that, for various volume and speed groupings, the distribution of the above distances was of an Erlang nature.

The probability of occurrence of each type of the process was determined for various traffic volume and speed groupings, based on the critical gap characteristics and the structure of the accepted gap.

DATA COLLECTION AND ANALYSIS

Data Collection

A northbound three-lane section (under automatic surveillance) of the Gulf Freeway in Houston was selected for this study. Data on traffic stream speeds and lane flow rates were collected by the control center, and an instrumented vehicle was driven by a test driver to obtain the following lane-changing data:

1. Delay in making a lane change once an instruction for a lane change has been given and
2. The distance traversed during the lane-changing process.

The critical gap characteristics of the test driver were determined through field measurements of his threshold angular velocity.

The characteristics of the headway distribution functions $\theta_i(t)$ on the freeway lanes were assumed to be of an Erlang nature, and the value of the parameters varied

according to flow rates, as shown by Drew, Buhr, and Whitson (4).

Data Analysis

The goodness of fit of the collected data to the model developed was determined by using the Kolmogorov-Smirnov test (5) and was found to exist at the 1 percent level of significance. Traffic data used in the analysis represented freeway levels of service B, C, and D.

APPLICATIONS

Application of the developed model can be found in reference (6) where the effectiveness of the location of directional signs, spaced as recommended by the Manual on Uniform Traffic Control Devices on a four-lane freeway, was analyzed. For the purpose of this analysis effectiveness was defined as the probability of completing a lane change within a distance determined by the location and size of a sign with respect to an exit ramp, for levels of service B and C. Based on this analysis it was found that for both levels of service the effectiveness of the sign at the gore area was between 0.35 and 0.45, while for the 1.6 and 3.2-km (1 and 2-mile) signs the effectiveness was very close to 1.0.

ACKNOWLEDGMENT

The contents of this paper reflect the views of the author who is responsible for the facts and accuracy of the data presented herein.

REFERENCES

1. M. Levin. Some Investigations of the Freeway Lane Changing Process. Texas A&M Univ., College Station, PhD dissertation, 1970.
2. R. M. Michaels and H. Weingarten. Driver Judgement in Gap Acceptance. U.S. Bureau of Public Roads, 1965.
3. R. D. Worrall and A. G. R. Bullen. Lane Changing on Multilane Highways. Northwestern Univ., Evanston, Ill., 1968.
4. D. R. Drew, J. H. Buhr, and R. H. Whitson. The Determination of Merging Capacity and Its Application to Freeway Design and Control. Texas Transportation Institute, Texas A&M Univ., College Station, Research Rept. 430-4, 1967.
5. B. Ostle. Statistics in Research. Iowa State Univ. Press, Ames, 1963.
6. M. Levin. Measuring Location Effectiveness of Freeway Directional Signs. Chicago Expressway Surveillance Project, 1975.

Regularity of Some Detector-Observed Arterial Traffic Volume Characteristics

William R. McShane and Kenneth W. Crowley, Department of Transportation Planning and Engineering, Polytechnic Institute of New York

An extensive data base was obtained for four two-lane arterial locations in Toronto. The data were 5-min detector counts in each base of each location for a total of 13 291 counts over 77 days. The paper reports on the regularity of the weekday pattern and of the time at which certain flow levels are reached. The patterns are quite regular, and the variance of the time at which specified peak levels are reached is rather small (from 4 to 6 min). A significant correlation between flows in the two lanes is found, particularly at higher volumes. The variation about the average pattern is also studied to assess the merits of using a predictor algorithm. The correlation between present and past variations (and thus future and present variations) is sometimes significant, but this correlation varies within a daily record. Further, it is not necessarily the same magnitude or sign at a given time on different days. It is nonetheless possible to construct a predictor that can extract information. The issue of the utility and the cost effectiveness of such prediction is noted, as is the fact that such an issue can only be resolved by considering the specific control with which the predictor is to be used. The issue of the typicalness or atypicalness of the Toronto regularity of pattern must also be considered.

The regularity of daily traffic patterns is of basic concern when various candidate control strategies are evaluated. At one extreme, if little variation exists from day to day, a time-of-day (minimal response) controller will undoubtedly suffice. At the other extreme, if no discernible pattern exists or if it arrives randomly each day, then a highly responsive control is appropriate.

Between these extremes, there is the possibility of a basic underlying pattern with substantial variation about it. There is also the possibility that on a given day there will be a major deviation from an otherwise extremely regular pattern. To accommodate such possibilities, traffic data can be collected in real time and deviations noted, or a predictor can be established to estimate future values.

Efforts to control congestion should consider (a) the regularity of daily patterns and (b) the regularity of the times at which various flow levels are reached. These considerations provide insight into the utility of minimal response control versus highly responsive control in this flow regime.

An extensive data base was acquired through the courtesy of the Metropolitan Toronto Department of Roads and Traffic. This data base was used to investigate

1. The regularity of the weekday pattern,
2. The regularity of the time at which certain flow levels are reached,
3. The correlation of the flow between lanes on the same approach, and
4. The potential for refining estimates of traffic volumes, based on observation of the deviations from the historic or nominal pattern.

DATA BASE

The Metropolitan Toronto Department of Roads and Traffic provided an extensive data base, consisting of 5-min samples of volume by lane at each of four sites over a 77-day period. The data were collected by computer from September 24 to December 10, 1973.

Figures 1 and 2 show the detector locations for each of the four sites.

The data were acquired from the detectors whenever possible. Table 1 gives the distribution of the samples by day of week and hour of day. Note that all periods of common interest are well covered. The data were output in 5-min counts, and the first counting period was initiated whenever the computer "came up." For simplicity, all data were shifted to standard 5-min periods; the first period was midnight to 12:05 a.m. The time shift thus introduced was uniformly distributed between -2.5 and +2.5 min. This could introduce a standard deviation of $\sigma = 1.44$ min in any time shift estimates.

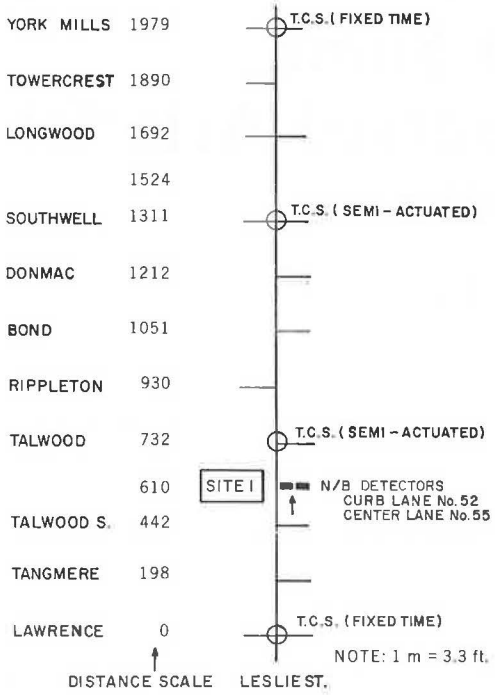
REGULARITY OF DAILY PATTERN

The average pattern was computed for each day by averaging the volume observed in each time slot. Because the weekdays did not differ substantially, a single average pattern was also computed by aggregating all weekdays at each detector.

Figure 3 shows the average weekday pattern at each

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

Figure 1. Location of detectors at site 1 (suburban).



of the four sites. The shaded area indicates the region within which one can expect 95 percent of the observations to fall. This is not a confidence bound on the average. The confidence bound is much tighter. It is an estimate of the fluctuations from day to day within a specified time slot.

Figure 3 shows that, although there is substantial variation most of the time, the peaking is quite sharp, and relatively little variation exists in the time at which certain levels are reached. That is, specified level X is reached at approximately the same time every day. One can even question whether the variation in the vertical dimension is as substantial as it appears: The 95 percent range is on the order of ±120 vehicles/h/lane or ±10 vehicles/lane in a 5-min period. This occurs when the average count is on the order of 40 vehicles/lane in a 5-min period. Further, the distribution is approximately normal, so that the deviations tend to be clustered near zero.

Figure 4 shows the distribution of the times at which certain volumes are reached at site 2. For instance, 720 vehicles/lane is first attained earlier than 6:55 a.m. only 10 percent of the time. By 7:07 a.m., there is a 90 percent chance that 720 vehicles/h/lane has already occurred. Note that the variation is rather small, leading to the conclusion that minimal response policies could be developed with some assurance that the onset of certain levels could be anticipated with some confidence.

Figure 2. Location of detectors at sites 2, 3, and 4 (central area).

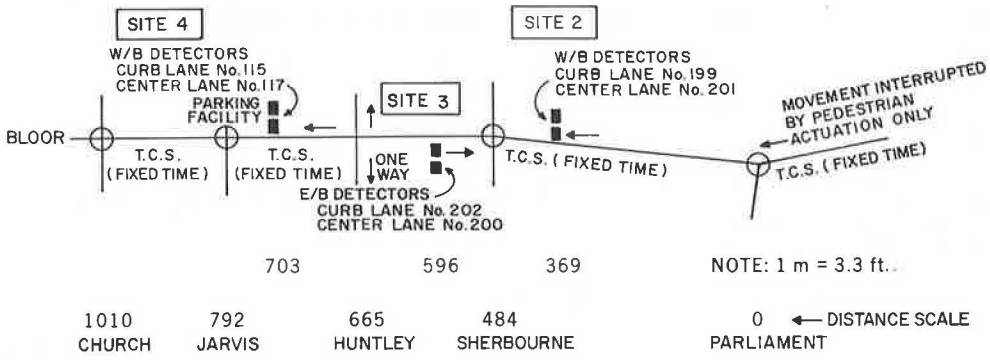


Table 1. Distributions of samples by day and time.

Time	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.	Total	Percent
Midnight	115	69	46	38	52	92	104	516	3.9
1 a.m.	69	0	0	12	12	26	83	202	1.5
2 a.m.	96	26	0	12	10	0	72	216	1.6
3 a.m.	96	56	0	35	13	14	72	286	2.2
4 a.m.	97	80	29	68	23	36	79	412	3.1
5 a.m.	129	116	74	89	99	75	96	678	5.1
6 a.m.	143	113	97	109	129	104	113	808	6.1
7 a.m.	144	132	117	120	132	108	120	873	6.6
8 a.m.	139	133	118	120	132	94	65	801	6.0
9 a.m.	135	128	120	120	131	82	0	716	5.4
10 a.m.	136	132	108	104	124	57	0	661	5.0
11 a.m.	132	131	108	96	116	43	0	626	4.7
Noon	67	44	45	47	61	48	0	312	2.3
1 p.m.	131	117	116	81	94	48	0	587	4.4
2 p.m.	133	120	120	86	101	48	0	608	4.6
3 p.m.	141	120	117	97	127	57	2	661	5.0
4 p.m.	144	132	132	120	132	77	24	761	5.7
5 p.m.	144	131	132	120	132	89	29	777	5.8
6 p.m.	118	114	114	100	126	95	72	739	5.6
7 p.m.	57	32	15	45	95	103	106	453	3.4
8 p.m.	15	0	0	1	91	108	117	332	2.5
9 p.m.	8	0	0	7	84	103	120	322	2.4
10 p.m.	52	30	8	45	86	96	119	436	3.3
11 p.m.	70	46	28	57	92	101	114	508	3.8
Total	2511	2002	1644	1729	2194	1704	1507	13 291	100.0
Percent	19	15	12	13	17	13	11	100	

It should be noted that there is no protection against statistical rare events, i.e., major deviations from the average pattern. Generally, these can be associated with weather or special community activities.

CORRELATION BETWEEN LANES

If a volume is to be observed, it is necessary to determine whether one lane will suffice or whether two or more lanes must be measured. For the data available, which contained only two-lane approaches, the question reduces to one or both.

It was found that all weekdays could be aggregated for a given site and that a two-regime curve relating total volume to curb-lane volume should be established. The breakpoint was established at 360 vehicles/h in the curb lane, and no data for a curb-lane volume of less than 240 vehicles/h were considered.

Data given in Table 2 show that the total volume is rather strongly correlated to curb-lane volume in the higher flow regime, with correlation coefficients on the order of 0.9 and more. The correlation equation used for the higher flow regime is

$$Y = A + C(X - 30) \quad \text{for } X \geq 30 \tag{1}$$

The correlation equation used for the lower flow regime is

$$Y = A + B(X - 30) \quad \text{for } X < 30 \tag{2}$$

where

Y = total 5-min count,
A = intercept at curb,
B = slope in regime for low flows,
C = slope in regime for high flows, and
X = curb-lane 5-min count.

A least squares fit was performed on the data for each site, with the two curves tied to a common point at the boundary between regimes. The results are summarized in Table 2 and are shown in Figure 5. Note that the slopes of lines in the $X \geq 30$ regime are comparable, except for site 1, the suburban site.

Clearly, the lane split tends to equalize as volume increases. The downtown sites tend to have greater concentrations in the outer (left) lane, although data are certainly not sufficient to identify location as a causative factor.

Based on these fits, the total volume can in fact be computed with some confidence if the calibrated lines are known.

It was determined that the weekend curves are not dissimilar to the weekday curves, within the range of the weekend data.

POTENTIAL FOR PREDICTION

On any given day, the actual pattern differs from the average pattern by some set of discrepancies ϵ_i. At a particular site, let

a_i = average count for time period i,
x_i = actual count for time period i, and
$\epsilon_i = x_i - a_i$.

If the set ϵ_i is serially uncorrelated, then there is no hope for predicting a future value x_j any better than simply saying that its expected value is a_j. If there is some correlation among the ϵ_i, however, then there is information contained in the sequence of past values of ϵ_i. This information may be extrapolated into the future to get some better estimate (i.e., prediction) of a value x_{i+k}.

Predictors may be based on nothing more than historic patterns, they may ignore the historic pattern and project forward based on current trends, or they may project forward a refinement to the historic pattern based on recent (i.e., real time) deviations from the historic pattern. Because the weekday pattern appears so regular (Figures 3 and 4), it appears most fruitful to investigate whether there is any additional information in the set of discrepancies.

The set of $\{\epsilon_i\}$, which is the outcome on any given day, may be viewed as the outcome of a zero-mean process. It is of interest to compute the autocovariance values at one and two lags: $R(1) = E(\epsilon_i \epsilon_{i+1})$ and $R(2) = E(\epsilon_i \epsilon_{i+2})$, as well as the variance $E(\epsilon_i^2)$. If a sample of N data points were available, these could be estimated by

$$\hat{R}(k) = (1/N - 1 - k) \sum_{i=1}^{N-1} \epsilon_i \epsilon_{i+k} \tag{3}$$

$$\hat{\sigma}^2 = (1/N - 1) \sum_{i=1}^{N} \epsilon_i \epsilon_i \tag{4}$$

respectively. If there were M days available, the estimates obtained from equations 3 and 4 for each day could then be averaged to improve the estimate.

The use of equations 3 and 4 assumes, however, that the quantities R(1), R(2), and σ^2 do not change as a function of time within the day. This is not necessarily true.

To ensure that such an assumption is not made unless justified, we used

$$\hat{R}_i(k) = \beta \hat{R}_{i-1}(k) + (1 - \beta)\epsilon_i \epsilon_{i-k} \tag{5}$$

$$\hat{\sigma}_i^2 = \beta \hat{\sigma}_{i-1}^2 + (1 - \beta)\epsilon_i \epsilon_i \tag{6}$$

as estimators, so that, if the quantities of interest do change, it will be reflected in the estimators. β controls both the responsiveness of the estimator to change and the variance (and thus confidence) of the estimate. $\beta = 0.8$ was used in the material presented here.

It may be shown that most predictors of interest depend on the quantities above. Indeed, if a predictor is being used such that one has past data through period K - 1 and is computing the values for period K + 1 during period K—the data for which are not yet in hand—then the quantity of interest is $E(\epsilon_{i+2}\epsilon_i)$, which indicates how much knowledge of period K + 1 can be extracted from period K - 1 and before.

The quantities $\hat{\rho}_i(1)$ and $\hat{\rho}_i(2)$ may be computed from

$$\hat{\rho}_i(k) = \left[\hat{R}_i(k) \Big/ \sqrt{\hat{\sigma}_i^2 \, \hat{\sigma}_{i-k}^2}\right] \tag{7}$$

Therefore, the autocovariance values are indeed functions of the time of day, and they should be treated as such. Moreover, there are times when they are sufficiently high that they can indeed be used effectively to refine the estimate of a future volume x_{i+k}.

Of course, it must be recognized that the effectiveness, while real, may not be cost effective. If $\hat{\rho}_i(2) \cong 0.7$, the variance reduction is on the order of $(0.7)^2 \cong 0.5$ or 50 percent. The standard deviation is thus reduced to $1/\sqrt{2}$ or 0.707 of its former value. Recall that the 95 percent bounds were ±120 vehicles/h/lane for a mean of about 480 vehicles/h/lane. They would now be reduced to ±0.707(120) or ±85 vehicles/h/lane. The net tightening of the range is (120 - 85)/12 = 3 vehicles per lane per 5 minutes. The question is whether this added precision is worth the cost.

Figure 3. Average weekday pattern and variation of individual flows.

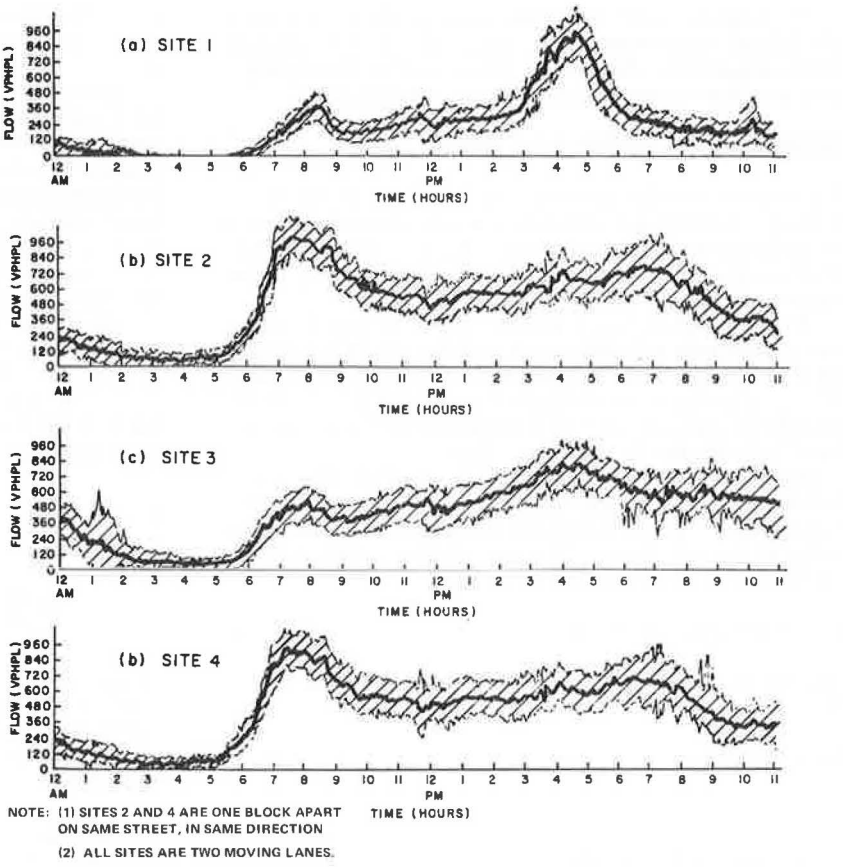


Figure 4. Times at which certain levels are reached (site 2).

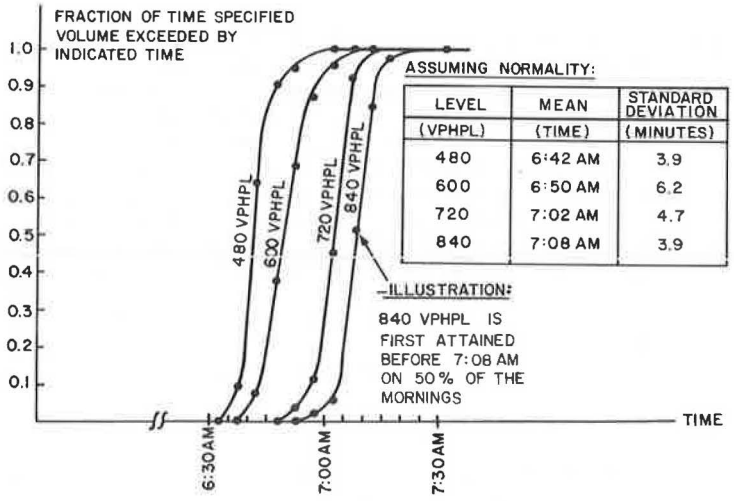


Figure 5. Least squares fit of total count as a function of curb-lane count, weekdays only.

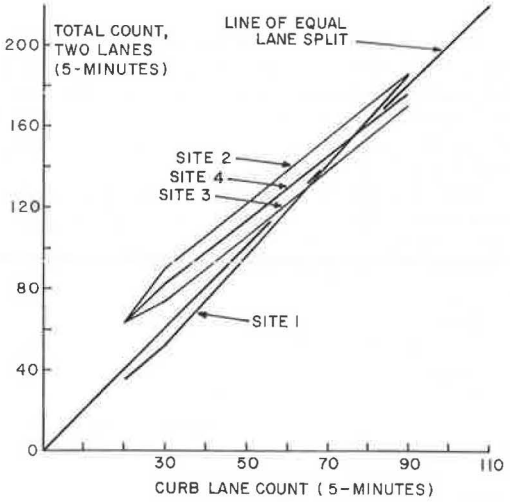


Table 2. Summary of correlations, weekdays only.

Site	Regime	Correlation: Total to Curb	Correlation: Center Lane to Curb	Sample Size	Slope in Regime	Intercept at Curb (vol = 30)
1	Low	0.67	0.37	2936	1.7	52.6
	High	0.98	0.93	3416	2.2	
2	Low	0.63	0.46	1640	2.5	89.4
	High	0.95	0.76	6311	1.6	
3	Low	0.27	0.08	1912	1.2	74.4
	High	0.88	0.59	6124	1.6	
4	Low	0.49	0.32	1633	2.1	23.5
	High	0.92	0.62	6400	1.5	

The correlations not only vary with time but also do not have comparable values at the same clock times from day to day. The indication is that, if a person decides to undertake a refined prediction of future volume levels, it should be undertaken with on-line estimation of the predictor coefficients. These coefficients vary with time. Moreover, this variation cannot be specified a priori from historic data as a simple or even complex function of time.

Note that this last conclusion can only be reached through a data base such as that used herein. Many daily sets $\{\rho_t(k)\}$ must be observed. Note also that the averaging of many sets $\{\hat{\rho}_t(k)\}$ or $\{\hat{R}_t(k)\}$ will not produce a better estimate of the true temporal correlations. Indeed, it will obscure it, for the average will tend to zero or at least to small values.

DISCUSSION OF RESULTS

The data were of particular interest because of the implications for the control of congested and oversaturated street networks (1). In regard to control of congested traffic flow, the following can be said about the Toronto data.

1. Because the daily pattern at a specific site is quite regular, minimal response (i.e., preplanned) signal control can be considered as a viable approach.
2. The times at which specific higher volumes are first reached is even more regular, and the initiation of those levels can be anticipated with some confidence. Although preplanned switching can be used with confidence, the variability that does exist lessens the benefits of multiple, rapid switches.
3. Single-lane detection provides good indications of the total approach volume, at least on two-lane approaches.
4. Variations about the average pattern are serially correlated, so that some information can be extracted from past variations to enhance volume predictions.

These results support the use of minimal response policies in many applications. Further, both these remarks and the discussion below are based on regularity as observed in Toronto. This has not been conclusively demonstrated to be the general pattern. Neither, however, are there strong arguments that it is atypical.

Before remarks are made about the entire range of flow levels exhibited in the data, two types of responsive systems must be distinguished: actuated equipment and coordinated system control algorithms. The time frame of the data studied herein (i.e., 5-min) is relevant only for the second type, and the remarks address only such systems.

It should be observed that the information contained to this point relates only to the inherent variability of the traffic flow and to potential for refining traffic predictions. It was noted that a two-step predictor can sometimes reduce the variance by 50 percent. This can result in a reduction in the deviation from the mean of 25 to 17.5 percent of the mean when the 95 percent range of volume values is considered.

It does not follow, however, that this reduction is important. The control algorithm may just not need such a precise prediction of volume. Even if there is some benefit to such added precision, there is the question of cost effectiveness: The detectors cost money, as does the maintenance of the detector system. Further, there are computation burdens on the computer system and storage requirements for intermediate results.

Because of these considerations, use of such predictors must be evaluated in terms of the cost effectiveness of the control function delivered, with predictor and control effectiveness integrally related. It is clear at this point that meaningful results can be obtained with only minimal response policies. It remains for advanced control projects such as the UTCS program to determine the cost effectiveness of various predictor-control policy combinations.

If open loop policies are used, however, two problems remain: (a) potentially severe impacts due to the rare event that would be ignored and (b) determining the average pattern, which may change slowly over time (i.e., months or years). The first problem can be addressed by limited surveillance, not for on-line adaptations but for a quality control chart type of monitoring on whether the average pattern assumed is indeed applicable.

The second problem can likewise be addressed by detectors placed for surveillance, and not necessarily for on-line control purposes. Such detectors could be sampled systematically, not all necessarily on the same day, so as to update average stored patterns.

ACKNOWLEDGMENTS

The authors thank Leonard Rach and Joseph Lamm of the Metropolitan Toronto Department of Roads and Traffic for providing the extensive data base and for their interest in the work conducted. Barbara A. Cicala of the Polytechnic Institute of New York assisted in the analysis, especially the many computer runs necessary.

This study was conducted under National Cooperative Highway Research Program Project 3-18(2). The opinions and findings expressed or implied in this paper are those of the authors. They are not necessarily those of the Federal Highway Administration, the American Association of State Highway and Transportation Officials, or the National Cooperative Highway Research Program.

REFERENCES

1. L. J. Pignataro, W. R. McShane, K. W. Crowley, T. W. Casey, and B. Lee. Traffic Control in Oversaturated Street Networks. Polytechnic Institute of New York, Final Rept., NCHRP Project 3-18(2), June 1975.
2. Urban Traffic Control System and Bus Priority System Traffic Adaptive Network Signal Timing Program—Software Description. Transportation and Environmental Operations, TRW, Aug. 1973.
3. E. B. Lieberman and others. Variable Cycle Signal Timing Program: Volume 4—Production Algorithms, Software and Hardware Requirements, and Logical Flow Diagrams. KLD Associates, Inc., June 1974.