HIGHWAY RESEARCH RECORD

Number 89

Traffic Flow Theory

3 Reports

Presented at the
43rd ANNUAL MEETING
January 13-17, 1964
and
44th ANNUAL MEETING
January 11-15, 1965

SUBJECT CLASSIFICATION
54 Traffic Flow
53 Traffic Control and Operations

1.20

HIGHWAY RESEARCH BOARD
of the
Division of Engineering and Industrial Research
National Academy of Sciences—National Research Council
Washington, D. C.
1965

Department of Traffic and Operations

Fred W. Hurd, Chairman
Director, Bureau of Highway Traffic, Yale University
New Haven, Connecticut

COMMITTEE ON TRAFFIC CONTROL DEVICES
(As of December 31, 1963)

William T. Taylor, Jr., Chairman
Assistant Traffic and Planning Engineer
Louisiana Department of Highways, Baton Rouge

Robert E. Conner, Secretary
Assistant Engineer of Traffic
Ohio Department of Highways, Columbus

W. C. Anderson, Chief Research and Development Engineer, Union Metal Manufacturing Company, Canton, Ohio
Donald S. Berry, Chairman, Department of Civil Engineering, The Technological Institute, Northwestern University, Evanston, Illinois
C. E. Billion, San Diego, California
James W. Booth, Traffic Engineer, Utah State Road Commission, Salt Lake City
Donald E. Cleveland, Assistant Research Engineer, Texas Transportation Institute, Texas A & M College, College Station
Abner W. Coleman, Traffic Engineer, Vermont Department of Highways, Montpelier
F. B. Crandall, Traffic Engineer, Oregon State Highway Department, Salem
J. E. P. Darrell, Traffic and Planning Engineer, Minnesota Department of Highways, St. Paul
Robert D. Dier, City Traffic Engineer, Long Beach, California
William H. Dorman, Lighting Product Development Laboratory, Corning Glass Works, Corning, New York
Daniel L. Gerlough, Head, Traffic Systems Section, Planning Research Corporation, Los Angeles, California
J. Al Head, Chief, Planning and Standards Division, Office of Highway Safety, U. S. Bureau of Public Roads, Washington, D. C.
J. T. Hewton, Operations Engineer, Traffic Engineering Department, Municipality of Metropolitan Toronto, Toronto, Canada
Rudolph Hofer, Jr., Highway Products and Structural Section, Sales Development Division, Aluminum Company of America, New Kensington, Pennsylvania
George W. Howie, Director of Public Utilities, Cincinnati, Ohio
Matthew J. Huber, Bureau of Highway Traffic, Yale University, New Haven, Connecticut
Rudolph J. Israel, Assistant Traffic Engineer, Traffic Department, California Division of Highways, Sacramento
James H. Kell, Assistant Research Engineer, Institute of Transportation and Traffic Engineering, University of California, Berkeley
Frank S. Kovach, Assistant Superintendent of Signal Systems, Akron, Ohio
Holden M. LeRoy, Traffic Control Engineer, Department of Streets and Traffic, Detroit, Michigan
J. Carl McMonagle, Institute for Community Development, Michigan State University, East Lansing
Phillip S. Mancini, Chief, Division of Traffic Engineering and Highway Planning, Rhode Island Department of Public Works, Providence
Fred J. Meno, II, Electrical Engineer, Public Lighting Commission, City of Detroit, Detroit, Michigan
William J. Miller, Jr., Director of Operations, Delaware State Highway Department, Dover
J. P. Mills, Jr., Traffic and Planning Engineer, Virginia Department of Highways, Richmond

James V. Musick, Winko-Matic Signal Company, Lorain, Ohio
A. R. Pepper, Traffic Engineer, Colorado Department of Highways, Denver
Marshall M. Rich, Wilbur Smith and Associates, Columbia, South Carolina
Edmund R. Ricker, Director, Traffic Engineering Bureau, Pennsylvania Department of Highways, Harrisburg
Frank G. Schlosser, Chief Engineer, Pfaff and Kendall, Newark, New Jersey
Rex G. Still, Traffic Engineer, Washington State Highway Commission, Olympia
Asriel Taragin, Special Assistant for Control of Research and Development, U. S. Department of Commerce, Washington, D. C.
Robert E. Titus, Planning Engineer, Planning and Traffic Division, West Virginia State Road Commission, Charleston
George M. Webb, Sacramento, California
Arthur M. White, Traffic Control and Safety Engineer, Mississippi State Highway Department, Jackson
Robert M. Williston, Chief of Traffic, Connecticut State Highway Department, Hartford
David K. Witheford, Bureau of Highway Traffic, Yale University, New Haven, Connecticut

COMMITTEE ON THEORY OF TRAFFIC FLOW

(As of December 31, 1963)

Daniel L. Gerlough, Chairman
Head, Traffic Systems Section, Planning Research Corporation
Los Angeles, California

Donald E. Cleveland, Acting Secretary
Associate Research Engineer, Texas Transportation Institute
Texas A & M College, College Station

John L. Barker, General Manager, Automatic Signal Division, Eastern Industries, Inc., Regent Street, East Norwalk, Connecticut
Martin J. Beckmann, Department of Economics, Brown University, Providence, Rhode Island
Don G. Capelle, Traffic Research Engineer, Automotive Safety Foundation, Washington, D. C.
J. Douglas Carroll, Jr., Deputy Director, Tri-State Transportation Committee, New York, New York
A. Charnes, Technological Institute, Northwestern University, Evanston, Illinois
Theodore W. Forbes, Department of Psychology and Engineering Research, Michigan State University, East Lansing
Herbert P. Galliher, Associate Director, Operations Research, Massachusetts Institute of Technology, Cambridge
Bruce D. Greenshields, Assistant Director, Transportation Institute, University of Michigan, Ann Arbor
Richard H. Haase, The RAND Corporation, Santa Monica, California
Frank A. Haight, Associate Research Mathematician, Institute of Transportation and Traffic Engineering, University of California, Los Angeles
Robert Herman, Research Laboratories, General Motors Corporation, Warren, Michigan
A. W. Jones, Bell Telephone Laboratories, Holmdel, New Jersey
James H. Kell, Assistant Research Engineer, Institute of Transportation and Traffic Engineering, University of California, Berkeley
Sheldon L. Levy, Director, Mathematics and Physics Division, Midwest Research Institute, Kansas City, Missouri
Russell M. Lewis, Assistant Professor of Civil Engineering, Rensselaer Polytechnic Institute, Troy, New York
E. W. Montroll, Washington, D. C.
Karl Moskowitz, Assistant Traffic Engineer, California Division of Highways, Sacramento

Joseph C. Oppenlander, School of Engineering, Purdue University, West Lafayette, Indiana
Martin C. Stark, Operations Research Analyst, Data Processing Systems Division, National Bureau of Standards, Washington, D. C.
Asriel Taragin, Assistant to the Director, Office of Research and Development, U. S. Bureau of Public Roads, Washington, D. C.
William P. Walker, Chief, Geometric Standards Branch, Highway Standards and Design Division, U. S. Bureau of Public Roads, Washington, D. C.
Martin Wohl, Lecturer and Director of Transport Research, Department of Economics, Harvard University, Cambridge, Massachusetts

Department of Traffic and Operations

Harold L. Michael, Chairman
Associate Director, Joint Highway Research Project
Purdue University, Lafayette, Indiana

COMMITTEE ON VEHICLE CHARACTERISTICS
(As of December 31, 1964)

K. A. Stonex, Chairman
Automotive Safety Engineer, Technical Liaison Section, Engineering Staff
General Motors Corporation, Warren, Michigan

W. A. McConnell, Secretary
Executive Engineer, Technical Analysis, Engineering and Research Staff
Ford Motor Company, Dearborn, Michigan

Kenneth Boldt, The Pure Oil Company, Research Center, Crystal Lake, Illinois
W. F. R. Briscoe, Manager, Road Testing, Product Development, United States Rubber Company, Detroit, Michigan
F. B. Crandall, Traffic Engineer, Oregon State Highway Department, Salem
Timothy F. Creedon, Highway Engineering Advisor, Automobile Manufacturers' Association, Detroit, Michigan
A. H. Easton, Department of Mechanical Engineering, The University of Wisconsin, Madison
D. M. Finch, Institute of Transportation and Traffic Engineering, University of California, Berkeley
Roy C. Haeusler, Automotive Safety Engineer, Engineering Division, Chrysler Corporation, Detroit, Michigan
L. C. Kibbee, Director, Engineering Department, American Trucking Associations, Inc., Washington, D. C.
John C. Kohl, Assistant Administrator, Transportation, Housing and Home Finance Agency, Washington, D. C.
J. Carl McMonagle, Institute for Community Development, Michigan State University, East Lansing
Alger F. Malo, Director, Department of Streets and Traffic, Detroit, Michigan
L. H. Nagler, Staff Engineer - Safety, American Motors Corporation, Detroit, Michigan
Walter C. Oram, Senior Highway Engineer, Paving Bureau, Portland Cement Association, Chicago, Illinois
F. William Petring, U. S. Bureau of Public Roads, Washington, D. C.
Roy B. Sawhill, Department of Civil Engineering, University of Washington, Seattle
Ross G. Wilcox, Executive Secretary, Safe Winter Driving League, Chicago, Illinois
Light B. Yost, Director, Highway and Traffic Safety Section, General Motors Corporation, Detroit, Michigan
Thomas E. Young, Principal Traffic Engineer, Division of Traffic Engineering, Cincinnati, Ohio

Foreword

The complexities of traffic behavior have been studied in varying degrees by investigators from various disciplines. The mathematical approach, aimed at deeper understanding of traffic phenomena, has resulted in a seemingly unending number of theoretical solutions to traffic problems. Admittedly, this mathematical approach, oriented as it is toward theoretical answers to traffic problems, is not always of immediate help to the practicing traffic and highway operations engineer. Yet, familiarity with the concepts brought out by traffic flow theorists is vitally needed in order to arrive at the practical solutions that must be found. The three papers in this Research Record offer some insight into these concepts.

Two of the papers were presented at the Board's 43rd Annual Meeting in January 1964, while the third was given at the 44th Annual Meeting. The reports are concerned with theoretical aspects of operating a freeway control system in peak periods, the use of traffic simulation to determine delay and fuel consumption of vehicles at intersections, and "spillback" resulting from queueing on highways. This Record should be of interest to both personnel specializing in freeway operations and to mathematicians concerned with theory of traffic flow.

The paper, "Peak-Period Control of a Freeway System—Some Theoretical Investigations," poses theoretical questions and then discusses them. For example, what is the objective of such control? How can congestion be predicted? What type of detection system is required? How can this complex system be considered and evaluated? How can these concepts be applied in the development of a flexible control system?

"Operating Costs at Intersections Obtained from the Simulation of Traffic Flow" relates how a computer program was used to simulate traffic. Delays and fuel consumption of vehicles passing through an intersection were calculated for vehicles traveling on the streets leading to the intersection. The program was operated at various combinations of traffic volume under different types of intersection traffic control. Graphs of costs were plotted comparing volumes and costs and preliminary warrants were determined for furnishing a traffic controls system from an economic viewpoint.

A New York researcher, in the last paper entitled "Spillback from an Exit Ramp of an Expressway" has investigated the progressive deterioration of a roadway system due to spillback from one section to the others. This so-called spillback is the result of queueing at certain points. The researcher questions and attempts to learn if judicious management of the inevitable queues might decrease the aggregate delay to the users of the entire system.

Contents

Peak-Period Control of a Freeway System—Some Theoretical Investigations

JOSEPH A. WATTLEWORTH*, Assistant Research Engineer, Texas Transportation Institute, and
DONALD S. BERRY, Chairman, Department of Civil Engineering, Northwestern University

This paper contains a series of theoretical considerations pertaining to the control of a freeway system during peak traffic periods. It attempts to answer such questions as the following: What are the objectives of operation of a highway system? What type of control technique should be used? What type of traffic detection is required? Where is it to be located? How should the entire freeway system be considered and controlled to produce optimal operation?

An arbitrary street and/or freeway system is analyzed to determine the objective function or goal of operation for the system. An input-output analysis is used.

A theory of flow at bottlenecks is developed to explain the reduction of flow rate at some bottlenecks during congestion while the flow rate at other bottlenecks remains at its capacity level during congestion. This is a macroscopic flow model based on basic continuity equations.

Other macroscopic models of traffic behavior at and upstream of a bottleneck are to determine what traffic variables are to be detected to (a) predict congestion and (b) indicate congestion and how far from the critical sections the detection must be made in order to allow control decisions to be made.

Several criteria are established for control techniques and several control techniques are examined in light of these criteria. The possible role of each in a final control system is also discussed.

Finally, a linear programming model of the operation of a freeway system is presented. This can be used as a descriptive model, but with some modifications could probably be used to provide the control actions required for optimal system operation. Interpretation of the dual variables and a sensitivity analysis are included and these provide many valuable insights into the operation of a freeway system. The linear programming approach suggested a method of determining demand at a certain freeway location.

•PEAK-PERIOD congestion is a frequent occurrence on many urban freeways. This is partly due to partial completion of planned freeway systems. It is doubtful, however, if enough freeways can be built (or should be built) to eliminate completely peak-period congestion. Hence, an operational or control means is needed to provide relief from congestion during the peak periods at each stage of freeway construction as well as for the completed system.

---

*Formerly with Expressway Surveillance Project, Illinois Division of Highways.

Paper sponsored by Committee on Theory of Traffic Flow and presented at the 43rd Annual Meeting.

There is evidence to indicate that congestion on a freeway can decrease the flow rate on the freeway, whereas congestion on an arterial street does not decrease the flow rates on those streets. For this and other reasons, prevention of congestion on freeways seems to be more important than its prevention on surface streets.

Congestion develops at a bottleneck[1] when the demand exceeds the capacity there. The controls then must increase the capacity of some bottlenecks and/or shift the demand, either spacially or temporally. The development of congestion in one part of the system may have quite different effects on various other parts of the system.

1. If the queue which develops at a bottleneck backs upstream past one or more exit ramps, the output rate at these ramps may be decreased.
2. The unusually early development of congestion at one bottleneck may cause the downstream flow rate to decrease earlier than on a "typical" day. This can either (a) decrease the output rate of the system of interest, or (b) delay or eliminate congestion at a downstream bottleneck, thereby increasing the output rate of the system of interest.

The interdependence of critical locations suggests that a systems analysis would be the only adequate analytical approach to the problem of reduction or elimination of congestion on a freeway system.

Because traffic conditions can change rapidly with time, freeway controls should be traffic-adjusted to respond to traffic conditions in a large area rather than only to conditions at individual locations. Thus, a control system is needed rather than a series of independently operated controls. A control system, properly designed and operated, could result in the optimal performance of at least a portion of the entire system, that is to say, a subsystem.

Before the development of a freeway control system, it is necessary to: (a) determine the objectives of the control, (b) understand traffic behavior so that the responses to the control system can be predicted, (c) understand traffic behavior since it affects detector locations, and (d) develop an analytical means of describing and considering a freeway system or subsystem both for predicting the effects of the control and for operating the controls.

SCOPE

This study was not intended to be the complete development of a new and different control system to be applied to streets and freeways. The intention is, rather, to undertake some of the theoretical investigations on which the final development depends. The controls which are investigated have as their purpose the improvement of traffic conditions on urban freeways. Controls to improve operation of arterial streets are not considered, and only special peak-period freeway controls are considered. Primary attention is given to normal operating conditions on the freeways.

Every control scheme has some philosophy behind it, even though it is usually not explicitly stated. Because only controls for the improvement of peak-period freeway traffic flow are considered, the basic philosophy of these controls is that improving operation on the freeways will result in a net improvement in the operation in the entire system of streets and freeways. There can be little doubt of this if the freeway can be operated optimally without diverting a significant portion of the demand to other parts of the system and if the freeway controls do not otherwise affect the operation of the streets. Hence, an emergent philosophy is that as little as possible of the freeway demand will be diverted to the arterial street system. The problem then becomes one of operating the freeway in an optimal manner with the given demand.

---

[1] In this paper, a location at which the capacity is lower than the possible upstream flow rate is called a bottleneck. Bottlenecks can be placed in two broad categories, permanent or geometric bottlenecks and temporary bottlenecks such as accidents and disabled vehicles. Two types of freeway operation result from these types of bottlenecks. Under "normal operation," traffic behavior is affected primarily by the geometric features of the freeway; under "reduced capacity operation" the capacity of one or more sections of the freeway has been reduced by a temporary bottleneck.

This type of optimal freeway operation does not assure the optimal operation of the entire street and freeway system. It does, however, assure an improvement. Perhaps the diversion of part of the freeway demand to the streets would improve the operation of the total system of streets and freeways. To determine the optimal control for the total system would require a much larger and more difficult study than is undertaken here.

The first part of this study is the development of a criterion function by which the performance of a freeway control system can be evaluated. This is accomplished by means of some theoretical, input-output considerations of an arbitrary highway system.

Also included is a discussion of the macroscopic behavior of traffic at a bottleneck and at points upstream of and downstream of the bottleneck. The applications and implications of these theoretical discussions to peak-period traffic control as well as to the type and location of traffic sensing or detection that would be required for the proper operation of a freeway control system are also discussed.

Finally, a linear programming model of the operation of a one-directional freeway subsystem is developed. This model presents many insights into the operation and control of the system. With some additional refinements it could possibly be used for control purposes to indicate the input rates at the entrance ramps which would yield optimal system operation for the reduced capacity situation.

In summary, this study is a series of theoretical investigations which are meant to provide some of the foundations upon which a traffic control system can be developed.

OBJECTIVES OF FREEWAY CONTROL

The discussions which follow are centered around the behavior of an arbitrary urban highway system. The arbitrary "system of interest" discussed could be the entire street and freeway system, the entire freeway system or any subsystem which does not violate the assumptions which will be presented. The system of interest, which is simply called the system in the following discussions, is visualized as being cut by a cordon line (or cordon lines) and is a closed system. Because it is a closed system, the cordon line(s) defines the locations of all of the system inputs and outputs.

Continuity Characteristics of Traffic Flow

In a closed system, the flow of traffic at all times satisfies the basic continuity equations. In terms of instantaneous rates, the input rate to the system equals the system output rate plus the rate of storage or accumulation within the system. Expressed mathematically, this becomes $i(t) = o(t) + s(t)$.

When used to describe a given time period, the continuity equation states that the number of vehicles entering the system equals the number leaving the system plus the change in the number within the system. For the time period from t_0 to t_1, $I(t_1) = O(t_1) + \int_{t_0}^{t_1} s(t)dt$. (See footnote 2.)

By similar reasoning, the number of vehicles in the system at time t equals the number in the system at time t_0 plus the difference between the number entering the system (between times t_0 and t) and the number leaving the system (between times t_0 and t). Expressed mathematically, $S(t) = S_0 + I(t) - O(t)$. (See footnote 3.)

System Travel Time

The number of vehicles in the system yields many interesting insights when considered over a period of time. If this number is known for each time, t, either as a graph

---

[2] $I(t) = \int_{t_0}^{t} i(\tau)d\tau$ = cumulative input from time t_0.

$O(t) = \int_{t_0}^{t} i(\tau)d\tau$ = cumulative output from time t_0.

[3] $S_0 = S(t_0)$ = number of vehicles in system at time t_0.

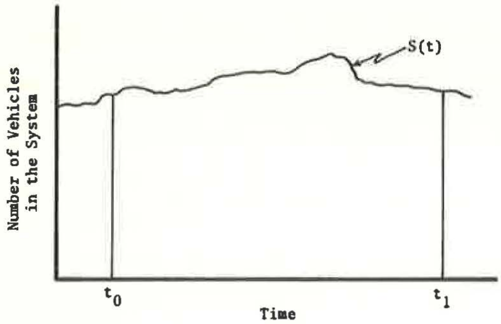


Figure 1. Number of vehicles in the system vs time.

or as a function, S(t), some interesting and important results can be obtained. Figure 1 shows the number of vehicles in a system versus time. The area under this curve between two times, t_0 and t_1, is the number of vehicle-minutes (if time is in minutes) accumulated in the system during this time. For the time period considered, this is the total "time of occupancy" (1) accruing to all vehicles while they are in the system. If there is no parking in the system, the "time of occupancy" is also the total travel time. The travel time in the system could also have been obtained by integrating S(t) between t_0 and t_1, i.e., travel time in the system from t_0 to t_1 = $\int_{t_0}^{t_1} S(t)dt$.

From this it can be seen that the total travel time in a given system can theoretically be determined from volume measurements alone. The number of vehicles in the system at the time t_0 (beginning of the period of interest) must be known. From that time until the end of the period an accurate record of all the vehicles entering and leaving the system must be known—according to time increments (such as one minute). For each time period then, the number of vehicles in the system, S(t), is known. This function can be integrated graphically or numerically to yield the total travel time in the system during the period of interest.

The preceding discussions were general and apply to any system or subsystem.

EQUIVALENCY OF MINIMIZING TRAVEL TIME AND MAXIMIZING SYSTEM OUTPUT RATE

The purpose of this section is to determine a criterion function or figure of merit to be used to guide in the operation of a freeway control system. Such a criterion function is essential in that it provides both a goal to strive for and a means of evaluating the operation of the freeway system. To optimize system operation, it is necessary to know what is to be optimized. Ideally, such a criterion function should be easily understandable and amenable to continuous measurement in the system.

Traditionally, travel time has been used to evaluate operation in transportation systems, subsystems and individual links. Because most drivers are trying to go between their origins and destinations in the shortest time, there is little doubt that the total street and freeway system travel time is a good figure of merit in examining operation of the entire system. Subject to the constraints of safety, comfort, convenience, etc., travel time in the system would ideally be minimized.

The use of total travel time in the evaluation of the operation of a subsystem or a link may not be quite so meaningful. In these, the total travel time accumulated in a given time period can be reduced by decreasing the per-vehicle travel time or by decreasing the number of vehicles using the facility during the time period. Even the average travel time on a link can be decreased by decreasing the volume on the link. This can be done by diverting some vehicles but, in this case, the problem may merely be shifted to adjacent links or subsystems. Total travel time is meaningful when no attempt is made to alter the input. Thus, the following discussions are limited to any system or subsystem for which the input rate is unaffected by congestion internal to the system. As pointed out previously, these are the only cases in which travel time comparisons are meaningful.

It was shown previously that the total travel time in a system or subsystem is the integral over time of the number of vehicles in the system. That is to say that, for the period from t_0 to t_2, the total travel time = $\int_{t_0}^{t_2} S(t)dt$. Since $S(t) = S_0 + I(t) - O(t)$, the total travel time $= \int_{t_0}^{t_2} [S_0 + I(t) - O(t)]dt = S_0(t_2 - t_0) + \int_{t_0}^{t_2} I(t)dt - \int_{t_0}^{t_2} O(t)dt.$

Since S_0, t_2 and t_0 are constants, the first term of the travel time expression is also a constant. On a given day, the cumulative input to the system, $I(t)$, is a fixed function of time and is assumed unchanged by any controls which are exerted on the freeway system. Thus, the second term in the travel time expression is also a constant. Hence, the total travel time = a constant $- \int_{t_0}^{t_2} O(t)dt$. This is the total travel time accumulated in the system under consideration in the time period from t_0 to t_2.

One objective of control or operation of a transportation system is to minimize the total travel time accumulated in the system in a certain critical period of time. Due to the previous considerations, for a fixed input function, minimizing the total travel time is equivalent to maximizing $\int_{t_0}^{t_2} O(t)dt$.

Figure 2 shows a hypothetical cumulative system output function plotted between times t_0 and t_2. If the time period under consideration is defined properly, there will be no appreciable congestion in the system at either time t_0 or t_2. The total system output in the time period, then, will be equal to the total demand for the period and will be a fixed value as shown in Figure 2. Since $O(t)$ is the cumulative system output, $\int_{t_0}^{t_2} O(t)dt$ is the area under this output curve in the period t_0 to t_2. The problem then becomes one of maximizing the area under a curve which passes through zero at t_0 and through a fixed point at t_2.

There are two major constraints which are placed upon this maximization. There is an upper limit on the slope of the curve because the slope of the curve is the output rate of the system. Hence, the maximum possible slope of the curve is the capacity rate of output. The height of the curve or the cumulative system output can never exceed the cumulative system input by more than the number of vehicles originally in the system.

Within these constraints, the area under this curve would be maximized by, starting at time t_0, maximizing the slope of the curve at each instant of time. If one were considering discrete time intervals, the object would be to increase the height of the curve by as much as possible in each time interval. The traffic interpretation of this is that in most cases the control strategy of maximizing the output rate at each moment of time (or the output in a given time period) is equivalent to minimizing the total travel time in the system for a fixed-system input function.

Several conclusions can be drawn from these considerations. In most cases in which the system input is not affected by controls or other system changes, maximizing the output rate at each moment of time is equivalent to minimizing the travel time in the system. If the controls or other changes cannot increase the output rate at some time during the period of interest, no decrease in travel time can be produced. Also, what happens within the system is of importance to total travel time only as it changes the output characteristics. Hence, in an analysis of this type, such items as speed are of importance (to system travel time) due only to their effect on the output rate of the system.

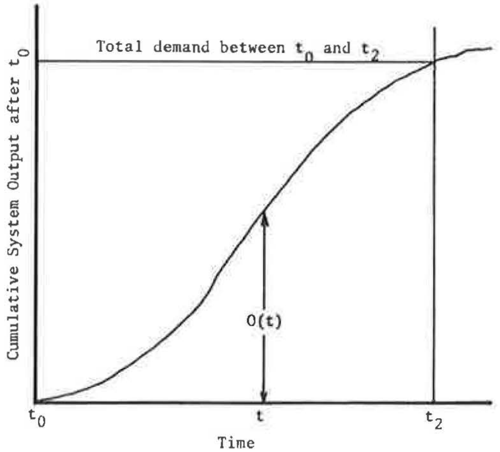


Figure 2. Cumulative system output vs time.

So far, all discussions have been made for large systems. Most of the same discussions hold for smaller subsystems or individual facilities, but a little more care must be exercised so that some of the assumptions are not violated. In some instances it is possible that an increase in the output rate of a freeway subsystem (a one-directional length of freeway with some on-and-off ramps) can cause downstream congestion which will restrict the output rate later, resulting in an increase in system travel time. (This is partly a problem of proper system selection.)

It is also possible that the optimal method of operation of the freeway and surrounding arterial streets is to change the freeway input rate by diverting some traffic. This paper concerns itself primarily with determining the optimal operation of a freeway subsystem for fixed inputs. The effect of this operation on the arterial streets is not considered. The philosophy which has been adopted in this paper is to accept the demand for freeway use as it occurs (or at least to store the excess demand for a period of time until it can be allowed on the freeway without causing congestion) and operate the freeway system optimally for that demand. This results in a minimum total travel time for all vehicles which demand use of the freeway in the time period being considered. It is not necessarily the optimal operating procedure when the surface street operation is considered along with the freeway operation. However, optimal operation of the freeways will almost certainly produce a substantially improved operation of the overall system.

EFFECT OF CONGESTION

There is considerable evidence which indicates that the development of congestion at a permanent freeway bottleneck can decrease the flow rate there (2 through 5). Since congestion can decrease the output rate of a freeway system and since one objective of freeway operation is to maximize the system output rate, another goal of freeway operation is to prevent congestion on the freeway. This is consistent with the overall objective of minimizing system travel time.

TYPE OF CONTROL

Criteria for Controls

The controls must be able to prevent or alleviate congestion in freeway system or subsystem without causing inefficiently low flows on the freeway. They must be flexible enough to respond to traffic conditions in the system. They must also be quite positive and firm so that the desired traffic behavior can be obtained and so that a given result can definitely be associated with a given control action. The controls must also be acceptable to the drivers, must not be hazardous and must fall within the limits of economic feasibility.

Entrance Ramp Metering

After considering many types of special peak-period freeway controls, ramp metering seemed to best meet the criteria which were established for such controls. Each of the other types of controls had some distinct advantages in some particular situations, but in the general situation, ramp metering seemed to hold the most promise.

Entrance ramp metering is a system by which the maximum flow rate at each entrance ramp is set based on freeway traffic conditions. This is done by releasing vehicles from the ramp to the freeway at the chosen time headways. In this way it is at least theoretically possible to maintain but not exceed the merging capacity. Ramp metering has been tested by the Congress Expressway Surveillance Project (6).

The possible benefits of metering are as follows:

1. Reduction of freeway congestion,
2. Increase of merging capacities,
3. Making merging maneuver easier for ramp vehicles, and
4. Diversion of some short trips from the freeways due to the time delay caused by the metering.

One potentially serious problem is the storage of queued vehicles on the ramps. This problem must be given a great deal of consideration in the design of a metering system.

TRAFFIC DETECTION IN A FREEWAY CONTROL SYSTEM

Purposes of Detection

Prediction of Congestion.—The first purpose of the traffic detection for a peak-period control system is to predict traffic conditions that will occur at bottleneck locations while there is still time to take corrective action. The detection system must have the capability of predicting the development of congestion at the bottlenecks so that the controls can be applied to prevent this congestion. The lead time of prediction must be sufficient to allow the ramp metering to respond so as to prevent the congestion.

Indication of Congestion.—Even with a peak-period control system in operation congestion will, at times, develop in the freeway system due to accidents or other unusual events. Congestion must be detected so that remedial controls can be initiated at once to minimize the effect of the unusual event. Thus, the second purpose of the detection system is that of providing an indication of congestion.

Variables to Be Detected.—An understanding of the behavior of each of the numerous freeway traffic variables (volume rate, speed, density, lane occupancy, etc.) prior to and during the peak period is important in the design of the detection system. One or more of these variables must be detected in order to accomplish the two objectives: (a) predicting traffic conditions at critical sections, and (b) indicating congestion.

The portion of freeway shown in Figure 3 is used as a framework for discussion, and a descriptive model of the behavior of the traffic variables in this zone is presented. Section B is the bottleneck or critical section and section U is assumed to be one-half mile upstream of section B. The assumed volume-density curves for these two sections are shown in Figure 4.

The behavior of three variables, volume rate (q), speed and density, is examined. These variables were chosen partially because of this interrelationship, $q = V \times k$. Lane occupancy behaves very much like density and, in fact, it could be considered a time-based density. Because of the similarity of behavior of lane occupancy and density, only density is considered.

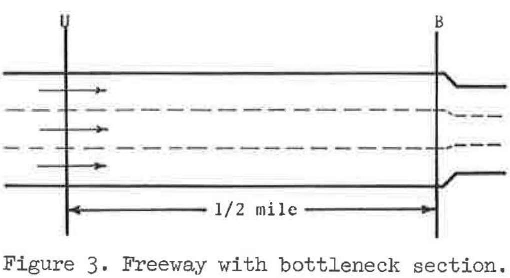


Figure 3. Freeway with bottleneck section.

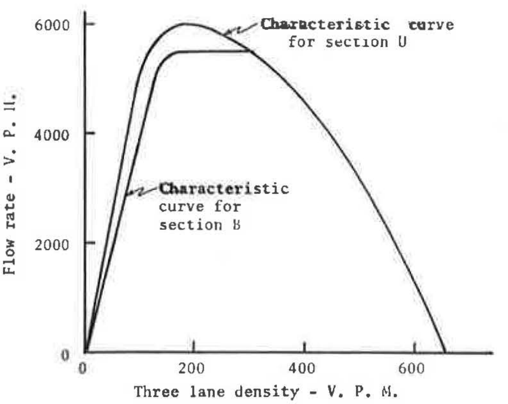
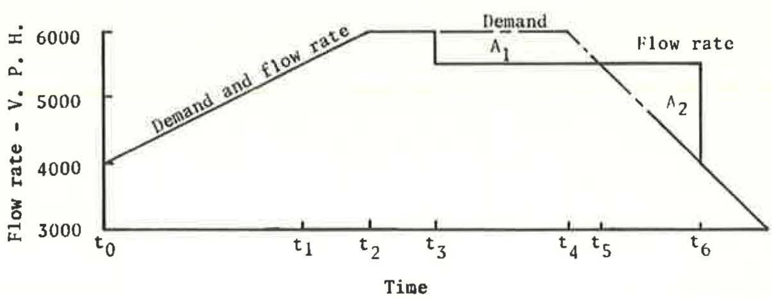


Figure 4. Flow rate-density curves for sections B and U.

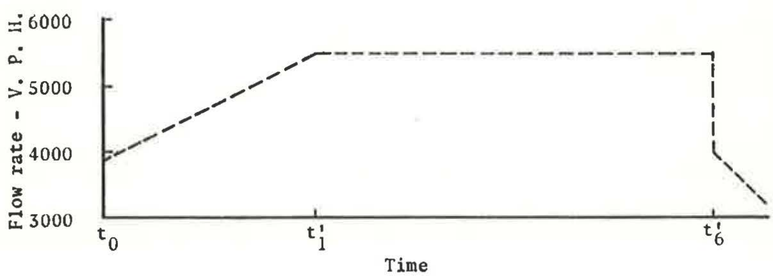
Figure 5a shows the assumed demand rate at section U. The following discussions do not depend on the exact shape of this curve so the somewhat linearized demand curve does not alter the results.

Because of the volume-density curves which were chosen (Fig. 4), the bottleneck capacity is assumed to be 5,500 veh/hr, while the capacity at section U is 6,000 veh/hr. It is also assumed that the density during congestion equals 300 veh/mi so the bottleneck flow rate does not in this case (by assumption) decrease due to congestion. This assumption was made to simplify the example. A similar, somewhat more difficult, analysis could be made in which the density during congestion could exceed 300 veh/mi thereby decreasing the flow rate at the bottleneck. For the purposes for which this analysis is used, however, the assumption that the density during congestion equals 300 veh/mi does not alter the results, mainly because the period of primary interest is that prior to congestion.

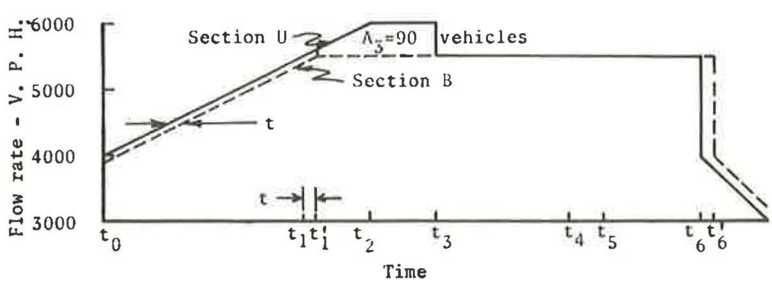
At time t_0 the volume rate at section U equals the demand rate of 4,000 veh/hr while the speed is 50 mph and the density is 80 veh/mi. The volume rate is also



a. Flow rate at section U versus time.



b. Flow rate at section B versus time.



c. Flow rates at sections U and B versus time.

Figure 5. Flow rates at two freeway sections vs time.

increasing because the peak period (it is assumed) is approaching. (Short-time variations in flow rate are not considered, only the overall volume rate trend which is increasing.) Due to the distance ($^1/_2$ mi) between sections U and B, there is a time lag between an increase (or decrease) at B. During the volume buildup prior to congestion, the volume rate curve at B will parallel the corresponding curve for section U but will lag by a time t (Fig. 5c). At time $t_0 + t$ the volume rate at B equals 4,000 veh/hr, the speed equals 40 mph and the density equals 100 veh/mi. Figure 6a and b shows the speeds and density plots at the two locations.

At time t_1 the volume rate at U reaches 5,500 veh/hr, the bottleneck capacity. The corresponding flow rate at section B is slightly lower than this (Figure 5a and c). The speeds at U and B are, respectively, 48 and 37 mph (Figure 6a) while the densities are 120 and 170 veh/mi, respectively, at U and B.

A short time after t_1, the flow at U exceeds 5,500 veh/hr. It was seen previously that the input rate minus the output rate equals the storage rate. In this case if the input exceeds the output capacity there is certain to be a positive storage rate.

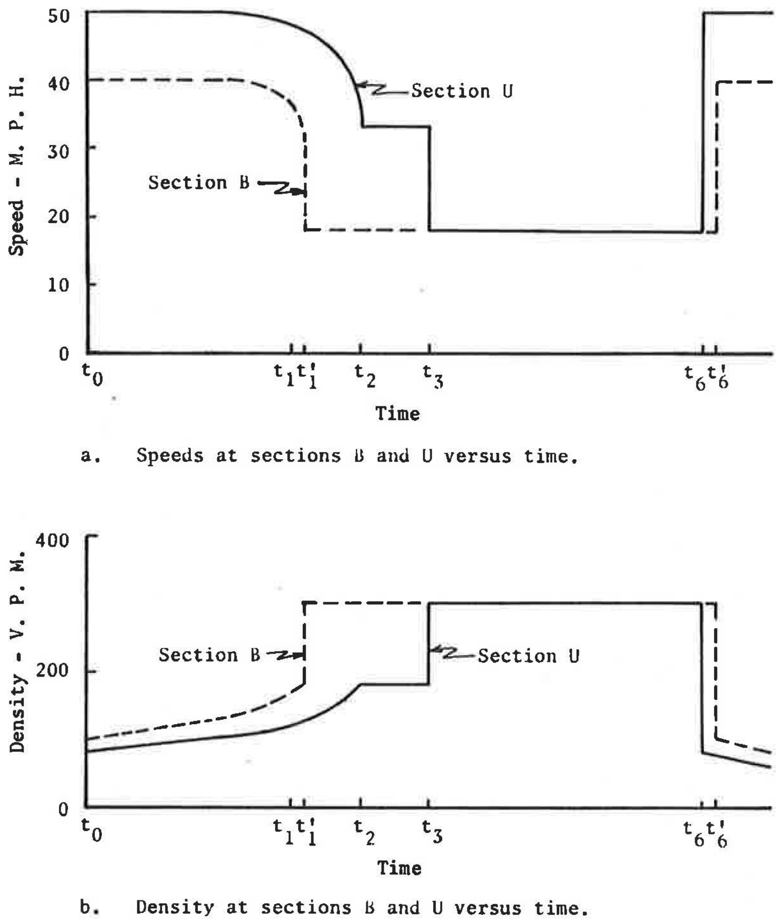


Figure 6. Speed and density at sections B and U vs time.

At time t_1' (t_1 + t) the volume rate at B reaches the capacity flow rate there. At this time storage or queueing occurs upstream of B and the density at B climbs to the (assumed) congested density of 300 veh/mi. The corresponding (Fig. 6a) speed at B decreases to 18 mph. The behavior of the speed and density at time t_1' at the bottleneck can be seen in Figure 6a and b (dashed lines). At this time, which is the beginning of congestion at B, the speeds drop sharply and the density rises sharply. The volume rate, however, remains constant after time t_1'. Thus, a measurement of volume alone would be a poor indicator of congestion, whereas speed and density are quite sensitive to congestion. Since lane occupancy is similar to density, it too, would be very sensitive to congestion. Hence, speed, density or lane occupancy could be used as indicators of congestion.

Due to the time lag between vehicles passing an upstream section and the same vehicles passing a bottleneck, the traffic behavior at the upstream section should provide predictive information on what is going to occur at the bottleneck. (To keep this description relatively simple, a constant time lag, t, is used; in reality, however, the time lag would vary, depending on the speeds in zone UB.)

It has been shown that when the volume rate at U exceeded the capacity flow rate at B, storage had to take place. Also, the excess of flow rate at U over the capacity flow rate at B led to the formation of the queue at B which caused the density to increase sharply and the speed to decrease sharply there. Thus, this congestion was caused by the flow rate at U exceeding the capacity flow rate at B. The volume rate at U exceeded 5, 500 veh/hr for a time period t prior to the development of congestion at B. Thus the upstream flow rate, when related to the bottleneck capacity, can be used to predict the development of congestion at the bottleneck. High volume rate can be considered the cause of congestion, whereas high density and low speed can be considered the effects of congestion.

At t_1 there are also certain values of speed (48 mph) and density (120 veh/mi) at section U which correspond to the 5, 500-veh/hr volume rate. It could be argued that an average speed of less than 48 mph or a density greater than 120 veh/mi could also be used to predict congestion at B. However, in this case speed and density are used to predict volume and would have no significance in themselves. A speed or density at U cannot be specifically related to a present or future speed or density anywhere else. The continuity equations hold true only for volumes. Indeed, Barker (7) studied the propagation of discontinuities of volume, speed and density and found that under free-flow conditions discontinuities in all three variables were propagated downstream. These discontinuities were followed through a series of detector stations about 400 ft apart. For a given variable, the downstream propagation of "waves" or discontinuities is a necessary but not a sufficient condition for use as a predictor of congestion at a downstream location. It is also necessary that the variable satisfy the continuity equations. These equations can be written for volume (or volume rate) for a closed system, namely the input equals the output plus storage. For a "straight pipe" length of freeway (such as in Fig. 3) this means the volume past the upstream section equals the volume past the downstream section plus the additional number of vehicles between the sections. Similar equations cannot be written for speed, density or lane occupancy.

Since congestion develops at a bottleneck when the upstream flow rate exceeds the bottleneck capacity, it seems even more logical to use the upstream flow rate to predict congestion. Capacity values of speed and density of lane occupancy have not yet been developed.

For the preceding discussions, one-to-one transformations between speed or density and volume rate were assumed—meaning that a given speed or density has one and only one corresponding volume rate. In reality, however, a range of volume rates would be associated with a fixed average speed or density and this would make prediction of volume rate from other variables more difficult. The relationships between volume rate and speed or density could also change from one upstream location to another, further contributing to the problem of predicting volume rate from the other variables.

It is also interesting that noncongested flow prevailed at the upstream location U after the bottleneck (section B) flow became congested. The time during which this holds true equals the time for the rear of the queue to reach the upstream location. The speed of the rear of the queue is a function of (a) the change in density from noncongested to congested operation, and (b) the storage rate. The storage rate equals the flow rate at U minus the flow rate at B. Hence, the speed of the rear of the queue equals $(q_u - q_B)/(k_B - k_u)$. On the volume rate-density curve, this would be the slope of a vector drawn between the operating points for sections U and B, as shown by Lighthill and Whitham (8).

If the density in zone UB were 120 veh/mi at time t_1', a total storage of 90 vehicles would be required to place the rear of the queue at U. This is because it was already assumed that the steady-state congested density is 300 veh/mi and zone UB has ½-mi length. Thus, the time T between the development of congestion at sections B and U is such that $\int_{t_1'}^{t_1' + T} (q_u - 5,500)dt = 90$. This is the area A_3 in Figure 5c. The queue reaches U at time t_3 ($t_3 = t_1' + T$).

At time t_3 zone UB is in a steady-state condition. That is, the density in zone UB is a constant 300 veh/mi. This requires that $q_u = q_B$ = 5, 500 veh/hr. Thus at t_3 the volume rate at U drops sharply from 6, 000 to 5, 500 veh/hr, the speeds drop sharply from 33 to 18 mph and the density increases from 180 to 300 veh/mi. At time t_6 all the congestion clears at U and at time t_6' it clears at B.

This analysis assumed that the flow rate at the bottleneck did not decrease due to congestion. If this assumption were not made it would mean that the steady-state congested operating conditions could be found in Figure 4 at the point corresponding to the congested density. Speeds and volume rates during congestion would be lower since the density is now greater than 300 veh/mi. Since the rear of the queue would travel upstream faster, the time required to travel the ½-mi distance would be less than T. All of these differences occur after congestion has set in and have no effect on the conclusions regarding the prediction of congestion.

This descriptive model was used to explain the behavior of three traffic variables—volume, speed and density—prior to and during congestion. Since it is a descriptive model, several assumptions can be made to facilitate the discussions. Probably the most important of these is the assumption of a known, fixed bottleneck capacity. Many things can change the capacity of a bottleneck or cause a bottleneck where none previously existed. The effect on the capacity of a freeway roadway of a disabled vehicle, an accident, adverse weather or other factor is an area needing a great deal of research. Perhaps, there is no one fixed capacity at a given location but rather there may be a probability of congestion developing due to a given flow rate. This probability may also change with different drivers, weather, etc.

The detection system which has been described is, perhaps, somewhat idealized and no consideration has been given to the economy of such a system. Such considerations might require that some compromises be made. For example, it may be necessary to sample a variable, such as volume, in one lane to estimate the variable across all lanes. It may also be necessary to estimate volume from speed, density or lane occupancy. In this way a one-variable system could possibly be developed. The accuracy of this sampling and estimating one variable from another should be carefully examined to determine if such procedures can be used successfully.

LOCATIONS OF DETECTORS FOR PREDICTION OF TRAFFIC CONDITIONS AT A SECTION

The traffic detection system must be capable of determining the proper volume rates to be allowed to enter at each entrance ramp in order to keep the freeway system operating in the best possible manner under existing conditions—either normal or reduced capacity operation. The volume rate on each ramp can be determined by the capacity of the merging section or by a downstream bottleneck. In either case, the detection of the upstream freeway volume can be used to determine the maximum allowable ramp volume. The metering rates at the ramps can then be set to allow no more than the predetermined rates of flow on the ramps.

The preceding section discussed a time lag between a certain volume passing an upstream location and its passing the bottleneck location. The purpose of this section is to examine this lag time with respect to other critical times of detection and control. The situation considered is that of a metered entrance ramp. Vehicles are detected

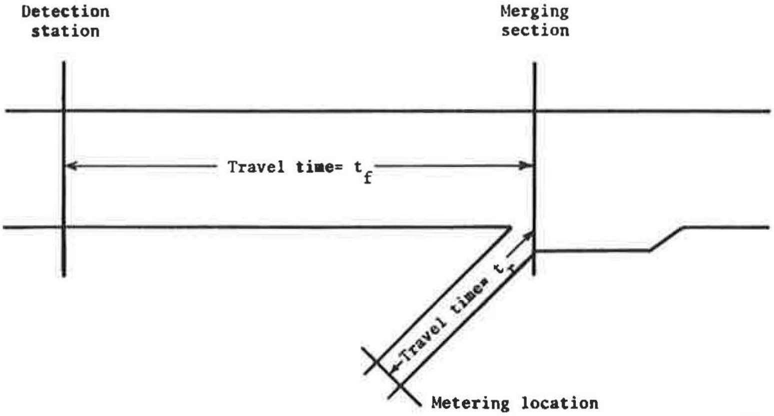


Figure 7. Schematic of metering and detection locations.

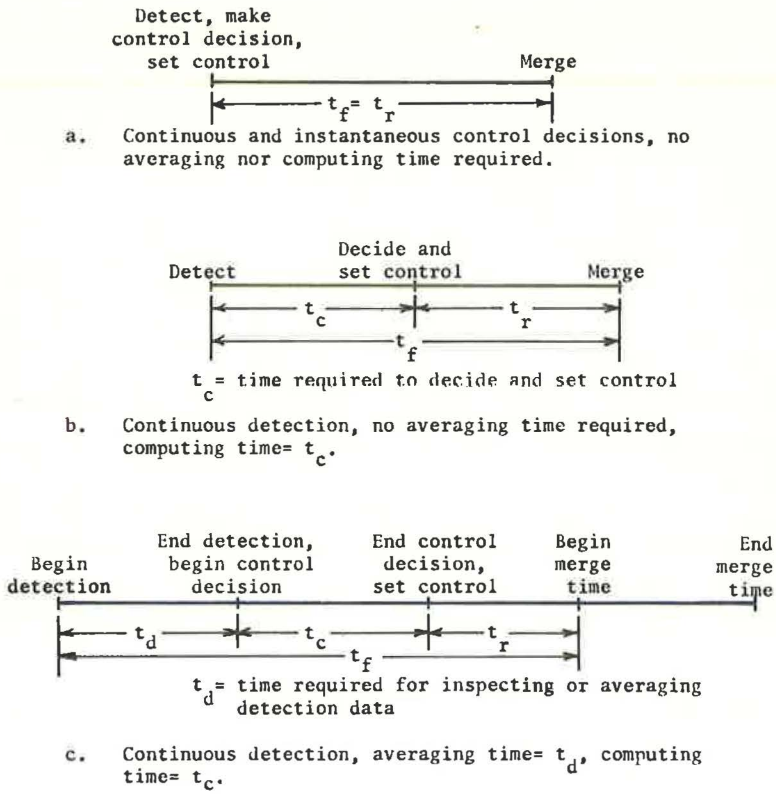


Figure 8. Relationships between required times for a detection station location.

upstream of the entrance ramp, and by metering, the total merging rate can be kept below the merging capacity. The problem is the determination of the time separation of the detection station and the merging section which is required to allow enough time for all of the control decisions and actions to take effect.

Figure 7 shows a schematic of the portion of freeway and ramp used for this discussion. The detection station is separated in time from the merging section by a travel time, t_f, on the freeway. The metering device on the ramp is temporally separated from the merging section by a travel time t_r. The attempt here is to determine the freeway travel time, t_f, which is required for the successful operation of the ramp metering.

The first and simplest case considered is one where the metering decision is made simultaneously with the detection. Thus, at each instant of time the control decision is made according to what is simultaneously detected. The purpose of the control is to control the volume rate of vehicles crossing the merging section. It has to know how many vehicles will be merging at any given time. This requires that the travel time between the detection station and the merging section equal the travel time from the metering device to the merging section, i.e., $t_f = t_r$ (Fig. 8a). If $t_f < t_r$ the metering decisions are made too late (according to what has already merged on the freeway). Hence in no cases should the upstream freeway detectors be located so that $t_f < t_r$.

In the second case considered, the control decision and control adjustment (if any) are assumed to take some amount of time, t_o. Figure 8b also illustrates the time relationships in this case. The time on the freeway between detection and merging is again t_f. After detection on the freeway, the decision and control time t_c must elapse before the control change takes place. Then a time t_r on the ramp is required before the merge takes place. In order that the detected freeway vehicles merge with the same ramp vehicles which had their control based on freeway vehicles' detection, $t_f = t_c + t_r$.

In all probability control decisions would not be made on instantaneous happenings at any detector location. It is more likely that the results of detection in a length of time would be used as a basis for control decisions. Short observation times would probably produce quite erratic results (for example, for a 1-sec time period the detection of no vehicles corresponds to a 0 veh/hr flow rate while the detection of one vehicle corresponds to 3, 600 veh/hr). Lengthening the observation time or averaging over a longer time would damp out much of the random fluctuation.

In the third case considered, the effect on the required freeway travel time of the time period, t_d, is used for detection or averaging of detector data. Figure 8c shows the time relationships in this case. A time t_d is required from the beginning to the end of the detection period. When the detection period is over a time t_c is needed to assimilate the data, make a control decision, and initiate the control. After a time t_r the first ramp vehicles released under the new control reach the merging section. They should be merging with the first vehicles to be detected on the freeway at a time t_f earlier. Hence, in this case $t_f = t_d + t_c + t_r$. For example, if the detection and averaging time is 45 sec, computation time is 15 sec and ramp travel time is 10 sec, t_f has to be greater than or equal to 70 sec.

It was seen that extra time required between the start of detection and the adjustment of the controls produces an increase in the required freeway travel time between the detection station and the merging section. Other time requirements would similarly increase t_f.

So far the considerations of various time requirements have been used to determine t_f. A fixed t_f or an upper limit on t_f can also be used to establish limits on other times. For example, if $t_f = 45$ sec and if $t_r = 10$ sec, $t_d + t_c$ must be less than or equal to 35 seconds.

Because congestion develops at a bottleneck and is propagated upstream, the ideal location for prompt detection of congestion would be at or slightly upstream of a bottleneck. As the distance upstream of the bottleneck increases the time lag between the development and detection of congestion also increases. Again this location immediately upstream of the bottleneck is the ideal location of the detectors. For reasons of economy it may be necessary to use the same detectors for prediction of traffic behavior and for detection of congestion. In some cases, one detector station can be located at one bottleneck to detect congestion and could also be used to obtain volume data for predicting traffic behavior at a downstream location.

In summary, it appears that short-period volumes upstream of a bottleneck provide the best prediction of impending congestion and are probably the best variable to measure for control purposes. Volume alone cannot be used to differentiate between congested and noncongested operation. Hence, speed, density or lane occupancy must also be measured at or upstream of the bottleneck in order to provide this information.

It is fortunate that volume is one variable to be measured for control purposes. It was previously indicated that maximizing the system output volume rate will lead to optimal operation of the system. Volume is also the only variable for which continuity equations can be written. Volume measurements are susceptible to point measurements that are much easier to accomplish than measurements over a length of roadway. The volume rate is also the most easily controllable variable and volume capacities can be established. Individual time headways at the input sources can be controlled and there is a one-to-one relationship between time headways and volume rate. Other advantages will appear later.

OPTIMAL OPERATION OF A FREEWAY SUBSYSTEM

The subsystem considered here includes storage areas for vehicles waiting to enter the freeway. Several entrance and exit ramps increase and decrease the volumes along the freeway. Figure 9 is a schematic of the freeway system (a portion of the westbound Congress Street Expressway in the Chicago area) which is used for the development of a prototype model. Four lanes of traffic enter at the Cicero Ave. end of the system and three lanes exit downstream of the Des Plaines Ave. entrance ramp. The demand at the (Cicero Ave.) freeway input source is not controlled, and thus, is one further re-

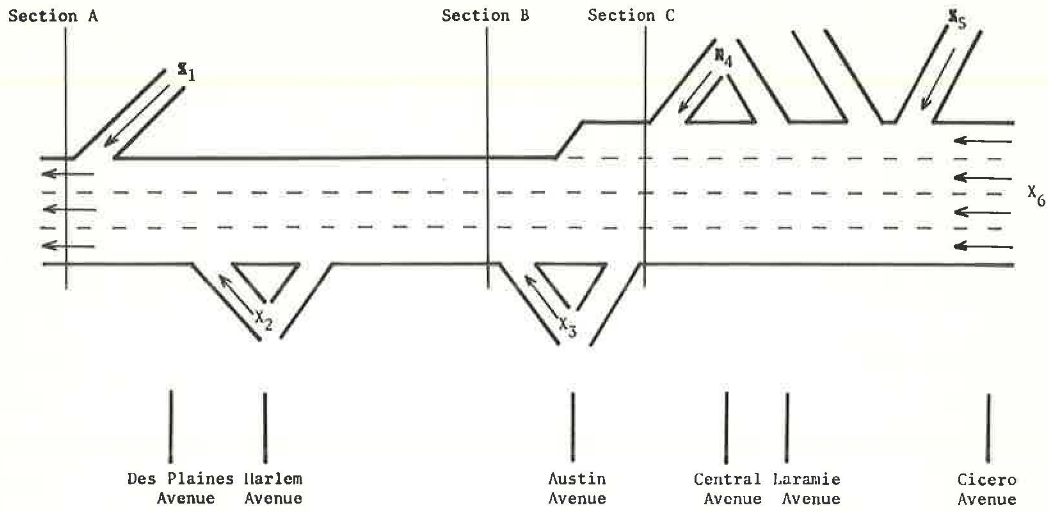


Figure 9. Schematic of freeway subsystem used in the development of the prototype linear programming model.

striction to the problem. Generally such a restriction will not affect the optimality of the operation, however. As used in this section, input refers to flows onto the freeway roadway and is a different use of the term than was made previously.

The controls which are considered are entrance ramp metering controls which can limit the inputs to the freeway from the various entrance ramps. The mathematical model yields the volume which is required or permitted on each of these ramps during the analysis period in order to obtain optimum performance of the portion of the freeway system under consideration. Metering devices are assumed to be in operation on the ramps to limit the ramp flow rate to the required level. Besides limiting the ramp flow rate, the metering system also serves to damp out large variations in demand on the entrance ramps. Hopefully, this will lead to smoother merging operations in the vicinity of the ramp and perhaps to higher merging capacity rates.

The objective function which was selected for optimization is the output of the system in a given time period. This is, of course, to be maximized. Since it is assumed that congestion can decrease the flow rate at a bottleneck, the controls must be operated so as to prevent the development of congestion at all bottleneck locations in the system in order to keep the output rate at its maximum level. Hence, critical points or potential bottlenecks in the section must be identified so that the demand on these sections can be kept below capacity levels.

The flow rate upstream of each bottleneck section must be so controlled as to prevent the development of congestion. The model which is developed tells how much flow from each source can be accommodated under these conditions during the analysis period. (The actual ramp flow rates would not be constant during the period.) The remainder is the amount which must be diverted or stored until the time period is over. The entire demand from the freeway input source (in this case at the Cicero Ave. four-lane section) is accepted into the system.

Origin-Destination Information

Since the volume at each critical or bottleneck section is composed of vehicles from several origins, the effect of altering the flow at one or more of the origins must be known. Therefore, for each input the percentage of its inflow vehicles crossing each critical section must be known.

Existing O-D data (9) for each entrance ramp provided part of this information. For the peak period at each entrance ramp the percent of entering vehicles destined for each of the exit ramps was available for the system of interest. These data are shown in lines 1 to 5 of Table 1.

TABLE 1

ORIGIN-DESTINATION DATA FOR THE FREEWAY SUBSYSTEM (FIG. 9)

Input Source	% Destined for:				
	Laramie	Central	Austin	Harlem	Through
1. Des Plaines	-	-	-	-	100.0
2. Harlem	-	-	-	-	100.0
3. Austin	-	-	-	5.1	94.9
4. Central	-	-	-	6.7	93.3
5. Cicero	0.9	2.2	4.7	9.8	82.4
6. Cicero	13.7	8.6	15.8	10.0	51.9

When combined with volume counts of vehicles entering and leaving the system, both on the mainline (the freeway roadway) and on ramps, these data were used to determine the destinations of the vehicles entering at Cicero Ave. on the freeway. At each output location, the total output volume during the period was known. From the input volume data for each location and the knowledge of the percent of this volume destined for each output location, it was possible to estimate, for each output, the volume coming from all the input sources except the mainline input. When this subtracted from the total volume at a particular output, the number of vehicles at the output coming from the mainline input is determined. For example, assume that the Laramie Ave. exit ramp volume (for the period considered) was 700 vehicles and the Cicero Ave. entrance ramp volume was 1,000 vehicles. It is known that 0.9 percent of the vehicles which enter the freeway from the Cicero on-ramp exit at Laramie Avenue. This means that the expected number of vehicles entering at the Cicero ramp and exiting at Laramie during the period is 9. Thus the other 691 exiting vehicles must have come from the freeway input. Since the total freeway input is known, the percent of this volume leaving at Laramie can be computed. The calculated percents of vehicles entering on the freeway at Cicero and destined for each of the outputs are shown in line 6 of Table 1.

Deterministic Linear Programming Model

The deterministic linear programming model discussed here optimizes the operation of a freeway subsystem or system subject to several constraints. The period considered in the optimization is one hour and it is selected for several reasons:

1. The ends of the section are temporally separated by approximately 6 to 8 minutes (the travel time from one end to the other); hence, extremely short time periods are not satisfactory.
2. It is a convenient time period for many data measurements.
3. It is a period that is shorter than the period of congestion in the specific instance under consideration.

Even at locations at which the congestion lasts for a shorter time, demand can be conveniently expressed as hourly rates.

All volumes, capacities, and demands used in the model are for a 1-hr period. This assumption can be, as it turns out, quite revealing. If the capacity of a critical location that now regularly experiences congestion is determined and the corresponding capacity restraint is not exceeded in the linear programming model, it simply means that metering the flows would have prevented congestion at this location and that the entire ramp demand would be satisfied within the hour. In other words, if the hourly capacity constraint is not exceeded, the congestion is caused by short-time surges of traffic or a downstream bottleneck.

Objective Function.—The model has as its objective the maximization of the output of the system of interest in the time period considered. The output is considered to be the volume leaving the system via the freeway mainline output and all of the exit ramps. The number of vehicles entering a closed freeway system during some time period equals the number leaving plus the number stored in the system during the same time. Thus, the input to the system equals the output plus storage. The storage rate can be positive or negative depending on whether vehicles are entering or leaving storage. As congestion develops and as congestion diminishes the absolute value of the storage rate increases. During steady-state conditions, either in free flow or congestion, the change in storage approaches zero, i.e., the number of vehicles in the system over the time period remains relatively constant. Since the prevention of congestion is incorporated into the linear programming model, the change in storage is considered to be zero. Hence, the input of the system is equal to its output, so input can be substituted for output in the criterion function. The objective of the model, then, is to maximize the input to the system. This can be interpreted on an intuitive basis; the model is maximizing the number of vehicles which can enter the system without encountering congestion before leaving. The variables of the model are the volumes at the input sources.

[It will be noticed that two different systems have been discussed so far. The first is the system in which the travel time is being minimized. This is the system consisting of the freeway and the areas where vehicles wait to be allowed to enter the freeway. The second system, which is considered in the linear programming model, consists of only the freeway and the entrance ramps up to the metering devices; thus, it does not include the waiting areas. If we call the first system the larger system and the second system the smaller system, the two are related as follows. (See Fig. 10.) We have seen that minimizing the travel time in the larger system is equivalent to maximizing the output (rate) in the larger system. Since the outputs of the larger and smaller systems are identical, this is equivalent to maximizing the output of the smaller system. If no congestion develops in the smaller system the input equals (very nearly) its output. Hence, maximizing the input (rate) to the smaller system without causing congestion is equivalent to minimizing the total travel time in the larger system.]

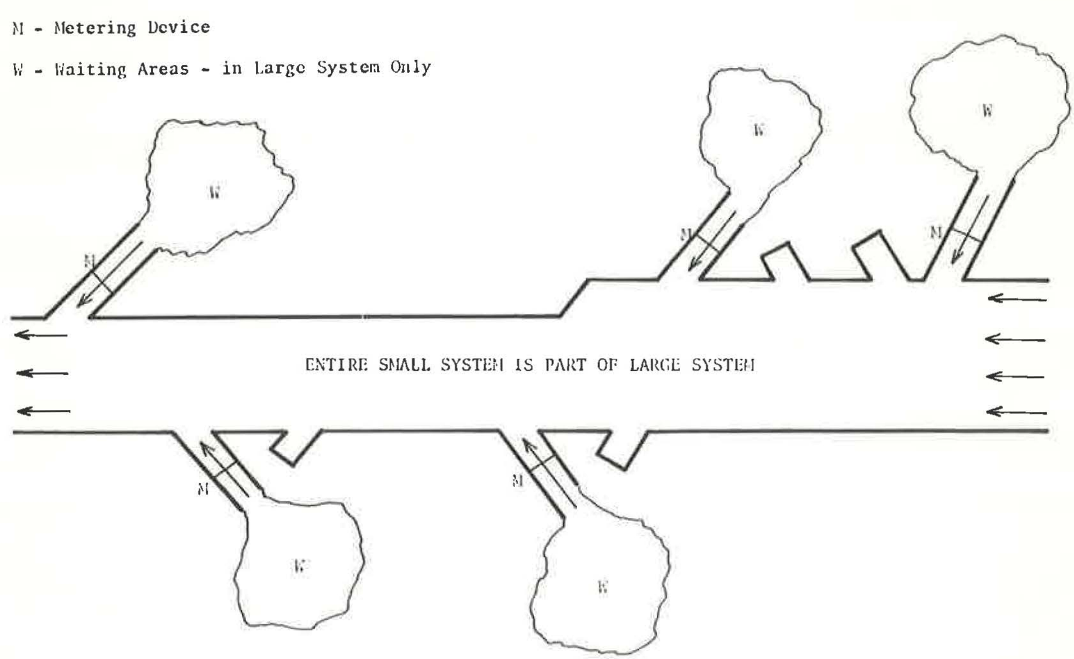


Figure 10. Two systems considered in the analyses.

Constraints.—The model has two types of restraints. Since the development of congestion at many freeway sections reduces the flow rate at these sections, one of the restraints is that congestion will not be allowed to develop at any location on the freeway. Alternately stated, the flow upstream of all critical sections on the freeway will be constrained so that the capacity flow rates at these critical locations will not be exceeded. In applying the model for normal operating conditions, constraints have to be established for only those locations which are, a priori, known to be likely sources of congestion.

The second restraint is that there are certain upper limits placed on the input volumes, because there is a limited demand or number of people desiring to use each ramp during any time period. This demand could, however, increase or decrease when a metering system is put into operation. If travel times on the freeway were lowered, the freeway would become more attractive to some motorists. This generated traffic must enter the freeway somewhere, so some ramp volumes could increase. However, some of the increased traffic on a given ramp might have formerly used another ramp in the system. It is also possible that many vehicles will be diverted from the freeway altogether, because of the delays at the entrance ramps which are caused by the metering operation. Thus an increase or a decrease in a ramp's volume is possible. In this model, it is assumed that the maximum demand is known for all ramp inputs as well as for the freeway input.

Statement of Model.—The deterministic model, then, yields the volume at each input source which maximizes the total input to the system subject to two types of constraints. First, a set of constraint equations is required to assure that congestion will not develop at any location. A second set of constraint equations restricts the inputs from each source so as not to exceed the demand at the source.

Development of Prototype of Deterministic Model

Objective Function.—The variables of the linear programming problem are the input volumes from each of the input sources. The variable corresponding to each input is shown in Table 2. The objective is to maximize $X_1 + X_2 + X_3 + X_4 + X_5 + X_6$.

Constraints.—The first set of constraint inequalities, those which require that demand not exceed the free-flow (or possible) capacity at each critical location on the freeway, utilizes the data from Table 1. It is necessary to know the percent of each input volume that will appear at each bottleneck location on the freeway. For example, the capacity at section A is 5,900 veh/hr, so the total demand at this section must be kept lower than 5,900 vehicles for the hour. The percent of vehicles crossing this section from each input is listed in the "through" column (Table 1). This volume is the freeway mainline output. One hundred percent of the Des Plaines and Harlem ramp traffic crosses this section, while 94.9 percent of the Austin ramp traffic, 93.3 percent of the Central traffic, etc., pass through this bottleneck. Expressed decimally, the first constraint is $1.00\ X_1 + 1.00\ X_2 + 0.949\ X_3 + 0.933\ X_4 + 0.824\ X_5 + 0.519\ X_6 \leq 5{,}900$. This assures that congestion will not develop at the Des Plaines Avenue entrance ramp merging section.

Another potential bottleneck location is the merge of the Austin Ave. entrance ramp (section B in Fig. 9); the capacity here is 6,000 veh/hr. All of the Austin and Central entrance ramp traffic crosses this section, while only the portion of the Cicero ramp and freeway input traffic which does not exit at Laramie, Central or Austin would pass section B. For the freeway input traffic, 38.1 percent leaves the freeway without reaching section B (13.7 percent at Laramie, 8.6 percent at Central, and 15.8 percent at Austin). The remaining 61.9 percent of this traffic

TABLE 2

VARIABLES AND CORRESPONDING INPUT VOLUMES

Variable	Represents Input Volume at
X_1	Des Plaines entrance ramp
X_2	Harlem entrance ramp
X_3	Austin entrance ramp
X_4	Central entrance ramp
X_5	Cicero entrance ramp
X_6	Cicero mainline

TABLE 3

MAXIMUM HOURLY DEMAND AT EACH INPUT

Input	Variable	Max. Hourly Demand
Des Plaines entrance ramp	X_1	600
Harlem entrance ramp	X_2	475
Austin entrance ramp	X_3	450
Central entrance ramp	X_4	500
Cicero entrance ramp	X_5	825
Cicero mainline	X_6	6, 800

($0.619\ X_6$) crosses section B. Similarly, 92.2 percent of the Cicero ramp traffic ($0.922\ X_5$) passes this bottleneck location. The constraint for the Austin Avenue merge is $1.00\ X_3 + 1.00\ X_4 + 0.922\ X_5 + 0.619\ X_6 \leq 6,000$.

The third and final bottleneck in the freeway system is the approach to the Austin Ave. exit ramp. This location is critical because of the transition from four lanes to three lanes. The capacity of this location (section C) is 6, 450 veh/hr, so the constraint for section C is $1.00\ X_4 + 0.969\ X_5 + 0.777\ X_6 \leq 6,450$.

These three constraints, when met, assure that congestion will not develop at any of the three bottleneck locations.

Another assumption was implicitly made in the formulation of these equations: there is a one-to-one tradeoff between ramp vehicles and freeway vehicles; that is, a ramp vehicle "uses only as much capacity" as a vehicle already on the freeway. If it is determined that this is not true (i.e., that the reduction of ramp traffic by one vehicle would allow more than one vehicle increase on the freeway) this can easily be put into the model provided the tradeoff is a constant. In the three constraints discussed so far, unity has been the coefficient of the variable corresponding to the traffic on the merging ramp. This would have to be changed to the correct value (the number of additional freeway vehicles which can be passed due to a one-vehicle decrease in ramp traffic) in the constraints. The capacity would also have to be increased to compensate for this change since the ramp volume is weighted more heavily. If 1.5 is found to be the correct coefficient, the second restraint (Austin entrance ramp merge) would be rewritten $1.5\ X_3 + 1.00\ X_4 + 0.922\ X_5 + 0.619\ X_6 = C$, where C is the modified capacity.

Table 3 gives the values of the maximum hourly demands at each input location. The constraint inequalities which prevent an input from exceeding the demand are

$$\begin{aligned} X_1 &\leq 600 \\ X_2 &\leq 475 \\ X_3 &\leq 450 \\ X_4 &\leq 500 \\ X_5 &\leq 825 \\ X_6 &\leq 6,800 \end{aligned}$$

Statement of the Prototype Model.—The prototype model, which maximizes the input to the system subject to constraints which (a) assure that congestion will not develop, and (b) assure that ramp volumes do not exceed the demand, is stated as follows:

$$\begin{array}{lcccccccccccl} \text{Maximize } X_1 & + & X_2 & + & X_3 & + & X_4 & + & X_5 & + & X_6 & & \\ \text{Subject to } X_1 & + & X_2 & + & 0.949\ X_3 & + & 0.933\ X_4 & + & 0.824\ X_5 & + & 0.519\ X_6 & \leq & 5,900 \\ & & & & X_3 & + & X_4 & + & 0.922\ X_5 & + & 0.619\ X_6 & \leq & 6,000 \\ & & & & & & X_4 & + & 0.969\ X_5 & + & 0.777\ X_6 & \leq & 6,450 \\ X_1 & & & & & & & & & & & \leq & 600 \\ & & X_2 & & & & & & & & & \leq & 475 \\ & & & & X_3 & & & & & & & \leq & 450 \\ & & & & & & X_4 & & & & & \leq & 500 \\ & & & & & & & & X_5 & & & \leq & 825 \\ & & & & & & & & & & X_6 & \leq & 6,800 \end{array}$$

This assumes a one-to-one tradeoff between ramp and freeway vehicle in the merge constraints.

TABLE 4

ORIGINAL SIMPLEX TABLEAU

c_i	X_i \ c_j	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	
	X_j	X_1	X_2	X_3	X_4	X_5	X_6	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	b_i
0	S_1	1	1	.949	.933	.824	.519	1									5900
0	S_2			1	1	.922	.619		1								6000
0	S_3				1	.969	.777			1							6450
0	S_4	1									1						600
0	S_5		1									1					475
0	S_6			1									1				450
0	S_7				1									1			500
0	S_8					1									1		825
0	S_9						1									1	6800
$Z_j - C_j$		-1	-1	-1	-1	-1	-1	0	0	0	0	0	0	0	0	0	$Z_0=0$

TABLE 5

OPTIMAL SIMPLEX TABLEAU

c_i	X_i \ c_j	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	
	X_j	X_1	X_2	X_3	X_4	X_5	X_6	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	
1	X_1	1						1		-.933		-1	-.949		.080	.206	447
0	S_2								1	-1			-1		.047	.158	213
1	X_4				1					1					-.969	-.777	367
0	S_4							-1		.933	1	1	.949		-.080	-.206	153
1	X_2		1									1					475
1	X_3			1									1				450
0	S_7									-1				1	.969	.777	133
1	X_5					1									1		825
1	X_6						1									1	6800
$Z_j - C_j$		0	0	0	0	0	0	1	0	.067	0	0	.051	0	.111	.429	$Z_0 =$ 9364

Computational Solution of Prototype Problem. — This problem can be solved quite easily by the well-known simplex computational technique (10). In order to do this, the inequalities must be converted to equalities. This is done by adding a slack variable to each constraint. The set of these slack variables are S_1, . . . , S_9. Table 4 shows the original simplex tableau.

In this tableau, the second row (X_j row) contains the designation of the variables corresponding to the particular column. The rate of change of the criterion function for each variable is located in the top row (C_j row) above the variable. The X_i column contains the basis variables and the rate of change of the criterion function for each of these is located adjacent to them in the C_i column. Hence the subscript designation j refers to any variable, whereas the subscript i refers only to the basis variables.

The slack variables, $S_1, \ldots, S_9$, constitute the original basis since they form an identity matrix. This can be seen in the X_i column since $S_1, \ldots, S_9$ appear in this column. The last column (b_i column) contains the current values of each basis variable. Originally, for example, $S_1 = 5,900$.

The $Z_j - C_j$ row contains the evaluators which are used to determine whether the introduction of a particular variable into the basis will produce an increase in the value of the criterion function. If a particular $Z_j - C_j$ is negative the introduction of the corresponding variable into the basis will produce this increase. As seen in Table 4, any of the six variables (X_1, X_2, X_3, X_4, X_5, X_6) will improve the solution when introduced into the basis.

The value of Z_0 is the value of the criterion function which is produced by the particular set of variables in the basis. Since, in the original tableau, all basis variables are slack variables and do not contribute anything to the criterion function, $Z_0 = 0$.

Six routine simplex iterations (one for each variable introduced into the basis) were required to reach the optimal solution for this model (Table 5). The values of the basis variables which yield the optimal solution appear in the column on the right. All of the non-basis variables equal zero, of course.

Interpretation of Solution.—In the formulation of the linear programming model it is possible to put in constraints which turn out to be redundant. For such equations the slack variables have positive values in the optimal solution. The interpretation given to the value of a slack variable in the optimal solution is the amount that a particular constraint would have to be reduced before it would cease to be redundant (10). In other words it is the excess "capacity" contained in the restrictions.

There are three slack variables in the optimal basis. They are S_2, S_4 and S_7. The slack variable S_2 is for the second equation, which is the capacity restraint at the Austin on-ramp merging section. Since $S_2 = 213$, 213 more vehicles could be passed through section B without developing congestion. The fourth equation states that the volume on the Des Plaines entrance ramp cannot exceed the demand of 600. However, the capacity of section A provides a greater restriction on the Des Plaines volume than this. This merging capacity restricts X_1 (the Des Plaines ramp volume) to 447 vehicles. This is 153 less than the demand on this ramp, so $S_4 = 153$ and there would be an unsatisfied demand of this amount at the Des Plaines ramp. Similarly, since $S_7 = 133$ there would be an unsatisfied demand of 133 vehicles in the hour period at the Central Ave. on-ramp. The capacity at section C provides a greater restriction to the variable X_4 than does the limit which is placed on the ramp demand. The total number of vehicles which must be prevented from entering the freeway during the hour is $S_4 + S_7 = 286$.

The effect on the value of the criterion function of unit changes in constraints is also interesting (Table 5). These are contained in the $Z_j - C_j$ row. The $Z_j - C_j$ at this stage are the optimal values of the dual variables (10). The dual variables are interpreted as the rate of change of the criterion function for a unit change in the corresponding constraint. For example, the dual variable for the first constraint is contained in the $Z_j - C_j$ row of the slack variable for the first equation (S_1). Its value is 1. This means that for a unit increase in capacity at section A, a unit increase in the system output (or input) is realized. The dual variable for the second constraint is zero. Since section B is not now (optimally) operating at capacity, increasing the capacity at this location would merely add to the overcapacity and would not increase the output of the system.

The dual variable for Eq. 3 must be interpreted as an expected value since its value is 0.067. If the capacity at section C were increased by one, an additional vehicle could be allowed to enter the freeway from the Central Ave. ramp without creating congestion at C. However, the vehicle could only be allowed on the freeway if it were going to exit before reaching section A or congestion would develop at that point. Since 6.7 percent of the vehicles entering the freeway at the Central Ave. ramp leave at Harlem Ave., the expected increase in input is only 0.067 vehicle—the value of the dual variable. This analysis indicates that remedial action in this subsystem should begin at the Des Plaines Ave. bottleneck since the value of its dual variable is higher than that of either of the other two bottlenecks.

For both Eqs. 4 and 5 the value of the dual variable is zero. These equations refer to constraints on demand at the Des Plaines and Harlem on-ramps. Since section A is operating at capacity, any increase in demand at either ramp could not be matched by an increase in the output (or input) volume.

Equation 6 placed an upper limit on the volume input at the Austin entrance ramp and the value of the dual variable for these equations is 0.051. This again is the expected value of the increase in system input with a unit increase in demand at this ramp. Of the vehicles which enter the freeway at Austin, 5.1 percent exit at Harlem. Hence, the probability that the additional vehicle would exit at Harlem is 0.051. Since it would be allowed to enter the freeway only if it were not going to pass through section A, the expected value of the increased input is 0.051.

The demand constraint at Central Ave. is redundant so its dual variable is zero. The Cicero ramp and Cicero mainline have dual variables equal to 0.111 and 0.429, respectively. Again these are interpreted as the expected values of the increases in system input for a unit increase in demand at these points. A closer look at the value of 0.111 at the Cicero ramp location is of some interest. The additional vehicle could only be allowed to enter the freeway if it were not going through section A. The probability of being allowed to enter the freeway is 0.176. However, if the vehicle crosses section C it will decrease by one the allowable number of vehicles from the Central ramp. Each vehicle entering at Central Ave. has only a 0.067 probability of being allowed to enter and the probability of a vehicle entering at Cicero Ave. crossing section C is 0.969. Hence, the expected increase in system input with a unit increase in demand at the Cicero ramp is $0.176 - (0.969)(0.067) = 0.111$ vehicle.

The optimal tableau shows a value of 9,364 for Z_0. This means that, in the hour considered, 9,364 vehicles could enter the system of freeway without encountering congestion and that this is the maximum number that can do so.

Extension of the Deterministic Model.—One of the potential drawbacks of a metering system is the buildup of queues of vehicles on the metered entrance ramps. One restriction that could be placed in the model is an upper limit on the number of vehicles which are not allowed to enter the freeway in the period considered. This can be alternately viewed as establishing the minimum number of vehicles using a given ramp in the time period or a lower limit on the metering rate. This is not necessarily the maximum queue length but could perhaps be related to this quantity without a great deal of difficulty. In any case it might be meaningful to place a limit on the number of vehicles in the hour period which are not allowed to enter the freeway from any input.

This is simply an upper limit on one or more of the slack variables and could be accomplished by adding inequalities of the type $S_j \leq Q_j$, where S_j is the jth slack variable and Q_j is the maximum number of vehicles on the jth ramp which are not allowed to enter the freeway in the time period considered.

Accidents, disabled vehicles, adverse weather, etc., frequently cause reduced capacity operation at one or more sections on the freeway. Hence, it would be desirable to somehow incorporate the effects of these events into the model so that optimal system operation (under the reduced capacity conditions) can be obtained. The discussions will be concerned with a capacity reduction at one section but this can readily be extended to cover the adverse weather situation.

In order to include the reduced capacity situation a capacity constraint will have to be placed on a section between each successive pair of ramps or, alternately viewed, on a section downstream of each location at which the volume can change. These would be similar to the constraints placed on sections A, B, and C in Figure 9. During normal conditions the normal capacity at each section would be used and many of the constraints would be redundant. However, in case an accident reduced the capacity at a given section, the reduced capacity would be used in the constraint for this section. The solution of this problem would yield the optimal inputs at each ramp under the conditions.

When thinking of using the linear programming model for control, one might wonder how the capacity at an accident location could be determined since an accident can have a wide range of effects—from virtually no effect to the closing of all of the freeway lanes (in one direction). If a detection station is located downstream of the accident it is possible to measure the capacity flow rate directly when congestion develops at the accident. The flow rate at the downstream detection station would be the capacity flow rate past the accident. This capacity could then be used in the linear programming model if it were being used for control purposes.

Limitations.—The use of the deterministic linear programming model assumes accurate knowledge of the O-D characteristics of each volume input source in the system. These data might change significantly with time on a given day or from one day to another. The implementing of a metering system would almost certainly change these characteristics so it would be necessary to obtain new O-D data for the system.

This model can be used only for time periods which are long compared to the travel time through the system. For this reason, it might be necessary to consider a dynamic model.

Obtaining O-D Data and Estimating Demand at a Section.—In view of the sensitivity of the model to changes in certain O-D data, it is quite important to have an accurate knowledge of these data. Since it might vary by time of day the data should be collected according to short time periods (such as 15 minutes). It is necessary to determine for each input source for each time period the percent of vehicles which exit at each output. This could be done in any one of several ways. The method discussed here consists of a 100 percent sample of vehicles entering each ramp on each of several days. It is quite a laborious method but it provides a great deal of information that more conventional O-D techniques could not provide. Sampling could be confined to time periods of interest.

The method was actually used by Brenner, et al. (11), but for different purposes. It consists of recording the time of arrival and the license number of each vehicle entering the freeway at each entrance ramp in the system, the time of departure and license number of each vehicle leaving the freeway at each exit ramp and counts of all vehicles entering and leaving the system via the freeway input and output. Matching the license numbers and times would yield the O-D data by time of day. As was done by Brenner, et al., these data could be used to determine travel times as well.

Another valuable by-product of these data would be the ability to estimate the demand on any section in the system. If the free-flow travel time between an entrance ramp and a bottleneck section is t, a vehicle entering the freeway at this entrance ramp at time t_0 represents one unit of demand at the bottleneck at time $t_0 + t$ (providing it does not exit before reaching the bottleneck). This is independent of the effects of intermediate bottlenecks. The sum of these demands over all inputs would yield the actual demand at a given bottleneck.

CONCLUSIONS AND RECOMMENDATIONS

This paper is primarily theoretical and presents many hypotheses which need to be tested. Subject to the validation studies suggested in the section on recommendations which follows, this theoretical study offers the following conclusions applicable to a somewhat idealized urban freeway system of the type which was analyzed.

Conclusions

1. Congestion develops at a bottleneck location when the upstream flow exceeds the bottleneck capacity for a sufficiently long period of time.

2. The development of congestion at a bottleneck causes high-density, low-speed operation upstream of the bottleneck.

3. There is evidence to indicate that the flow rates at many freeway bottlenecks are lower when there is congestion upstream than during some periods of free flow. The reduction in the flow rate at a bottleneck under normal operating conditions may be due primarily to the inability of the congested upstream freeway to supply vehicles to the bottleneck at its capacity flow rate. Under these conditions the start-and-stop flow upstream of the bottleneck may be the factor limiting the flow rate to the bottleneck.

4. Since congestion upstream of a freeway bottleneck can cause the bottleneck flow rate to decrease, the output of the freeway system can be increased (or maintained at its maximum level) by the prevention of congestion at all locations in the system. One goal of a control system, then, is to prevent the development of congestion everywhere in the freeway system.

5. In most cases, for a given demand on a street and/or freeway system, the peak-period objective of maximizing the output of the system is equivalent to minimizing the total travel time in the system.

6. Traffic control on freeway systems holds promise for reducing freeway congestion and reducing travel time in the total street and freeway system. Such controls include both (a) controls on the freeway and (b) control of the inputs to the freeway.

7. Control of certain inputs to the freeway system seems to be the most effective method of preventing congestion on the freeway during normal operating conditions and minimizing the effects of reduced capacity operation. Of the various input controls, ramp metering appears to hold the most promise. By allowing ramp vehicles to enter the freeway at the maximum rate that will not cause congestion, it should be possible to obtain the best use of the freeway system.

8. At most metered entrance ramps, the vehicular storage capacity probably is insufficient to store the maximum queue which develops at the ramps. The storage of queued vehicles is one major problem of metering.

9. Volume measurements should be very useful as predictors of developing congestion, as long as detection takes place upstream of all bottlenecks and entrance ramps. However, measurements of lane occupancy, speed or density are needed in addition as indicators of congestion, such measurements preferably to be made at bottleneck locations.

10. The use of volume measurements in a freeway control system also has other advantages: (a) such measurements provide a check to determine whether the output volume rate of the system is being maximized; and (b) the continuity characteristics of volume make it the only variable which is well suited to theoretical system analyses.

11. Application of a control at one location affects traffic operations at many other locations. The entire system should be studied, not the isolated locations. For this reason, a systems analysis is perhaps the most adequate analytical technique for predicting the effect of a control or control system on the system under consideration.

12. Linear programming provides a valuable tool for describing the operation of a freeway system or subsystem. It is possible that it could be used in reduced capacity situations to determine the proper controlled ramp inputs to provide optimal peak-period operation of a freeway system.

13. Demand on a freeway section can be estimated by sampling the O-D characteristics and volume-time characteristics of free-flowing system inputs upstream of the section.

Recommendations

Probably the most important empirical study that should be undertaken is the study of the behavior of the macroscopic traffic variables at and upstream of various types of bottlenecks in order to determine the effect of control on traffic operation at these locations. The questions of whether or not congestion normally decreases the flow rate, in what situations the volume rate decrease takes place, and how much the volume decreases due to congestion must be answered. In situations in which congestion does not decrease the flow rate, freeway travel time under normal operations can be significantly decreased only by decreasing the inputs. In this case, the output rate of the freeway is little affected by freeway storage, so the prevention of freeway congestion is not necessarily the peak-period objective (although it probably would still be desirable).

The flow rate away from a queue should also be investigated because it will furnish some information on a steady-state congested flow rate. Perhaps there is no single value, but if there is, our knowledge of it will contribute greatly to the evaluation of the possible effects of a control system.

Shock wave development and propagation should also be studied. Queueing forms and dissipates at bottlenecks prior to congestion but finally flow breaks down when excessive queueing takes place. The causes of queueing and the behavior of the shock waves at a bottleneck should be studied, because such study will furnish information on the causes of congestion and for what time period and by how much upstream flow rate can be allowed to exceed the bottleneck capacity. In other words, it will help to determine the probability of congestion which is associated with a particular set of upstream flow conditions.

Since many freeway bottlenecks are at entrance ramp locations, the merging maneuver could be studied along with the shock waves and queues at bottlenecks. The capacity of the merging areas under different conditions must be obtained in order to establish the upper limit of the ramp input rate for a given upstream flow rate. It is frequently assumed that there is a fixed merging capacity and that the proportions of vehicles merging from the ramp and freeway do not affect the capacity value. This assumption of a one-to-one tradeoff between ramp and freeway vehicles should be investigated.

The effect of entrance ramp metering on the merging capacity is also needed. If metered arrivals from the ramp allow a larger upstream freeway flow rate, an additional benefit of metering will have been realized.

A method for the estimation of demand must be evaluated. The method of recording license numbers and arrival times of vehicles at the upstream entrance ramps as well as a free-flowing freeway section should be tested. The duration as well as the severity of the control at a given location depends on the demand function at this location. If it is not known, the selection of the proper control may be difficult. The changes in demand caused by upstream controls must also be determined.

While the license numbers and arrival times are being recorded on the ramps, another important study should be conducted. The linear programming model assumed that the freeway O-D pattern remained constant during the peak period. If license numbers of exiting vehicles are recorded at each exit ramp, the changes in the O-D patterns with time can be obtained.

The control system proposed here would work best under "normal" traffic conditions (i.e., no accidents, disabled vehicles or other "unusual" events), since the full capacity of the freeway could be used. However, the "unusual" situation can also be taken care of since ramp closure is possible as part of a flexible metering system. In this case some vehicles would be diverted around the capacity reductions on the freeway and onto those surface streets which have remaining capacity. The frequency of these events under various volume rates and congestion conditions should be examined. Even more basic and important, the effects of a traffic accident, tire changer, disabled vehicle and other "unusual" events on traffic behavior and especially on the flow rates should be studied. The effect of adverse weather on the capacity flow rate of bottleneck sections also warrants intensive investigation. The complexity of the final control system may depend on the outcome of these studies.

The cost of the final detection system could be substantially reduced if it is possible to use measurements of speed, lane occupancy or density to estimate volume or if it is possible to estimate the traffic variables for all lanes by sampling detection in one lane. These possibilities should be thoroughly investigated to determine the sacrifice of accuracy that accompanies an economic savings.

The philosophy of this paper is to accept the total demand on the freeway and to operate the freeway system in an optimal manner. No vehicles (or at least as few as possible) would be diverted from the freeway. They could be delayed from entering by means of the metering system. This set of conditions permits only the optimal operation of the freeway system but does not assure optimal operation of the total system which includes the streets as well as the freeways. The next logical step is the development of a model which would yield the optimal operation of the entire system. Perhaps the Charnes-Cooper multi-copy model (10) with capacitated entrance ramp links (to produce travel time increases with volume increases) would fill this need.

ACKNOWLEDGMENTS

This paper is based on part of a doctoral dissertation at Northwestern University. Wattleworth was the author and Berry was chairman of the faculty committee in charge of evaluating the dissertation. Other members of this committee, whose contributions certainly deserve mention, are Professors Abraham Charnes, Adolf D. May and Loring G. Mitten.

The work on this material was also part of Wattleworth's duties at the Expressway Surveillance Project of the Illinois Division of Highways. Discussions with the staff of this project, including Project Director Adolf D. May, also proved valuable in the preparation of this material.

REFERENCES

1. Rothrock, C. A., and Keefer, L. E. Measurement of Urban Traffic Congestion. Highway Research Board Bull. 156, pp. 1-13, 1957.
2. Edie, L. C., and Foote, R. S. Traffic Flow in Tunnels. Highway Research Board Proc., Vol. 37, pp. 334-344, 1958.
3. Edie, L. C., and Foote, R. S. Effect of Shock Waves on Tunnel Traffic Flow. Highway Research Board Proc., Vol. 39, pp. 492-505, 1960.
4. Wattleworth, J. A. Bottleneck Throughput Study. Unpublished report presented to the Freeway Operations Comm. of the Highway Research Board, 1963.
5. Highway Capacity Manual. U. S. Dept. of Commerce, Bureau of Public Roads. 1950.
6. May, A. D. Experimentation with Manual and Automatic Ramp Control. Highway Research Record 59, pp. 9-38, 1964.
7. Barker, J. L. Determination of Discontinuities in Traffic Flow as a Factor in Freeway Operational Control. Inst. of Traffic Engineers Proc., pp. 25-40, 1961.
8. Lighthill, M. J., and Whitham, G. B. On Kinematic Waves II. A Theory of Traffic Flow on Long Crowded Roads. Royal Soc. of London Proc., Series A, Vol. 229, pp. 317-345, 1955.
9. Characteristics of Traffic Entering the Westbound Congress Expressway Within the Pilot Detection System Study Section. Unpublished project report of the Expressway Surveillance Project, Chicago, 1962.
10. Charnes, A., and Cooper, W. W. Management Models and Industrial Applications of Linear Programming. 2 Vols., Wiley, 1961.
11. Brenner, R., Telford, E., and Frischer, D. A Quantitative Evaluation of Traffic in a Complex Freeway Network. Highway Research Board Bull. 291, pp. 163-206, 1961.

Operating Costs at Intersections Obtained From The Simulation of Traffic Flow

EARL R. RUITER, Traffic Planning Analyst, Chicago Area Transportation Study, Chicago, Illinois*, and
PAUL W. SHULDINER, Associate Professor of Civil Engineering, Northwestern University, Evanston, Illinois

A computer program has been written which simulates the traffic at an urban intersection, and determines both delays and fuel consumption of vehicles passing through the intersection. By placing typical unit costs on hours of time and gallons of fuel, operating costs are determined for each vehicle and then averaged for all vehicles traveling on each of the two streets. The variable inputs to the program include type of intersection control (two-way stop or semi-traffic-actuated signal), volume levels, turning percentages, critical lag at the stop sign or signal phasing and detector locations for the traffic signal, sampling time, and vehicle fuel consumption characteristics. The program is written for an IBM 704-709 computer and has an approximate real time to computer time ratio of four to one.

To illustrate the usefulness of the program in the economic analysis of intersections, the program was run at various combinations of main street and side street volumes under both traffic signal and stop sign control. The cost contours for each type of intersection control were compared to find areas where stop sign control resulted in the lowest operating costs, where traffic signal control was cheapest, and where the two types of control resulted in equal operating costs. The line of equal operating costs can be considered a warrant line separating traffic signal preferability from stop sign preferability.

•AS AN AID in the selection among alternative transportation improvements, highway and traffic engineers have made extensive use of a form of economic analysis which involves the evaluation of the anticipated effects of each alternative upon road-user costs. Direct vehicle operating costs constitute a major element in such analyses, and much effort has, therefore, gone into determining how these costs vary with speed, gradient, curvature, and pavement and vehicle type (1, 10, 15). The excess cost of stopping over that of traveling at various constant speeds has also been studied (1).

A gap in the knowledge exists, however, in the case of predicting operating costs at intersections. Besides the previous factors, it appears that operating costs vary with volume and type of control. Because these relationships are not known exactly, traffic engineers now resort to noneconomic methods of justifying expenditures at intersections. One such method is the use of warrants based on engineering judgment and on observations of intersection performance. Warrants have been developed for stop sign and traffic signal intersection control (11, 13). Another noneconomic method is the

*Formerly, Graduate Student and Research Assistant, Civil Engineering Department, Northwestern University.

Paper sponsored by Committee on Traffic Control Devices and presented at the 43rd Annual Meeting.

use of sufficiency ratings which gave an indication of the priority and need of various improvements (14).

A method of accurately predicting operating costs at all intersections would make it possible to replace warrants based on judgment and observation of performance with warrants based on minimizing the total costs associated with the intersection. Also, individual projects proposed to improve a given intersection could be compared by means of the normal methods used in engineering economy.

The purpose of this paper is to report on a method of predicting operating costs at intersections using traffic simulation techniques. The use of the method is illustrated by the development of warrants based upon the minimization of operating costs.

Prediction of Operating Costs at Intersections

In the past, intersection studies have concentrated on describing vehicle behavior such as average headways, delays and queue lengths, as well as the variations in these factors with changes in volume level. Time-lapse photography and various types of delay meters have been used to study vehicle operating characteristics in the field. More recently, analytical models have been developed using probability and queueing theory (12, 13) and Monte Carlo methods for the simulation of vehicle behavior (5, 9). Each of these techniques has its own inherent advantages and disadvantages. However, Monte Carlo simulation holds, perhaps, the greatest promise through its ability to deal with complex probabilistic situations for which no direct analytical method of solution is known.

For this reason, the simulation method was chosen as a basis for the development of a method for predicting operating costs at intersections. The model which was developed is a combination of two previously developed computer programs. The simulation program which was used is that developed by Lewis (8, 9). Fuel costs are obtained using the methods developed by Robbins (6, 7). The combined program predicts fuel and time costs, the two largest factors in operating costs. Among the factors which are not considered are oil, tire, maintenance and depreciation costs.

Lewis' program simulates the operation of the intersection of a four-lane and a two-lane street. The choice of traffic control is limited to either stop signs on the minor street or a semi-actuated traffic signal.

Robbins' program calculates the speed profile and fuel consumption of a representative vehicle traveling over a given highway alignment. The speed profile is limited by driver preferences and vehicle characteristics. Fuel consumption is determined by calculating piston speed and brake horsepower required per square inch of piston area for each time interval. A value of fuel per brake horsepower hour can then be read from a fuel map relating this quantity to piston speed and brake horsepower required per square inch of piston area.

Description of the Modified Program

In order to obtain operating costs for vehicles passing through intersections, it was decided to modify the intersection simulation program written by Lewis so that it would calculate the fuel consumption of each vehicle as it moves through the intersection area. The method of calculating fuel consumption is essentially the same as that developed by Robbins (6). Total operating costs are obtained for each vehicle as it is released from the intersection by adding its accumulated fuel costs (fuel consumption multiplied by gasoline cost) to its time cost (total time spent by the vehicle in the system multiplied by the value of time). These total operating costs are then accumulated for all vehicles starting in a given lane and performing a given turning maneuver.

Simplifying Assumptions

A number of the assumptions used to simplify the model are those employed by Lewis in formulating his simulation program:

1. Vehicles travel so as to minimize their delays.

2. Factors such as minimum spacing of vehicles, maximum speeds, acceleration and deceleration rates, and acceptable gaps are constants for all drivers and all vehicles.
3. Pedestrians have no effect on drivers.
4. The opportunity to pass is limited to straight through vehicles following turning vehicles.

Other simplifying assumptions were necessary in order to make the Lewis' and Robbins' programs compatible. These assumptions include the following:

1. All operating costs except fuel and time costs can be ignored.
2. The effects of vertical grades and curve resistances on fuel consumption at intersections can be ignored.
3. Vehicles are capable of performing according to their drivers' preference; they are not limited by the vehicles' capabilities as in Robbins' program.
4. All vehicles using an intersection can be represented by one vehicle type, with one set of vehicle characteristics.

Resulting Program

The major addition to the Lewis program is the provision of a method of calculating fuel consumption for each vehicle during each time interval. By applying the simplifying assumptions to the Robbins' procedure, the following method for computing fuel costs was obtained:

1. Determine the acceleration rate, average speed, and distance traveled during the time interval, based on the maximum desired speed, and limited by spacing, acceleration, stopping, and turning restrictions.
2. Determine which gear the vehicle will be in.
3. Calculate the speed of the engine in revolutions per minute.
4. Calculate air, rolling, and acceleration resistances.
5. Calculate the brake horsepower required per square inch of piston area.
6. If the vehicle is idling at a stop or coasting, use a linear equation relating fuel consumption to engine speed to calculate the fuel consumption. If the vehicle is not idling or coasting go to step 7.
7. Calculate the rate (feet per minute) of piston travel.
8. Use the results of steps 5 and 7 to find from the fuel map the amount of fuel per brake horsepower hour which will be consumed.
9. Multiply the results of step 8 by the brake horsepower and the time increment to determine the amount of fuel which will be consumed during the current time interval.

Lewis' input routines were modified so that the vehicle data needed to calculate fuel consumption could be read in. Output routines were modified so that they would calculate and print out operating cost data in addition to the delay data given by the original program.

The modified program has a real time to computer time ratio of four to one, using an IBM 709 computer.

USE OF THE PROGRAM

Selection of the Input Data

Most of the input data were chosen to correspond to that used by either Lewis or Robbins in their individual programs. The fuel map and vehicle type (1960 Plymouth station wagon) were those used by Robbins. The intersection parameters were those used by Lewis. A summary of these data is given in Table 1.

Computer test runs were made to insure that the action of the vehicles had not been changed from the experience in the unmodified program. Fuel consumption rates were determined and checked for reasonableness. Also, the variability or ratio of standard deviation to means of the operating costs for individual vehicles was

TABLE 1

PARAMETERS AND INPUTS USED TO OBTAIN VOLUME WARRANTS

Parameters:	
Maximum desired speed	44 fps
Maximum acceleration rate:	
Normal conditions	3 fps^2
Starting from stop	6 fps^2
Maximum deceleration rate:	
Normal conditions	6 fps^2
Stopping at amber light	12 fps^2
Arrival distribution	Modified binomial
Minimum vehicle spacing	22 ft
Inputs:	
Fuel map	Typical for gasoline engines
Vehicle	1960 Plymouth station wagon
Gasoline price	\$0.33/gal
Time cost	\$1.50/hr
Transient time	300 sec
Sample time	Variable
Distance of detectors from stop lines	21 ft
Critical lags	5.8 sec
Lane volumes	Variable
Traffic signal controller intervals:	
Main street	
Minimum green	30 sec
Amber	3 sec
Side street	
Initial green	2 sec
Extension green	4 sec
Maximum green	30 sec
Amber	3 sec
Directional distributions	60%-40%
Lane distribution, 4-lane streets:	
Outside lane	60%
Inside lane	40%
Turns, % of total volume:	
Main street, both turns	7 each
Side street, both turns	14 each

checked so that production run time could be chosen which would result in a uniform level of accuracy from one operating cost figure to another.

Selection of Computer Running Times

As a first step in running the modified program, it was necessary to determine the duration of run required at each volume level to attain a preselected level of significance. An equation relating duration of run to volume was derived (see Appendix). The sample time for each run was determined by use of this equation for both main and side street volumes. The largest of the two durations prescribed was then selected. The resulting savings in machine time amounted to approximately 30 percent when compared with the commonly used constant sample time of one hour.

Warrants for Intersection Control

One of the underlying purposes of the modified program is the developing and testing of intersection control warrants based on minimum average vehicle operating costs. Two types of intersection control (stop sign and semi-actuated signal) and a range of main street volumes (400 to 1,400 veh/hr) were tested at side street volume levels chosen so as to lie on both sides of the minimum delay warrant line developed by Lewis. Additional side street volume levels were tested in those instances where the initial pair of volumes did not define the preference boundary.

The results of these runs are given in Tables 2 and 3. Table 2 summarizes the values derived assuming traffic signal control; the results for the stop sign condition are given in Table 3. Figure 1 presents these data in terms of equal cost contours for both types of control for all levels of side street and main street volumes which were considered. The intersection of equivalent contour lines indicates combinations of side street and main street volumes for which vehicle operating costs are equal for both traffic signal and stop sign control. A warrant line may be drawn through these points representing the minimum vehicle operating cost boundary between these two types of traffic control. Such a curve is shown in Figure 2 (solid line) along with the minimum delay warrant line developed by Lewis.

For the most part, the equal cost contours indicate that average operating costs increase with both types of control as either side street or main street volumes increase. At high levels of side street volume, however, the apparently anomalous situation exists of average vehicle costs decreasing with increasing main street volumes. If it is recalled that we are dealing with average vehicle operating costs, the explanation becomes fairly obvious. At constant side street volumes (and constant side street costs), the average side and main street costs decrease as a result of the increased proportion of lower main street costs brought about by increasing the number of main street vehicles. These contours would begin to slope downward to the right as congestion on the main street increases the main street vehicle operating costs.

Since there is an added cost associated with installing a signal light at an intersection instead of a stop sign, the warrant line in Figure 2 is not strictly applicable. However, when the cost of a signal light is capitalized over its useful life and the cost per vehicle is determined, it will be very low.

TABLE 2

WARRANT PRODUCTION RUNS—TRAFFIC SIGNAL CONTROL

Nominal Volumes (veh/hr)		Actual Volumes (veh/hr)		Avg. Operating Costs ($/veh)		
MS[a]	SS[b]	MS[a]	SS[b]	MS[a]	SS[b]	Both[c]
400	400	410	403	0.0139	0.0224	0.0181
	500	418	482	0.0135	0.0234	0.0188
	600	418	518	0.0140	0.0241	0.0196
	650	403	619	0.0158	0.0220	0.0195
600	150	624	160	0.0117	0.0234	0.0140
	200	622	206	0.0125	0.0244	0.0155
	250	624	216	0.0129	0.0212	0.0150
	350	657	342	0.0143	0.0219	0.0169
	400	630	414	0.0149	0.0229	0.0181
800	150	832	160	0.0126	0.0234	0.0143
	250	822	216	0.0132	0.0212	0.0149
1,000	75	1,068	80	0.0121	0.0222	0.0128
	175	1,062	158	0.0124	0.0214	0.0136
1,200	50	1,241	64	0.0122	0.0206	0.0126
	100	1,288	105	0.0129	0.0233	0.0137
	150	1,272	160	0.0134	0.0234	0.0145
1,400	25	1,428	21	0.0110	0.0194	0.0111
	75	1,494	80	0.0129	0.0221	0.0134

[a]MS = main street.
[b]SS = side street.
[c]Both = both main and side streets.

TABLE 3

WARRANT PRODUCTION RUNS—STOP SIGN CONTROL

Nominal Volumes (veh/hr)		Actual Volumes (veh/hr)		Avg. Operating Costs ($/veh)		
MS[a]	SS[b]	MS[a]	SS[b]	MS[a]	SS[b]	Both[c]
400	400	389	363	0.0100	0.0214	0.0155
	500	385	508	0.0099	0.0247	0.0183
	600	385	580	0.0099	0.0254	0.0192
	650	385	608	0.0099	0.0257	0.0196
600	150	642	165	0.0101	0.0256	0.0132
	200	619	207	0.0101	0.0318	0.0155
	250	623	263	0.0101	0.0309	0.0163
	350	672	309	0.0101	0.0331	0.0174
	400	645	353	0.0101	0.0369	0.0196
800	150	877	165	0.0098	0.0293	0.0129
	250	834	254	0.0099	0.0430	0.0176
1,000	75	1,039	71	0.0100	0.0349	0.0116
	175	1,073	179	0.0100	0.0581	0.0169
1,200	50	1,235	48	0.0100	0.0428	0.0112
	100	1,221	99	0.0101	0.0520	0.0132
	150	1,246	170	0.0102	0.1791	0.0305
1,400	25	1,421	21	0.0102	0.0542	0.0108
	75	1,432	71	0.0102	0.1073	0.0148

[a]MS = main street.
[b]SS = side street.
[c]Both = both main and side streets.

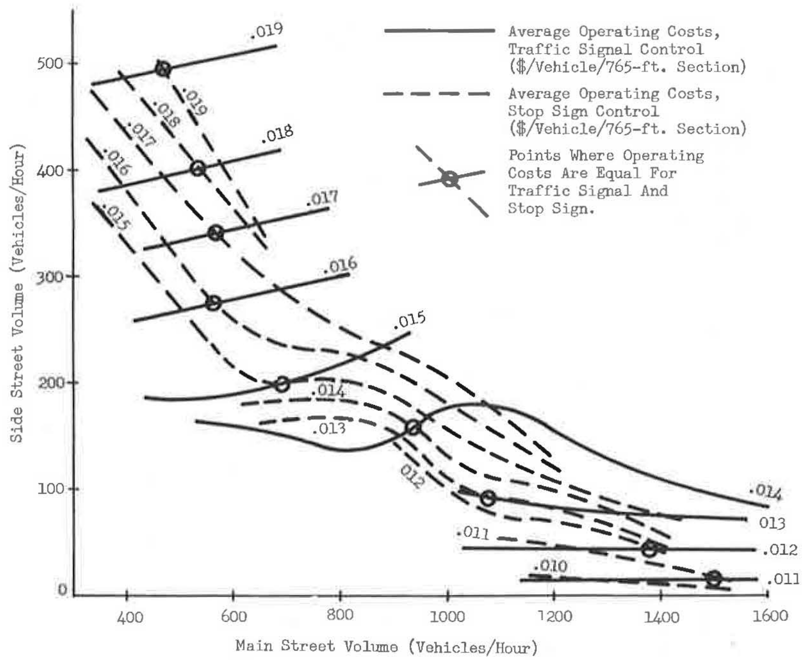


Figure 1. Contour map of vehicle operating costs at an intersection.

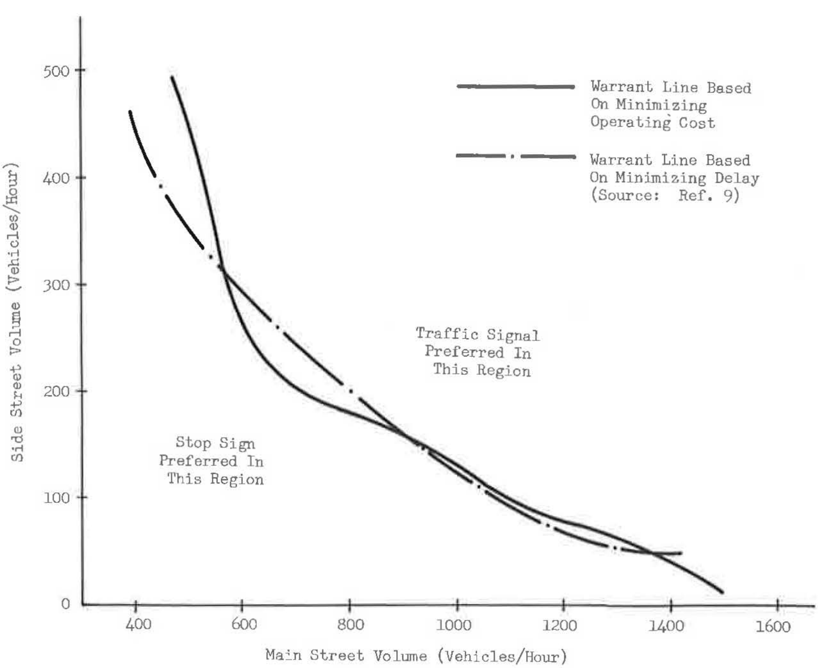


Figure 2. Comparison of warrant lines—minimum cost and minimum delay.

TABLE 4

VOLUME WARRANTS FOR PRETIMED SIGNALS[a]

Number of Lanes		Volumes	
MS	SS	MS	SS
(a) Warrant I—Minimum Volume			
4 or more	2	600	250
2	4 or more	500	333
(b) Warrant II—Interruption of Continuous Traffic			
4 or more	2	900	125
2	4 or more	750	167

[a]Derived from Manual on Uniform Traffic Control Devices, pp. 185-186; pretimed signals are warranted whenever the intersection volumes exceed those given for 8 hr per day.

Comparison with Other Warrants

Figure 2 shows graphically the difference between the minimum delay warrant developed by Lewis and the minimum operating cost warrant. The curves are nearly the same for main street volumes higher than 900 veh/hr. The entire operating cost curve, however, is more sharply "kinked" and therefore lies below Lewis' curve in the 550 to 900 main street volume range and above it for lower volumes.

The warrants based on operating costs can also be compared with those given in the Manual on Uniform Traffic Control Devices (11) for pretimed traffic signals (Table 4). The side street volume figures are given for both directions of approach, obtained from the Manual warrants by assuming a 60 to 40 percent directional distribution. If these warrants are interpreted strictly, the warrant line separating stop sign preference from signal light preference appears as a series of right-angled steps (Fig. 3). The lower corners of these steps are the points specified in Table 4, with changes of designation so that the main street is always the one with four traveled lanes and the side street the one with two traveled lanes. If the warrants are interpreted more loosely, the warrant lane can be obtained by drawing a smooth curve through the points given in Table 4. Such a curve is also shown in Figure 3.

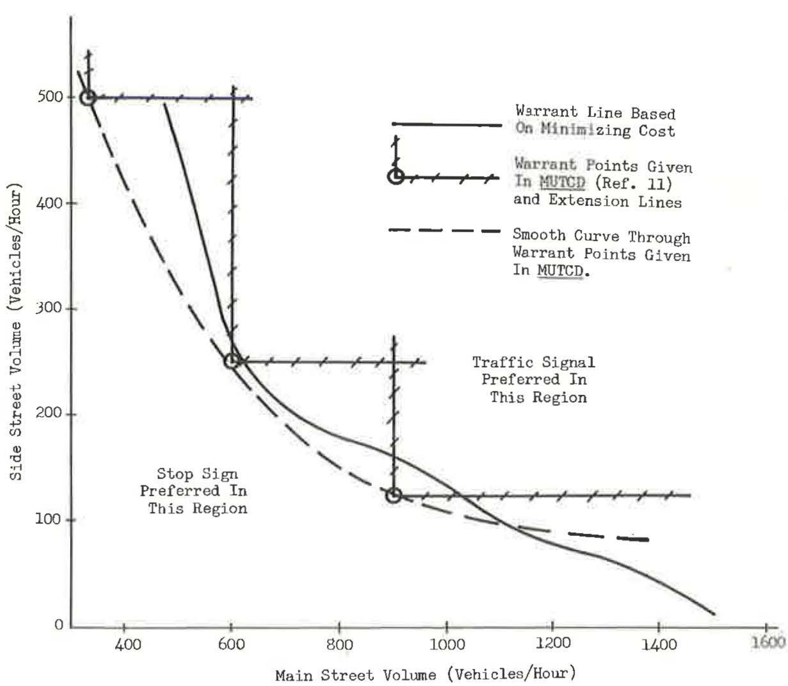


Figure 3. Comparison of warrant lines—minimum cost and Manual on Uniform Traffic Control Devices.

When discussing traffic-actuated signals the Manual states that they are warranted "at intersections where the volume of vehicular traffic is not great enough to warrant pretimed signals, . . . if other conditions indicate the need for traffic control signals and justify the cost of installation." (11, p. 200) This seems to indicate that if the Manual gave specific volume warrants for actuated signals, they would be lower than those given for pretimed signals.

Figure 3 indicates that the warrant for semi-actuated signals based on operating costs and the volume warrants for pretimed signals given in the Manual would result in the same choice of intersection control (stop vs signal) in most cases. However, if the Manual curves were shifted downward to any great extent to serve as actuated signal warrants, the result would be that at many volume combinations at which signals would be chosen they would result in higher operating cost than would stop signs.

A number of factors must be kept in mind before applying the warrant for intersection control based on operating costs. One of these factors is that the warrant is based on only one criterion of many possible criteria. Operating costs are minimized, but there is no recognition of such factors as pedestrian volumes, accident experience, and the need of progressive movement. Of course, the desire to minimize delays is taken into account by assigning a cost to a vehicle's time.

Another factor which must be recognized is that the warrant is based on a host of assumptions as to drivers' characteristics, traffic characteristics, and fuel consumption characteristics of a representative vehicle. Changes in any of these parameters will affect the warrant line obtained.

The warrant line based on minimizing total operating costs is presented, therefore, not as the answer to the problem of what type of control to install at a given intersection, but as an example of how the operating costs at intersections program can be used. Once satisfactory values are found for all the parameters involved, similar warrants could be developed which could be combined with warrants based on other criteria in a handbook such as the Manual on Uniform Traffic Control Devices.

CONCLUSIONS

The addition of the calculation of operating costs to an intersection simulation model has provided a model which enables the engineer to analyze more accurately the operating costs associated with intersections. These data are especially useful when determining the type of intersection control which should be used at intersections on major highways, whether this "control" is a stop sign, traffic signal, or the elimination of the intersection by interchange.

The volume warrants based on minimizing operating costs provide an economic method of determining whether traffic signals or stop signs should be used at intersections. This economic method can easily be improved by adding other costs (accident, oil consumption, etc.) as they become available. The warrants developed here are in general agreement with the existing warrants given in the Manual on Uniform Traffic Control Devices.

ACKNOWLEDGMENTS

This work was performed as part of project IHR-55, Spacing of Highway Grade Separations, conducted at Northwestern University Civil Engineering Department and financed jointly by the Illinois Division of Highways and Northwestern University. A large share of the computer time used was donated by the Northwestern University Computing Center.

The authors wish to acknowledge also the assistance of Professor Russell M. Lewis who consented to the use and modification of his intersection simulation program. The Civil Engineering Systems Laboratory at the Massachusetts Institute of Technology was helpful in providing details on the vehicle operating cost method developed by David H. Robbins.

REFERENCES

1. AASHO, Committee on Planning and Design Policies. Road User Benefit Analyses for Highway Improvement. Washington, D.C., 1960.
2. Berry, D.S. Field Measurement of Delay at Signalized Intersections. Highway Research Board Proc., Vol. 35, pp. 505-522, 1956.
3. Dixon, W.J., and Massey, F.J., Jr. Introduction to Statistical Analysis. New York, McGraw-Hill, 1957.
4. Greenshields, B.D., Schapiro, D., and Erickson, E.L. Traffic Performance at Urban Street Intersections. Tech. Rept. No. 1, Bureau of Highway Traffic, Yale Univ., 1947.
5. Kell, J.H. Analyzing Vehicular Delay at Intersections Through Simulation. Highway Research Board Bull. 356, pp. 28-39, 1962.
6. Lang, A.S., and Robbins, D.H. A New Technique for Predicting Vehicle Operating Cost. Highway Research Board Bull. 308, pp. 19-35, 1962.
7. Lang, A.S., Roberts, P.O., and Robbins, D.H. An Evaluation of Techniques for Highway User Cost Computation. Highway Research Board Bull. 320, pp. 1-11, 1962.
8. Lewis, R.M. The Simulation of Traffic Flow to Obtain Volume Warrants for Intersection Control. Ph.D. Thesis, Purdue Univ., Sept. 1962.
9. Lewis, R.M., and Michael, H.L. Simulation of Traffic Flow to Obtain Volume Warrants For Intersection Control. Highway Research Record No. 15, pp. 1-43, 1963.
10. Lieder, N. Passenger Car Fuel-Consumption Rates. Public Roads, Vol. 32, No. 5, pp. 113-120, Dec. 1962.
11. National Joint Committee on Uniform Traffic Control Devices. Manual on Uniform Traffic Control Devices for Streets and Highways. U.S. Dept. of Commerce, Bureau of Public Roads, Washington, D.C., June 1961.
12. Newell, G.F. Queues for a Fixed Cycle Traffic Light. In Annals of Mathematics and Statistics, Vol. 31, pp. 589-597, 1960.
13. Raff, M.S., and Hart, J.W. A Volume Warrant for Urban Stop Signs. Eno Found. for Highway Traffic Control, Saugatuck, Conn., 1950.
14. Schenler, W.W., and Michael, H.L. A Sufficiency Rating Method for Urban Intersections. Proceedings of the 48th Annual Road School, Purdue Univ., Eng. Extension Series No. 1, pp. 239-265, March 1962.
15. Woods, K.B. Highway Engineering Handbook. New York, McGraw-Hill, 1960.

Appendix

Selection of Run Times

The method used to determine the sample time necessary to achieve a desired degree of accuracy follows:

Since we are concerned with comparing sample means with population means when the standard deviation of the population (σ) is unknown, the two-tailed t test is applicable (3).

$$t_{(\frac{1}{2}\alpha)\ (N-1)} = \frac{\bar{x} - M}{s/\sqrt{N}} \tag{1}$$

in which

- t = statistic used to test for equality of population mean and sample mean when σ is unknown;
- α = the probability of rejecting a true hypothesis;
- $N-1$ = degrees of freedom of t distribution;
- x = sample mean;
- M = population mean;
- s = sample standard deviation; and
- N = number of observations in the sample.

We wish to keep the difference in means ($\bar{x}$ - M) less than or equal to a given fraction (p) of $\bar{x}$. Therefore, let

$$d = (\bar{x} - M) = p\,\bar{x} \tag{2}$$

Substituting Eq. 2 into Eq. 1 and solving for N,

$$N = \left(\frac{t_{(\frac{1}{2}\alpha)\ (N-1)}\ s}{d}\right)^2 \tag{3}$$

N is related to the volume level Q (veh/hr) and the elapsed time T (seconds) by

$$N = QT/3,600 \tag{4}$$

The variability of the sample data (V) may be defined in the following manner:

$$V = \frac{s}{\bar{x}} = \frac{ps}{d} \tag{5}$$

Substituting Eqs. 4 and 5 into Eq. 3 and solving for T, we obtain the final relationship for the sample time required to give a desired level of accuracy:

$$T = \frac{3,600}{Q}\left(\frac{t_{(\frac{1}{2}\alpha)\ (N-1)}\ V}{p}\right)^2 \tag{6}$$

TABLE 5

DETERMINATION OF SAMPLE TIME REQUIRED

Control	Street[a]	Volume Level (veh/hr)	Avg. Total Cost ($/veh)	Std. Dev. of Total Cost ($/veh)	Variability	Eq. for Sample Time Required[b]
Semi-actuated traffic signal	MS	252	0.0151	0.00687	0.455	T = 192,000/Q
	SS	365	0.0203	0.00762	0.375	T = 130,000/Q
Stop sign	MS	252	0.00991	0.00639	0.648	T = 388,000/Q
	SS	367	0.01961	0.00976	0.497	T = 229,000/Q[c]

[a]MS = main street; SS = side street.
[b]T = sample time (sec); Q = traffic volume (veh/hr).
[c]Later modified to T = 412,000/Q, as described in the text.

In order to use Eq. 6 to determine the sample time required at a given volume level, values must be specified for α, p, and V. An α value of 0.05 was chosen so that the results would be significant at the 95 percent level. The corresponding value of t depends on N - 1, the degrees of freedom, and N is unknown. However, t varies only slightly from a value of 2.0 for all values of N between 20 and infinity when $\alpha = 0.05$. Since it seemed likely that more than 20 vehicles would have to be sampled, t was assumed to have a constant value of 2.0.

A value of p, the allowable fraction of deviation in $\bar{x}$, of 0.125 was chosen since this is approximately the accuracy of the operating cost calculation method (6).

An analysis of the test runs indicated that the variability of the total operating costs depends both on the type of signal control and on which street is being considered. The results of this analysis are given in Table 5. Also given are the resulting equations for T obtained by substituting (into Eq. 6) the previous values given in Table 5.

If the equations for T given in Table 5 are accepted, the implicit assumption is made that the variability found at the volume levels used in the test runs would remain constant, regardless of volume level. This assumption was checked by making a second set of test runs, with volumes of 1,400 and 25 on the main street and side street, respectively. The variabilities for these volume levels were all lower than those given in Table 5, except for the stop sign side street case, where the new variability was 0.667, higher than the 0.497 given in Table 5. The equation for the sample time required for this case was therefore revised to T = 412,000/Q. Although the equation for T in each case could be further modified by making V a function of the volume level Q, this was not done because only limited information was available on the variation of V with Q. Since the two sets of test runs indicated that V tended to be a maximum at the intermediate volume levels, it was decided to use the maximum V's found in these tests and assume them to be constant for all values of the volume level. The net effect is to provide a factor of safety for high and low volumes to overcome the ignorance of the true value of V at these volumes.

Discussion

RUSSELL M. LEWIS, Associate Professor of Civil Engineering, Rensselaer Polytechnic Institute—The authors have cleverly combined the work of D. H. Robbins on predicting operating costs of vehicles and the efforts of this writer in the development of a simulation model of a traffic intersection. They developed curves for direct operating costs based on the costs of fuel consumption and time. As an example of the use of these data, minimum volume warrants were presented for an actuated traffic signal.

A word of caution should be given in regard to the direct use of the operating cost data as given in Figure 1. The validation of simulation models, such as the one used, is a most difficult if not impossible task. To minimize the effects of inaccuracies in the formulation of the model, however, the procedure of model comparison may be used. Insofar as possible, identical models were used to represent the studied intersection as operated under the two types of traffic control—the two-way stop sign and the semitraffic-actuated signal. Any distortions present in the models are thus reflected in a similar manner in the results obtained from each model. The differences in operating cost, therefore, are more reliable than the absolute values of operating cost as obtained for either type of control. The use of model comparison also permits the elimination as direct considerations of such cost producing variables as pedestrian movements, parking interference, and local intersection characteristics.

One of the several advantages of the simulation method is that all variables may be precisely controlled. Traffic is generated by a Monte Carlo process using a probability distribution function and a pseudorandom-number series that can be reset at the beginning of each run. Since only the central tendency is specified, the traffic volume that actually occurs during a run will vary somewhat with different lengths of the run. By using the same length runs for the two different control types, identical traffic volumes occur. (Actually slight variations in traffic volume may occur due to differences in the pattern of releasing vehicles from the system at the beginning and end of runs under the two types of traffic control; such variations are very small for runs that simulate one hour of real traffic.) Furthermore, not only are the traffic volumes the same, but the exact pattern of vehicle arrivals is duplicated. The simulation model employed in this study contained a separate traffic generation and random number routine for each street, enabling the volume level to be varied on one street while retaining the identical traffic on the other street.

The variability of the results obtained from the simulation model is a function of both traffic volume and control type. As volume levels increase on either street, the variability decreases; also the variability is less for signal control than for stop sign control. The use of a constant run time for each set of parameters, therefore, may appear wasteful of machine time. The authors developed a procedure which related the duration of a run to the two street volumes and the type of control. Unfortunately, the employment of variable run times mitigates a most important advantage of the simulation method.

The use of a constant run time would have assured that comparisons in operating cost could be performed independently of any differences in the pattern of traffic that occurred during the periods sampled. In addition, constant run times yield traffic volumes that may be held constant on one street and varied in reproducible increments on the other street. This control over traffic volumes greatly assists in the analysis of the simulation data.

An analysis of the direct operating costs (which include time costs) and the published delay data (9) was performed by the writer. The cost of time represented a nearly constant amount of 70 percent of the cost of operation. Furthermore, the remaining operating costs (that due to fuel consumption alone) exhibited a wide amount of scatter. Therefore, it is indicated that not only is travel time the foremost factor, but also that it is more difficult to draw conclusions from operating cost when time is excluded. The value of time used by the authors was $1.50/veh hr. If a persons per vehicle ratio of 1.8 was assumed, this figure corresponds to $0.83 per person hour. Although it is most difficult to establish a monetary value of time, it is obvious that any increase in the value of time would further decrease the significance of fuel consumption as a factor of operating cost.

It is felt that the apparent differences between the two warrant lines shown in Figure 2 have been over emphasized by the authors. If a smooth curve were used for the warrant based on minimizing operating cost, it would be almost identical to the curve based on minimizing delay originally developed by the writer. The discrepancies as shown in Figure 2 may be largely due to the sampling procedures used, rather than to any basic divergence in warrant principles.

The operating cost information presented by the authors is of great interest and should prove useful in economic studies. For the purpose of developing warrants for intersection control they aptly point out that many other factors (such as accident potential, pedestrian movement, and control at adjacent intersections) must be considered. Delay is recommended as generally preferable to operating cost as the basis for intersection control volume warrants for the following reasons:

1. Delay represents the major portion of operating cost, and the inclusion of other direct operating costs does not materially affect the conclusions that would be drawn from delay alone.
2. Delay is the more readily measured quantity.
3. Delay is the most identifiable factor by the motorist and is dominant in his determination of acceptable intersection control techniques.

EARL R. RUITER and PAUL W. SHULDINER, Closure.—The authors wish to express their appreciation to Professor Lewis for his continued interest in the work using his simulation model. The points brought out in the discussion are conducive to a better understanding of the paper and of the problems involved in simulation in general and in the simulation of operating costs in particular.

Professor Lewis advocates the use of constant run times so that the problem of different patterns of traffic at constant nominal volumes does not arise. However, this problem does arise in reality. The authors feel that the statistical analysis provides a satisfactory method of dealing with the problem, whereas the use of constant run times ignores the problem. If the problem is ignored, the model is removed one more step than is necessary from the reality of random traffic.

Spillback from an Exit Ramp of an Expressway

D. C. GAZIS, IBM Watson Research Center,
Yorktown Heights, New York

A discussion is given of the problem of control of an oversaturated system comprising an expressway, a highway, and an exit ramp leading from the expressway to the highway. A traffic light is assumed to control the intersection of the exit ramp and the exit highway. When this intersection becomes oversaturated, the queue along the ramp may spill back into the expressway causing a reduction of its throughput. Any improvement in the service rate of the exit ramp can be effected at the expense of causing some additional delay to the traffic on the exit highway. The operation of the traffic light serving the intersection of ramp and highway is determined, which minimizes the delay of the vehicles served by the entire system.

•ONE very common feature of congestion is the progressive deterioration of various sections of a roadway system due to the "spillback" from one section to its neighbors. Spillback is the result of queueing at certain points coupled with the troublesome fact that automobiles have nonzero length, and sometimes appreciably so. Given this reality, spillback could only be avoided by providing ample parking space for the queueing vehicles. In practice, such parking space is limited or even nonexistent. The question arises whether or not judicious management of the inevitable queues might decrease the aggregate delay to the users of the entire system.

In two previous papers (1, 2), examples were given of oversaturated systems in which the aggregate delay could be reduced by an appropriate allocation of the green time of the intersection signals throughout the period of oversaturation. The previous theory (1, 2) is used here for the treatment of the problem of optimization of an oversaturated system involving an expressway, 1, an exit ramp, 2, and the exit highway, 3 (Fig. 1). The intersection of 2 and 3 is controlled by a traffic light, 4, as in the case of an observed real situation. It is assumed that this intersection is oversaturated during a rush period. A queue may then build along the exit ramp, 2. When the length of this queue exceeds the storage capacity of the ramp, it spills back into the expressway. The spillback ties up at least one lane of this expressway. In practice, it ties up probably more than one lane, because drivers desiring to use the exit ramp may drive for a while along the lane next to the right lane and then slow down and try to find an opening into the queue. At the same time, some through traffic is invariably trapped in the right lane and fights its way out, very likely reducing the efficiency of the neighboring lane in this process. In any event, a substantial reduction of the throughput of the expressway is caused which frequently results in queueing along this expressway.

In what follows, this spillback problem is treated as one of optimization of an oversaturated system involving three traffic streams along 1, 2, and 3. The control parameter is the split of the green of the traffic light, 4, the operation of which is to be optimized during the rush period.

Paper sponsored by Committee on Vehicle Characteristics and presented at the 44th Annual Meeting.

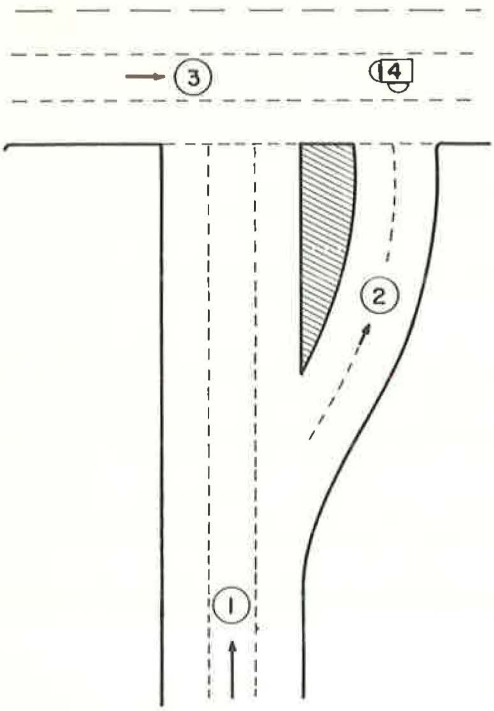


Figure 1. Configuration of the system comprising an expressway, 1, an exit ramp, 2, and an exit highway, 3; the intersection of ramp and highway is controlled by a traffic light, 4.

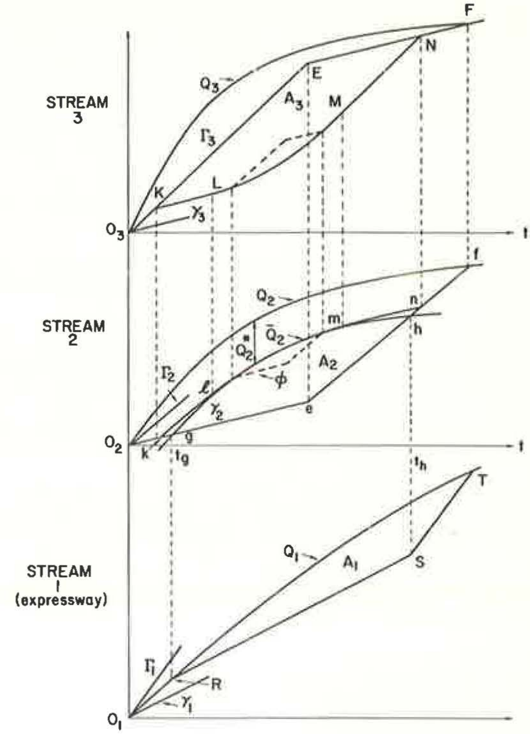


Figure 2. Optimum control of the system when spillback can be avoided altogether.

SOLUTION OF THE OPTIMIZATION PROBLEM

The cumulative demand curves Q_1, Q_2, and Q_3, of the streams 1, 2, and 3 are shown in Figure 2. The maximum and minimum service rates for streams 2 and 3 are determined from the operation of the traffic light 4. Thus, Γ_2, γ_3 correspond to allocation of maximum green to stream 2, whereas γ_2, Γ_3 correspond to maximum green for stream 3. The light cycle is, for the moment, assumed constant. Also shown are two service rates for stream 1, Γ_1 and γ_1. The former is obtainable when the expressway is unobstructed, and the latter is the reduced expressway throughput in case of spillback. It is assumed that Q_1 can be adequately served by the normal service rate Γ_1.

More often than not the saturation flows s_2 and s_3 are such that

$$s_2 < s_3 \tag{1}$$

According to the theory (1), the optimum operation of light 4 alone would be a two-stage operation involving the service curves O_3 EF and O_2 ef, with the highway stream 3 receiving preferential treatment. However, in the present case one must take into account that the intersection 4 is not isolated, and a large enough size of the queue of stream 2 will cause additional delays on the expressway. Let us draw the curve

$$\overline{Q_2} = Q_2 - Q_2^* \tag{2}$$

where Q_2^* is the maximum acceptable queue which does not cause spillback. Let the curve $\overline{Q}_2$ intersect the service curve O_2 ef at points g and h. This means that if one accepts the service curve O_2 ef for stream 2, he will cause spillback during the time

$t_g \leq t \leq t_h$. The result will be a delay for stream 1 proportional to the area between Q_1 and the service curve RST, which will be denoted by A_1.

Assuming now that we operate the light 4 so that the queue along 2 remains smaller than or equal to Q_2^* at all times, any such service curve of stream 2 must be between the curves Q_2 and $O_2k\ell mnf$. The portions $k\ell$ and mn of the latter curve are tangent to the curve $\bar{Q}_2$ and correspond to service rates Γ_2 and γ_2, respectively. The curve $O_2k\ell mnf$ corresponds to the minimum possible service rate of the stream 2 which prevents spillback, with full utilization of the green light in both directions 2 and 3. The complementary service curve of stream 3 is O_3KLMNF. Choosing these two service curves rather than the curves O_3EF and O_2ef involves an increase of the aggregate delay at intersection 4 equal to the difference between the areas A_2 and A_3 which are contained between the pairs of curves (KEN, KLMN) and (ken, $k\ell mn$), respectively. It may be seen that any trade-off of delay between streams 2 and 3 involves quantities proportional to the saturation flows s_2 and s_3. This is so because the trade-off is accomplished by taking green time from stream 3 and giving it to stream 2. The utilization rate of this green time is then reduced from s_3 to s_2 cars per second of green. Accordingly, the ratio A_2/A_3 is given by

$$A_2/A_3 = s_2/s_3 \tag{3}$$

The total change in the aggregate delay of all three streams is given by

$$\delta = A_1 + A_2 - A_3 \tag{4}$$

or, in view of Eq. 3,

$$\delta = A_1 + A_2\left(1 - \frac{s_3}{s_2}\right) \tag{5}$$

A net reduction of delay results if δ is positive. In this case it pays to adopt the strategy of keeping the queue along the ramp below the critical value Q_2^*. If δ is negative, then spillback is not as damaging as it appears, at least in terms of total delay, which is minimized by an optimum operation of the traffic light 4, assumed isolated (1). However, it may still be desirable to prevent congestion on the expressway for safety reasons which may override delay considerations. If this is the case, one may accept a small negative delay trade-off, δ.

Assuming that the delay criterion is the dominant one, we find that a critical constant rate, q_1, of demand along the expressway exists, which is related to A_2 according to a relationship obtained by setting δ in Eq. 5 equal to zero. Thus,

$$\frac{(\Gamma_1 - \gamma_1)(q_1 - \gamma_1)}{\Gamma_1 - q_1}\frac{\tau^2}{2} = A_2\left(\frac{s_3}{s_2} - 1\right) \tag{6}$$

where the left-hand side of Eq. 6 is equal to A_1, and

$$\tau = t_h - t_g \tag{7}$$

Solving Eq. 6 for q_1,

$$q_1 = \frac{2A_2(s_3/s_2 - 1)\Gamma_1 + \gamma_1(\Gamma_1 - \gamma_1)\tau^2}{2A_2(s_3/s_2 - 1) + (\Gamma_1 - \gamma_1)\tau^2} \tag{8}$$

If the demand rate is smaller than q_1, then spillback is the lesser of two evils, since it corresponds to minimum total delay.

By similar arguments we may investigate the possibility of allowing spillback during a portion of the interval $(t_h - t_g)$. If the rate of demand along the expressway falls sufficiently below Γ_1, it may be profitable to adopt a strategy such as that corresponding to the dashed line ϕ in the middle diagram of Figure 2, and the complementary service

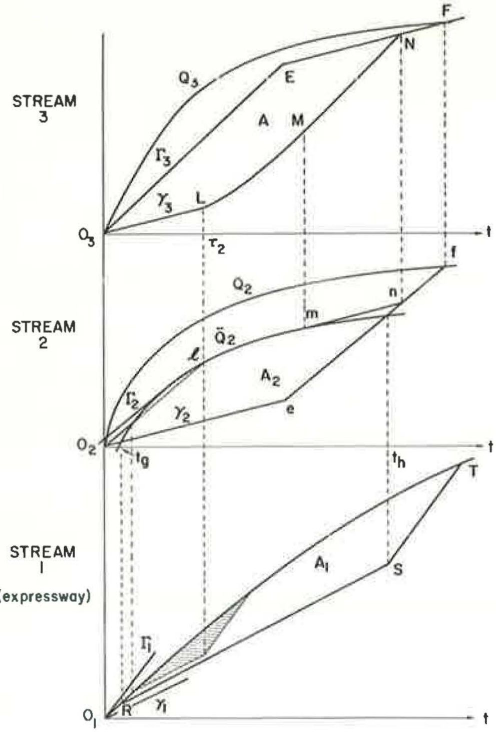


Figure 3. Optimum control of the system when spillback is unavoidable for at least a short period between the onset of oversaturation and the time τ_2.

curves for streams 1 and 3 (the last one not shown in Fig. 2). This policy will introduce some additional delay to the stream 2, and because of spillback it will also delay some vehicles in the stream 1. It will reduce somewhat the average delay to stream 3. The net change can be computed by an expression similar to Eq. 5, if the exact shapes of Q_2 and Q_1 are known. Finally, if the demand rate along the expressway falls below γ_1, it is always profitable to allow spillback.

So far, we have discussed only the case when it is possible to prevent spillback during the entire rush period, if so desired. This is not the case if the tangent to the curve $\bar{Q}_2$ with slope equal to Γ_2 intersects the abscissa axis to the left of O_2, as shown in Figure 3. In this case spillback is inevitable, due to a very fast rise in demand along the exit ramp. The best one can do is maintain maximum service for the stream 2 until the spillback is eliminated at time τ_2. This alternative is to be compared with that corresponding to the service curves O_3EF, O_2ef, and O_1RST. The net change in total delay, δ, is again given by Eq. 5, where A_1 and A_2 now denote the total delay to streams 1 and 2 minus the inevitable one shown by the shaded areas of Figure 3. If δ is positive, then spillback should be prevented after τ_2.

The preceding discussion has certain similarities with the examples of References 1 and 2 and certain differences. As in those examples, the solution given is a deterministic one depending on the demand during the entire rush hour rather than the instantaneous sizes of the queues. Also, the need for anticipating the critical behavior of queues, on the basis of available data regarding recurrent demands, is shown in Figures 2 and 3. Thus, if spillback is to be avoided, one must sometimes act before the queue along the exit ramp attains the critical size Q_2^*. Thus, the optimum strategy calls for maximum service of this queue starting at t_k (Fig. 2) and at 0, i. e., the onset of oversaturation (Fig. 3).

One special feature of the present problem is that the size of the queue along the exit ramp affects the service rate of the expressway 1. This was not the case in the problems of References 1 and 2 where it was pointed out that the asymptotic behavior of the demand curves, near the end of the rush period, might be sufficient for determining the optimum operation of the traffic lights. In the present problem, however, the exact shape of $Q_2(t)$ is needed. Moreover, the solution is more sensitive to fluctuations of demand along 2. In any case, the discussion given can be used as a guide for designing an adaptive control system which takes into account fluctuations of demand. For example, if spillback is to be avoided, the system must keep the size of queue along 2 below the critical size Q_2^* at all times.

VALUE OF PARKING SPACE

Let us try to get a gross estimate of the value of an increase of the parking space along the exit ramp, or equivalently of the critical queue size Q_2^*. An increase of Q_2^* by one car permits a reduction of the area A_2 (Fig. 2) by

$$\delta_2 \approx (t_h - t_g) \tag{9}$$

Assuming that spillback is to be avoided, the increase in Q_2^* will result in total reduction of delay for streams 2 and 3 equal to

$$\delta_{23} = (t_h - t_g)\left(\frac{s_3}{s_2} - 1\right) \tag{10}$$

Consider an example: assume $s_3/s_2 = 2$, $t_h - t_g = 1$ hour, and a cost of delay equal to $1.50 per hour. Also, assume that oversaturation occurs approximately once a day or, say 300 times per year. During a 30-yr amortization period, the increase in capacity by one car will result in a reduction of delay which may be valued at $13,500. A highway planner may take such an estimate into account in deciding whether to increase the capacity of an exit ramp or not. It should be pointed out, however, that the return per unit increase of Q_2^* diminishes as one approaches the maximum vertical distance between the curve Q_2 and the line O_2ef (Fig. 2). The decrease in rate of return is equal to the decrease of $(t_h - t_g)$, or roughly linear.

Incidentally, a substantial decrease in delay may also be accomplished by a drastic reduction of the average length of the automobile. This will be the case, provided that the overall performance of the automobiles is not affected by the reduction of their size, a conjecture which will be easily refuted by Detroit.

OPTIMUM LIGHT CYCLE

Up to this point, the light cycle at intersection 4 has been assumed constant. It should be interesting to find the value of the light cycle which optimizes the overall performance of the system in terms of delay. The light cycle influences delay in the following ways:

1. A long cycle decreases the delay by increasing the utilization rate of the cycle, assuming that the lost time due to acceleration and clearance is essentially independent of the light cycle length.
2. A long cycle increases the additional per cycle delays due to intermittent service. These delays are proportional to the sawtooth areas of Figure 4.
3. A long cycle decreases the effective parking capacity, Q_2^*, of the exit ramp. This is so because Q_2^* is the actual capacity of the exit ramp minus the extra queue length built during the red phase of the cycle along 2. A decrease of Q_2^* causes additional delays as seen in the preceding section.

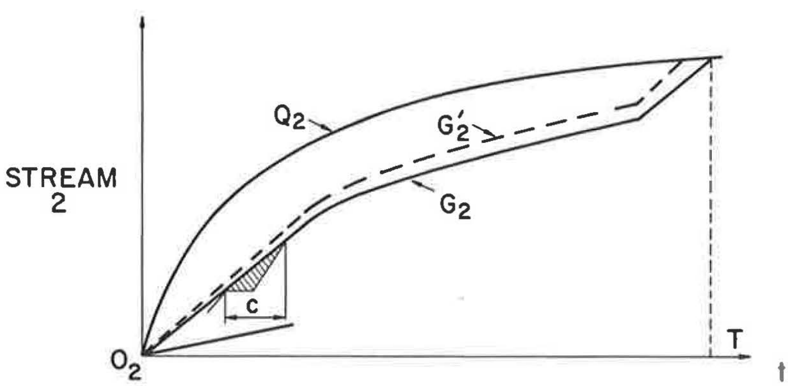


Figure 4. Influence of a variation of the light cycle on the aggregate delay of stream 2.

The optimum cycle is such that small variations about its value produce essentially zero variation of delay due to the three previous factors. An estimate of the optimum light cycle is obtained as follows:

Let the cycle be denoted by c, the total lost time per cycle by L, and the percentage of effective green allocated to direction 2 by p(t), where t is the time. The service rates along 2 and 3 are

$$\bar{\gamma}_2(t) = (1 - L/c)\, s_2\, p(t) \tag{11a}$$

$$\bar{\gamma}_3(t) = (1 - L/c)\, s_3[1 - p(t)] \tag{11b}$$

The decrease of delay due to an increase of the light cycle, Δc, is approximately equal to

$$(\Delta D)_a = \Delta c \left\{ \frac{LT}{2c^2} \int_0^T [\, s_2 p + s_3 (1 - p)\,]\, dt \right\} \tag{12}$$

where T is the duration of the rush period. It is assumed that p(t) and T are essentially unaffected by a small change Δc. An approximate value of the integral in Eq. 12 is $(s_2 + s_3)T/2$, assuming a more or less symmetric distribution of the values of p(t) about the value $1/2$, during T. Using this approximation,

$$(\Delta D)_a \approx \frac{LT^2}{4c^2} (s_2 + s_3)\, \Delta c \tag{13}$$

The delay corresponding to the sawtooth area of Figure 4 is

$$a_2 = \frac{1}{2}\, p(1 - p)\, s_2\, (c - L)\, c \tag{14a}$$

per cycle, for stream 2, and

$$a_3 = \frac{1}{2}\, p(1 - p)\, s_3\, (c - L)\, c \tag{14b}$$

per cycle, for stream 3. An increase in c produces an increase of this delay equal to

$$(\Delta D)_2 = \Delta c \left\{ \frac{s_2 + s_3}{2} \int_0^T p(1 - p)\, dt \right\} \tag{15}$$

We need an estimate for the integral of Eq. 15. The integrand is equal to 0.25, for p = 0.5 and 0.09 for p = 0.1. In view of the fact that the optimum control calls for extreme values of p, we shall assume (a better estimate may be obtained if one has, from a trial solution, a good approximation for the function p(t)) for the integral, the value 0.12 T, in which case

$$(\Delta D)_b = 0.06\, (s_2 + s_3)\, T \Delta c \tag{16}$$

Finally, the decrease in Q_2^* due to an increase dc is equal to

$$\Delta Q_2^* = \Delta c \left[\, p(1 - p)\, s_2 \right] \tag{17}$$

Hence, according to the preceding section, assuming

$$t_h - t_g \approx T \tag{18}$$

we find an increase in total delay

$$(\Delta D)_c = \Delta c \left\{ (s_3 - s_2) \int_0^T p(1 - p)\, dt \right\} \approx \Delta c \left[0.12 (s_3 - s_2)\, T \right] \qquad (19)$$

Now setting

$$(\Delta D)_b + (\Delta D)_c - (\Delta D)_a = 0 \qquad (20)$$

we obtain an equation for c, namely,

$$- \frac{LT^2}{4c^2} (s_2 + s_3) + (0.06)\, (s_2 + s_3)\, T + (0.12)\, (s_3 - s_2)\, T = 0 \qquad (21)$$

which yields

$$c \approx \left[\frac{LT\, (s_2 + s_3)}{0.24 (3 s_3 - s_2)} \right]^{1/2} \approx 0.2 \left[LT \frac{\rho + 1}{3\rho - 1} \right]^{1/2} \qquad (22)$$

where $\rho = s_3/s_2$. For example, assuming $\rho = 2$, $T = 1$ hour, and $L = 4$ sec, we find $c \approx 3$ min.

It will be noted that although the exact value of the coefficient multiplying the square root in Eq. 22 will vary after a more accurate computation of the integrals of Eqs. 12 and 15, the dependence of c on the square root of T and L remains as an intrinsic feature of the present theory. It should be remarked that the capacity of the exit ramp imposes an upper limit on the light cycle which may be of primary importance in the case of a very short ramp. Thus, the queue buildup during red must not exceed the actual ramp capacity. If this capacity is Q_{max}, then

$$Q_{max} \geq p(1 - p)\, s_2\, (c - L) \qquad (23)$$

Hence,

$$c \leq L + \frac{Q_{max}}{s_2\, p(1 - p)} \qquad (24)$$

For example, assuming $p = 0.2$, $Q_{max} = 10$ cars, $s_2 = 0.3$ cars/sec, and $L = 4$ sec, we find that c must be at most $3\frac{1}{2}$ minutes.

The preceding discussion assumes, of course, that the queue 2 can be served critically by an appropriate choice of p.

CONCLUDING REMARKS

The discussion is based on the assumption that the system comprising the expressway, the ramp, and the highway is isolated from other oversaturated regions. If this is not the case, one must consider an enlarged oversaturated system which can be considered isolated. For example, the stream 3 along the highway may contain a large amount of traffic coming from an exit ramp of the direction of the expressway opposite to direction 1. In this case, both exit ramps may be likely to produce spillback if the expressway is heavily traveled both ways. One general rule, in such a case, is that queueing can be permitted where there is greater parking capacity. A more detailed investigation is needed to determine where spillback, if inevitable, must be allowed in order to minimize the delay in the entire system.

Perhaps an explanation is due regarding the meaning of the demand curve $Q_2(t)$ used in this paper. This curve represents all the cars which would have demanded ramp service at time t, had the approach to the ramp been completely open. In practice, if spillback takes place, a large number of these cars will be mingled with through traffic in the queue formed along the expressway. Therefore, care must be exercised in ascertaining the appropriate value of $Q_2(t)$ by observing the composition and length of the queue along the expressway.

REFERENCES

1. Gazis, D. C., and Potts, R. B. The Oversaturated Intersection. Proc., Second International Symposium on Traffic Theory (to be published).
2. Gazis, D. C. Optimum Control of a System of Oversaturated Intersections. Operations Research, 12, 815-831 (1964).