TRANSPORTATION RESEARCH

RECORD

No. 1324

Highway Operations, Capacity, and Traffic Control

Communications, Traffic Signals, and Traffic Control Devices 1991

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1991

Transportation Research Record 1324
Price: $26.00

Subscriber Category
IVA highway operations, capacity, and traffic control

TRB Publications Staff
Director of Publications: Nancy A. Ackerman
Senior Editor: Naomi C. Kassabian
Associate Editor: Alison G. Tobias
Assistant Editors: Luanne Crayton, Norman Solomon
Graphics Coordinator: Diane L. Snell
Production Coordinator: Karen S. Waugh
Office Manager: Phyllis D. Barber
Production Assistant: Betty L. Hawkins

Printed in the United States of America

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

Communications, traffic signals, and traffic control devices, 1991.
p. cm.—(Transportation research record ISSN 0361-1981 ; no. 1324)
ISBN 0-309-05170-3
1. Traffic engineering. 2. Highway communications. 3. Traffic signs and signals. I. Series: Transportation research record ; 1324.
TE7.H5 no. 1324
[HE333]
388 s—dc20 92-5623
[388.3′ 122] CIP

Sponsorship of Transportation Research Record 1324

GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF TRANSPORTATION FACILITIES
Chairman: H. Douglas Robertson, University of North Carolina–Charlotte

Facilities and Operations Section
Chairman: Lyle Saxton, Federal Highway Administration

Committee on Communications
Chairman: Philip J. Tarnoff, Farradyne Systems Inc.
Secretary: Gerard J. Kerwin, New Jersey Department of Transportation
Walter A. Albers, Jr., E. Ryerson Case, Kan Chen, Min I. Chung, Robert L. French, Charles J. Glass, David W. Goettee, L. F. Gomes, Robert L. Gordon, Kevin Kelley, Wesley S. C. Lum, Roger D. Madden, Said Majdi, Frank J. Mammano, Corwin D. Moore, Jr., Michael A. Perfater, John J. Renner, Richard E. Stark, S. J. Stephany, Robert B. Weld

Committee on Traffic Control Devices
Chairman: Jonathan Upchurch, Arizona State University
Secretary: W. Scott Wainwright, Montgomery County Department of Transportation
Ronald M. Cameron, Robert L. Carstens, Benjamin H. Cottrell, Jr., Charles E. Dare, P. Norman Deitch, Robert E. Dewar, Paul H. Fowler, Robert L. Gordon, Robert David Henry, Richard P. Kramer, Feng-Bor Lin, Richard W. Lyles, Hugh W. McGee, Zoltan A. Nemeth, Errol C. Noel, A. Essam Radwan, Lewis Rhodes, Robert K. Seyfried, Harry B. Skinner, Howard S. Stein, Dwight L. Stevens, James A. Thompson

Committee on Traffic Signal Systems
Chairman: Herman E. Haenel, Kimley-Horn & Associates Inc.
Secretary: Alberto J. Santiago, Federal Highway Administration
William D. Berg, A. Graham Bullen, E. Ryerson Case, Edmond Chin-Ping Chang, David J. Clowes, Roy H. Fielding, Nathan H. Gartner, Peter Hakkesteegt, H. Milton Heywood, Paul P. Jovanis, James H. Kell, Les Kelman, Ken F. Kobetsky, Joseph K. Lam, Feng-Bor Lin, Adolf D. May, Jr., James L. Powell, Raymond S. Pusey, Dennis I. Robertson, Lionel M. Rodgers, Stephen Edwin Rowe, Philip J. Tarnoff, James A. Thompson, Charles E. Wallace, Roy L. Wilshire

Richard A. Cunard, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1990.

Transportation Research Record 1324

Contents

Foreword

The papers in this Record discuss some of the problems and issues facing urban traffic engineers as they try to optimize transportation systems to address increasing traffic congestion. From determining the communications needs for the development and implementation of the intelligent vehicle-highway systems (IVHS) to operating traffic signal systems safely and efficiently, the papers in this Record provide information and guidance.

Readers with a specific interest in IVHS communication needs will find papers pertaining to the development of traffic message channels for providing up-to-date traffic information to motorists. Issues related to the development of IVHS, particularly advanced vehicle control systems, are also discussed.

Readers with an interest in traffic signal control and traffic signal systems will find papers about choosing appropriate left-turn signal phasing, use of speed-location diagrams for the intergreen interval, traffic signal progression systems and demand-responsive systems, development of knowledge-based expert systems for isolated intersection signal control, evaluation of isolated traffic signal delay models for use with isolated at-grade light rail crossings, and development of a downtown area traffic signal system using distributed processing in modular traffic signal systems. In addition, a paper on the operational effectiveness of post-mounted delineators and retroreflective raised pavement markers is included.

Field Trials and Evaluations of Radio Data System Traffic Message Channel

PETER DAVIES AND GRANT KLEIN

The Radio Data System Traffic Message Channel (RDS-TMC) will be introduced in Europe in the mid-1990s. RDS provides for the transmission of a silent data channel on existing VHF-FM radio stations. TMC is one of the remaining RDS features still to be finalized. It will enable detailed, up-to-date traffic information to be provided to motorists in the language of their choice, thus ensuring a truly international service. As part of the European DRIVE program, the RDS-ALERT project has carried out field trials of RDS-TMC. Testing was undertaken prior to and during the RDS-ALERT project, and implications for the TMC service throughout Europe were considered. TMC offers an exciting prospect of a practical application of information technology suitable for the 1990s and into the next millennium. Further TMC developments will provide interfaces linking it to other intelligent vehicle-highway system technologies currently under development in the United States, Europe, and Japan.

Traffic congestion is one of the most serious problems affecting transportation in the United States today. The volume of traffic is increasing at an alarming rate and will continue to rise into the next century. These problems are, of course, not unique to the United States. Growth in international traffic is increasing congestion throughout Europe as the border controls between European Community member states are reduced.

A standardized system for providing detailed, up-to-date traffic information to motorists is paramount in an integrated approach to solving these congestion problems. Provision of this information in Europe must allow for the different languages spoken in the member states. The ability to present detailed information in the driver's preferred language would greatly enhance the effectiveness of a traffic information system.

The Radio Data System Traffic Message Channel (RDS-TMC) will provide such a traffic information service to European motorists before the end of the century. The RDS facility, defined by a European Broadcasting Union (EBU) specification (*1*), provides for the transmission of a silent data channel on existing VHF-FM radio stations. Its primary purposes are to identify radio broadcasters and to allow self-tuning receivers to automatically select the strongest signal carrying a particular program. One of the most popular RDS facilities is the program service name. This facility gives the listener a display of up to eight characters showing the name of the program being received, such as "BBC R4" for Radio 4 in the United Kingdom or, potentially, "WCXR" for the Washington classic rock station.

Castle Rock Consultants, Heathcoat Building, Highfields Science Park, University Blvd., Nottingham NG7 2QJ, England.

Many RDS features have already been defined and implemented in most parts of Europe. One additional feature of RDS not yet finalized is the Traffic Message Channel (TMC) for digitally encoding traffic information messages. Group Type 8A, one of 32 possible RDS data groups, has been reserved for the TMC service. It will provide continuous information to motorists through a speech synthesizer or text display in the vehicle. TMC will improve traffic data dissemination into the vehicle by several orders of magnitude over conventional spoken warnings on the radio. By linking with intelligent vehicle-highway system (IVHS) technologies, TMC will interface with on-board computers, creating an additional tool for direction and control of traffic movements.

BACKGROUND

The concept of using a subcarrier to convey additional information on a VHF-FM broadcast dates back many years. By the mid-1960s, many FM stations in the United States used subcarriers to convey a subsidiary audio program signal. This "storecasting" was used to play background music in restaurants and shops (*2*). These subcarrier systems were subject to a Federal Communications Commission regulation known as Subsidiary Communications Authorization. These systems were not suitable for use in Europe because of the level of crosstalk from the subcarrier into the main audio program.

The Autofahrer Rundfunk Information (ARI) system was developed in West Germany in the early 1970s. ARI is a relatively simple tone-signaling system which requires only a simple decoder. It indicates which programs carry traffic announcements as part of the audio, when a traffic announcement is currently being broadcast, and the geographical area to which the announcement applies.

By the mid-1970s several European organizations were working toward the development of an FM subcarrier system using data to modulate the subcarrier. In 1978, EBU began working toward a standard for a station and program identification system for FM broadcasts. This work resulted in the RDS specification in 1984 (*1*).

The basic structure of the RDS service is shown in Figure 1. The data rate for RDS is 1187.5 baud, divided into groups of 104 bits. Each group is made up of four blocks of 26 bits each, of which 10 bits are used as a checkword. An RDS group is the smallest package of data that can be defined within the system. RDS is broadcast using a 57-kHz subcarrier; when an ARI signal is broadcast on the same subcarrier, the two signals are broadcast 90 degrees out of phase.

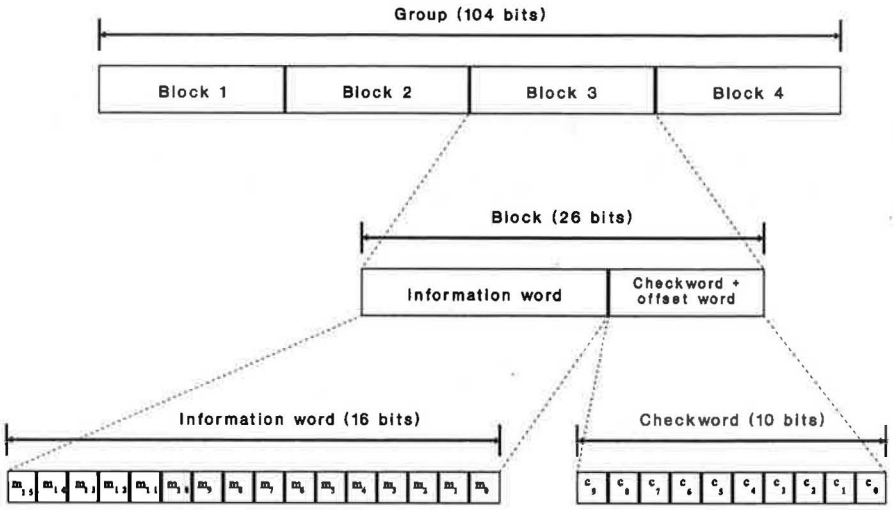


FIGURE 1 RDS group structure.

From the outset of the RDS development, it was foreseen that the system might be used as a channel to convey traffic messages. Three broad approaches were initially investigated for this service. They considered text-to-speech, phoneme transmission, and the use of a fixed repertoire of speech synthesizer messages. The third option, which uses a fixed repertoire of basic traffic messages stored as a dictionary in the receiver, was considered the most practical. The addition of supplementary dictionaries to include a store of place names would be possible by mass storage such as CD-ROM.

Since the development of the RDS specification, a number of proposals for the TMC feature have been made (*2*). The Dutch RVI project was one of the first. It began to investigate broadcasting traffic information via RDS in 1985. The project comprised the definition and implementation of a prototype system and the development of technical facilities to transmit traffic information using RDS. The RVI demonstration concluded that although RVI was not the optimal solution for a TMC service, it would be feasible to use speech synthesizers in vehicles to introduce a similar service throughout Europe.

The second main development in TMC came from Philips and Blaupunkt, who had been working separately on concepts for broadcasting traffic information via RDS. In 1986, the two manufacturers started to work together to develop a common proposal for the traffic message format. The proposed structure used a two-group RDS message to give incident cause, effect, location, and stop time, and to provide advice for the driver. This proposal was presented to the European Conference of Ministers of Transport (ECMT) in 1987.

From early 1987, TMC research in West Germany concentrated on reducing the redundancies in previous proposals. Blaupunkt and the West German Ministry of Transport (BAST) worked on a coding structure based on the messages used in the ARI system. The result was a one-group structure incorporating 90 percent of the ARI messages. This development increased the efficiency of earlier coding structures, but considered only those messages required in West Germany.

The CARMINAT project was launched in 1986 to investigate technological advances for improving car travel. The project focused on the implementation of an experimental area in France and was intended to evaluate and validate prototype systems. A protocol was developed to incorporate the traffic data collection, message generation, and message broadcast features of an RDS-TMC service. This protocol and others developed in CARMINAT are still evolving through demonstrations in the experimental area.

The Commission of the European Communities funded a study in 1987 and 1988 to examine how actual traffic messages would fit into a standard RDS-TMC format. A primary objective of the study was to bring together all of the national traffic messages developed to date and set up an agreed international set. An important aspect of the research was to canvass opinion in order to get a representative range of views.

The study, carried out by Castle Rock Consultants, led to a final report on the so-called CRC protocol for TMC in October 1988 (*3*). It defined an allocation of bits within RDS-TMC that offers the flexibility for covering a full range of situations while retaining the potential benefits of single-group messages for the majority of actual events. The first objective of the study was to resolve location coding issues. Second, it defined a comprehensive message coding structure to allow for basic message texts, advice, and message quantification such as queue length or incident duration. The third area of the study addressed the message management aspects of RDS-TMC.

The CRC protocol significantly increased the efficiency of the message coding developed in previous proposals. In conjunction with the revised location codes, it allowed the majority of messages to be broadcast using a single RDS group. The few multigroup messages containing additional, detailed information would be wholly compatible with the recommended one-group structure through the use of virtually identical fields in the first group.

RDS-ALERT

Leading on from the TMC coding studies, the European Community's DRIVE (Dedicated Road Infrastructure for Vehicle Safety in Europe) program addressed the development of

RDS-TMC. DRIVE I was a precompetitive, prenormative research and development program involving collaboration among over 1,000 transportation-related experts from 300 European organizations (*4*). Its conceptual goal was a fully integrated road transport environment (IRTE) in which intelligent vehicles are linked to an intelligent road network. This goal was to be reached by developments in information technology and telecommunications applied to road transport. Such technological developments are called Road Transport Informatics (RTI). DRIVE II will focus on field trials of the various technologies.

RDS-ALERT (Advice and Problem Location for European Road Traffic) was one of about 70 projects within DRIVE I. Its goal was establishing standards in RDS-TMC location coding, message content, and message management acceptable throughout Europe. An additional objective of the project was to ensure compatibility with other in-vehicle equipment. RDS-ALERT was undertaken by a consortium of manufacturers and broadcasters coordinated and led by Castle Rock Consultants. It was a 2-year project which started at the beginning of 1989.

Within RDS-ALERT, current proposals for single-group and multigroup messages were reviewed and evaluated in order to reach a starting consensus for experimental evaluation. A liaison group structure was established with representatives from EBU, ECMT, and the RDS-ALERT consortium to ensure the necessary integration with other RTI developments and among principal actors in RDS-TMC. Liaison was carried out through this group, with other programs, and through contacts with traffic authorities and broadcasters throughout Europe.

The RDS-ALERT project sought to establish internationally accepted standards in RDS-TMC coding, format, and operation for use as part of a road traffic information system. In finalizing a standard protocol for approval through this project, a comprehensive, flexible, and efficient approach has been sought. *Comprehensive* coding ensure that all required locations and messages can be coded. *Flexible* coding deals with current situations and allows for future changes in the highway networks. *Efficient* coding enables the limited capacity available for RDS-TMC to be used to its best advantage.

One of the major goals of the project was to provide for maximum flexibility of message coding in order to permit future, but unforeseen, TMC developments. Also, the project aims to develop consistent and agreed location coding strategies for countries wanting to implement RDS-TMC. In developing standards, RDS-ALERT defined optimal message management strategies and limits to permissible variations, taking into account the different situations of participating countries.

The RDS-ALERT project made substantial progress toward developing an international RDS-TMC standard. Field tests were carried out to evaluate the RDS broadcast and reception conditions in a number of countries. These field tests demonstrated the requirements for an efficient TMC coding structure. Building on the work of Castle Rock Consultants (*3*), a message set was derived to incorporate the messages required in each country interested in implementing an RDS-TMC service. This development is at the stage where messages can be input quickly and easily, resulting in a final RDS code for broadcast.

The final field trials in RDS-ALERT started early in 1990. These field trials, carried out on test routes in southeast England, used signal-quality results from a number of European countries to ensure that reception conditions were representative of those experienced on the European highway network. The results of the field trials were used to evaluate all aspects of the proposed TMC protocol. The draft protocol has been drawn up and circulated as part of the consensus-building process. This draft will require final modifications and refinements before full documentation of the protocol can be submitted to the relevant standard-setting bodies.

FIELD TRIALS

Testing of RDS reception has been carried out under many conditions in recent years. In 1987, Philips carried out extensive RDS field tests in West Germany (*2*). These tests evaluated the percentage of RDS blocks received with and without error-correction facilities, and looked at the loss of block reception prior to synchronization onto the RDS data stream. The test results showed that between 62 and 85 percent of blocks were received correctly. Using the error-correction facilities provided for in RDS, the number of correctly received blocks increased to between 73 and 89 percent.

In order to decode a single-group TMC message, all four blocks from the group must be correctly received. Because of the nature of the RDS data stream, bit errors are more likely to come in bursts than to be randomly distributed. Hence, if one block is received correctly, the probability of correctly receiving the subsequent block is increased. In Germany, BAST carried out studies of the RDS reception characteristics for single-group and multigroup TMC messages. The results of these studies showed that 80 percent of single-group messages could be correctly received under good reception conditions, decreasing to 18 percent for four-group messages.

The RDS and ARI signals are both broadcast on a 57-kHz subcarrier. When both systems are provided on the same station, interference is kept to a minimum by broadcasting the two signals 90 degrees out of phase. However, the two signals still interfere in mountainous areas where multipath effects are experienced. The level of interference for areas with both RDS and ARI broadcasts was tested by Bosch. These tests indicated that the ARI signal reduces the proportion of error-free blocks by up to 20 percent.

The RDS-ALERT project included four phases of field trials. The first phase was carried out in Germany using a professional Blaupunkt receiver and Schuemperlin decoder. The second phase of the testing took place in France with 10 transmitters equipped to broadcast RDS-TMC data. The third testing phase evaluated aspects of the TMC protocol based on the ISO seven-layer communications model. Finally, the fourth phase investigated RDS reception conditions in a number of countries over various terrains.

Phase One

Phase One tests used static data transmission techniques to broadcast a fixed TMC data set. These techniques allowed

the same sequence of messages to be used under varying reception conditions. On test routes south of Hanover, Germany, results were recorded for

- Field strength,
- Signal quality,
- Group and block reception statistics, and
- Time for message reception.

The tests showed that even under relatively poor reception conditions it would be possible to receive TMC messages, although the delay might be as much as 5 minutes between initial broadcast and correct decoding in the vehicle (*5*). This finding highlighted the requirements for an efficient coding structure with as many messages as possible broadcast in a single RDS group.

Phase Two

Phase Two testing was carried out south of Paris, France. Traffic information was broadcast in real time using a draft TMC protocol. The tests examined

- Reception quality,
- Proportion of messages received incorrectly,
- Proportion of messages not received, and
- Data flow requirements for a fully implemented system.

These test results noted significant differences in reception conditions between urban and rural areas. The proportion of blocks correctly received was reduced by about 5 percent on entering an urban area (*6*). Overall, however, the results were favorable for successful operation of a TMC service.

Phase Three

Testing in *Phase Three* was based on test routes in southeast England. The field testing used wide-band recorders to record the full FM multiplex from two stations, both on-air in a test vehicle and directly from the 240-kW transmitter. Comparison of the on-air and direct recordings gave an error file that can be applied to any RDS data stream to simulate real conditions. The error file could therefore be used to evaluate draft TMC protocols under identical conditions. Preliminary results from these tests included the following findings:

Number of RDS Groups per Message	*Percentage of Messages Correctly Received*
0.5	86
1	84
2	77

These findings relate to the condition with no ARI signal on the same broadcast and with a deviation (width of the RDS signal, usually between 1.2 and 7.0 kHz) of 2 kHz. The tests indicated that no false messages were received, implying that all corrupted groups were rejected.

Phase Four

In *Phase Four*, Castle Rock Consultants conducted RDS signal-quality tests under the varying reception conditions experienced across Europe. Results were recorded with RDS decoders from Blaupunkt and Philips. The tests examined the range of field strengths and data error rates experienced in the areas where TMC will be implemented.

The testing using the Blaupunkt receiver covered Switzerland, Belgium, Germany, the United Kingdom, France, and the Netherlands. These tests were conducted in mountainous regions as well as flat land. Also, a number of RDS deviations were included in the tests, ranging from 1.2 kHz in Germany to 7 kHz in parts of France. The deviation refers to the width of the RDS signal on either side of the 57-kHz subcarrier. The tests in Germany covered radio stations using the ARI system as well as RDS.

The results from the Blaupunkt decoder comprised field strength and error rate readings along each of the routes for a particular program. About 8,000 readings were taken using this decoder, each providing an average over a 20-sec period. The results of the field strength tests are shown in Figure 2. The figure shows the percentage of readings in each country at particular field strengths. The service area for an FM transmitter is defined in the United Kingdom as the area in which the field strength is at least 54 dB relative to 1 uV/m (*7*). Figure 3 shows the bit error rate results for the same routes. The results shown in both figures are summarized in Table 1. The following paragraphs briefly analyze these results.

A number of important conclusions can be drawn from the field strength results. The mean field strengths show low values for France and the Netherlands and higher values for the other countries. Similarly, the variation in the field strengths for France and the Netherlands are larger. The main factors causing this variation are the transmitter power and spacing. In France, radio transmitters are not significantly lower in power than the rest of Europe, but their spacing is considerably greater.

The distributions differ significantly for the other four countries, although the mean field strengths are similar (59.3 to 60.0 dB). The United Kingdom field strengths, taken mostly from BBC stations, have the most consistent values with a standard deviation of 5.9 dB. Switzerland, Belgium, and Germany have standard deviations between 6.8 and 7.5 dB. Again, transmitter power and spacing are the main parameters affecting these results. The United Kingdom is covered by the BBC by over 30 main transmitters with powers up to 250 kW. In addition, nearly 100 smaller transmitters cover areas missed by the main network. The other countries considered do not have such comprehensive coverage.

The bit error rate results show some different trends. Generally, the bit error rates are expected to follow trends opposite those of the field strengths. However, in Switzerland and Germany, both the field strengths and the bit error rates are some of the highest. Also, although France and the Netherlands have relatively low field strengths, their bit error rates are around the average for all routes.

Terrain is one of the factors affecting the bit error rates in RDS reception. The Swiss and German results were obtained mostly from mountainous regions, typical of those countries.

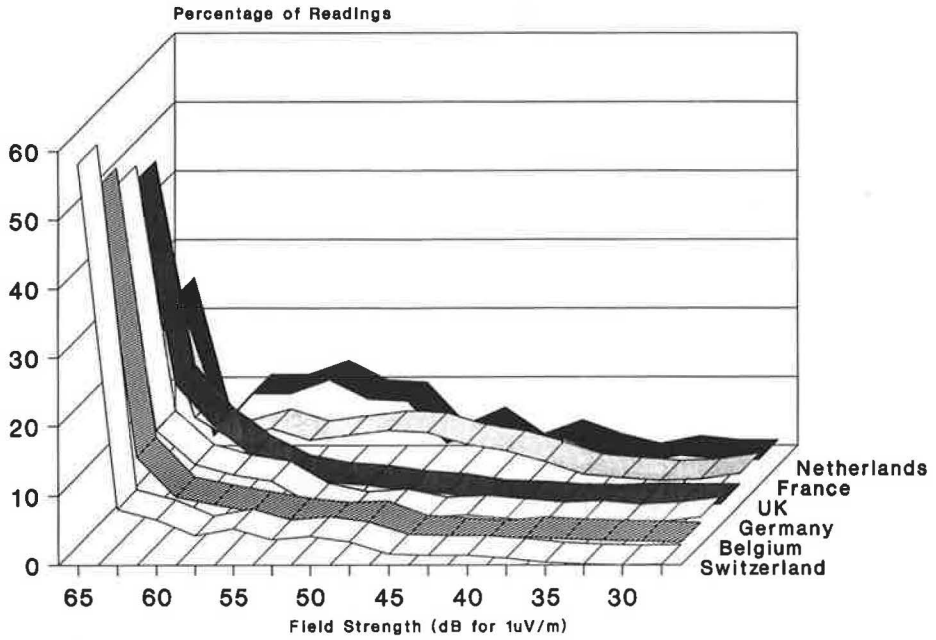


FIGURE 2 RDS field strength results.

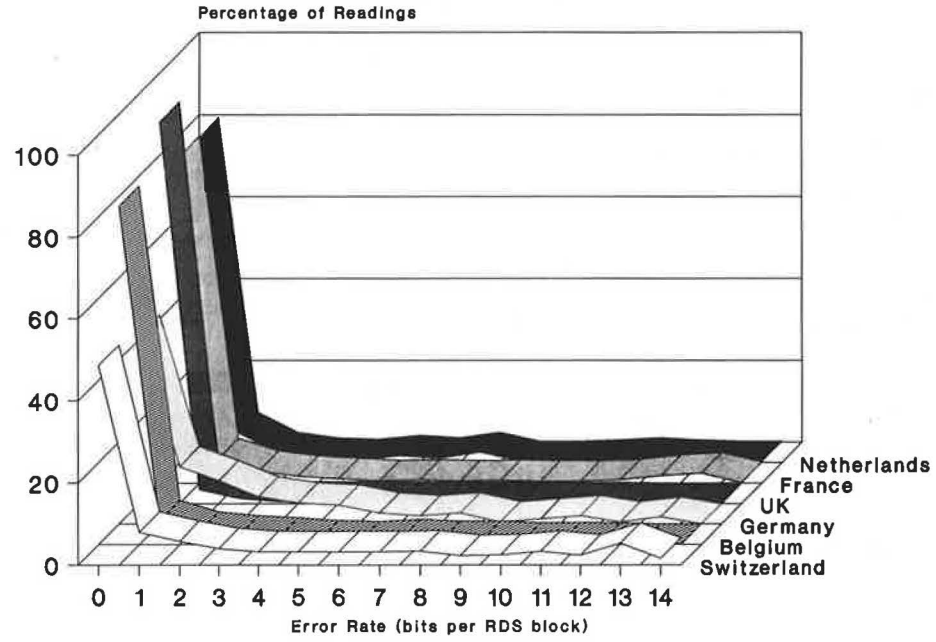


FIGURE 3 RDS bit error rate results.

This type of terrain results in multipath effects caused by signal reflections. Although multipath does not significantly affect the field strength, it can cause data errors. Conversely, the Netherlands and much of France consists of relatively flat land, resulting in lower bit error rates than normally expected for their mean field strengths.

The second main parameter affecting RDS bit error rates is the deviation used. In Germany the deviation is set at 1.2 kHz; in France many stations use up to 7 kHz. A high deviation means a wide RDS signal and therefore a lower probability of receiving data errors. The low deviation value is used in Germany to reduce the interference between the RDS and ARI signals. This combination of low deviation and ARI interference makes German reception conditions among the most testing in Europe.

The two receiver manufacturers in the RDS-ALERT consortium, Philips and Blaupunkt, adopt different transmitter switching strategies in their existing RDS receivers. The Blaupunkt receiver periodically tests some of the alternative frequencies in order to determine the strongest signal carrying a particular station. It switches to an alternative frequency only when the alternative has shown a consistently higher field

TABLE 1 SUMMARY OF RDS FIELD STRENGTH AND BIT ERROR RATE RESULTS

Country	Mean field strength (dB)	Standard deviation	Mean error rate (Bits/ block)	Standard deviation	Number of readings
Switzerland	59.8	6.8	3.3	4.4	910
Belgium	59.3	7.5	0.6	1.8	1415
Germany	59.3	7.3	2.2	3.1	409
UK	60.0	5.9	0.2	1.0	2383
France	54.3	9.2	1.1	3.0	2829
Netherlands	54.8	8.3	0.9	2.5	355
All	57.7	7.7	1.0	2.6	8301

strength. The Philips receiver, on the other hand, tends to switch to an alternative frequency if the alternative has shown an instantaneously higher field strength.

Although both of these strategies have advantages for audio reception, neither of them is the optimal solution for TMC frequency switching. The Blaupunkt strategy would result in TMC messages transmitted from behind the vehicle. These messages would therefore relate to an area already passed through by the vehicle. The Philips strategy would overcome this problem to a degree by switching more readily. It would, however, be less efficient in that each frequency switch results in data losses before continuity can be gained.

Figure 4 shows results of tests using the Philips receiver to monitor four transmitters broadcasting the same BBC program in north England. Figure 5 shows the route followed during the test. The results highlight some of the disadvantages of each switching strategies. The Blaupunkt strategy would keep the receiver tuned to the Keighley transmitter until around 25 miles. It would then tune to Holme Moss and remain on that until the end of the route. The Philips strategy would tune to Holme Moss at 18 miles, but would also select Wharfedale for miles 49 and 50. The ideal strategy, however, would tune to Holme Moss at 18 miles, but would ignore the brief field strength peak from the Wharfedale transmitter because this would cause two data continuity losses within a short time.

This brief analysis highlights some of the differences in RDS reception across Europe and in current RDS receiver capabilities. TMC will need to operate successfully under all of these conditions. The results indicate the variations within

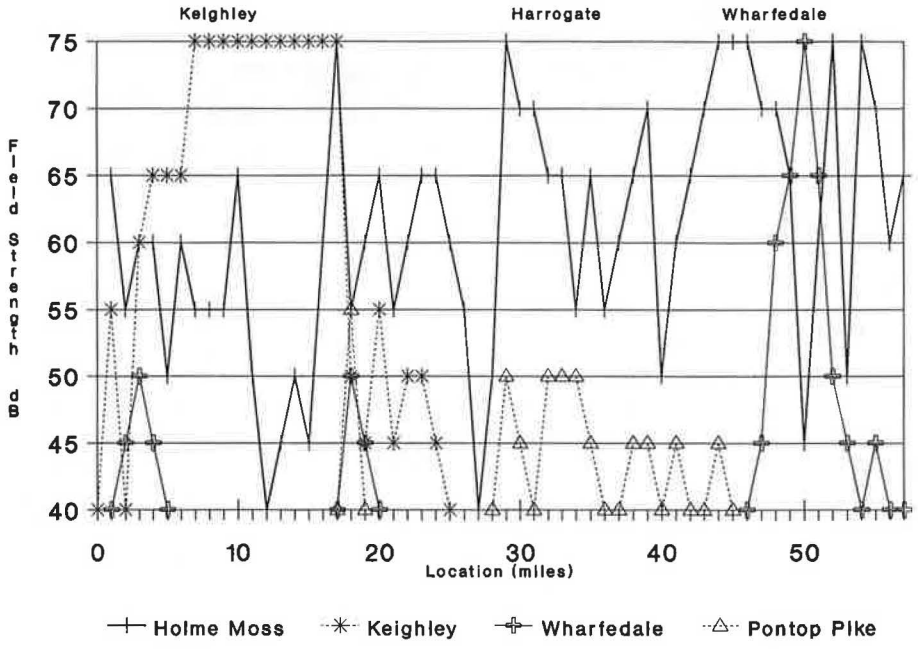


FIGURE 4 RDS test results.

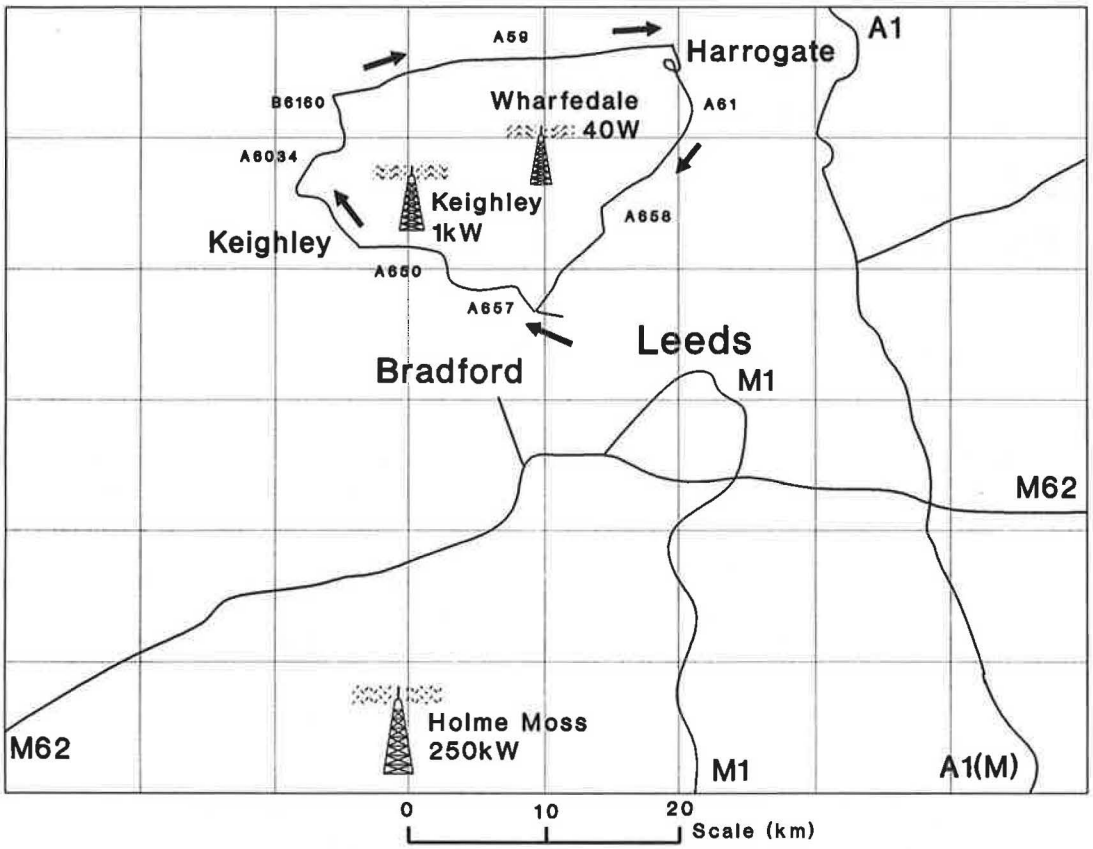


FIGURE 5 Map of RDS test circuit.

each country and throughout Europe. However, more detailed analysis is still required to investigate the rate at which signals reduce in strength over distance in order to develop optimal transmitter switching strategies for TMC receivers. Castle Rock Consultants has performed some analyses of these strategies but further work is still required to determine a near-optimal solution.

CONCLUSIONS

The TMC protocol has evolved through tests and evaluations in the RDS-ALERT project. Final modifications of the protocol will significantly increase its efficiency and flexibility, while maintaining a comprehensive structure capable of satisfying the requirements of a driver information system suitable for the 1990s and into the next millennium. RDS-TMC offers the exciting prospect of a practical application of information technology on the verge of implementation.

ACKNOWLEDGMENT

The RDS-ALERT project is being undertaken by a consortium of partners on a cooperative basis. The input of all partners to the project is duly acknowledged.

REFERENCES

1. European Broadcasting Union. *Specification of the Radio Data System (RDS) for VHF/FM Sound Broadcasting*. Doc. Tech. 3244-E, European Broadcasting Union, Brussels, Belgium, March 1984.
2. RDS-ALERT Consortium. *Review of Progress to Date on RDS-TMC*. DRIVE V1029, Deliverable 1, June 1989.
3. Castle Rock Consultants. *Radio Data System (RDS) Traffic Message Channel (TMC)*. Final Report to the Commission of the European Communities, Nottingham, England, Oct. 1988.
4. Commission of the European Communities. *The DRIVE Programme in 1989*. DRI 200, Brussels, Belgium, March 1989.
5. RDS-ALERT Consortium. *Report on Message Coding Field Tests*. DRIVE V1029, Deliverable 9, Jan. 1990.
6. RDS-ALERT Consortium. *Report on Message Reception Field Tests*. DRIVE V1029, Deliverable 11, Jan. 1990.
7. *Radio Transmitting Stations*. BBC Engineering Information, BBC Engineering Publications, London, England, 1989.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the consortium of partners in the RDS-ALERT project.

Publication of this paper sponsored by Committee on Communications.

Issues in Communication Standardization for Advanced Vehicle Control Systems

STEVEN E. SHLADOVER

Several issues must be addressed when establishing intelligent vehicle-highway system (IVHS) communication standards, particularly for the most advanced of these systems, the advanced vehicle control system (AVCS). Each of the three stages of AVCS evolution have separate issues that must be addressed in order to define communication standards. It is recommended that measured progress toward the development of IVHS communication standards be based on a solid foundation so that they do not risk becoming impediments to progress. Given the current state of development of IVHS technology, more questions are raised than are answered.

The intelligent vehicle-highway system (IVHS) is an assortment of communication, computer, sensing, and control technologies that can be applied to observe, guide, or control the movement of vehicles in a traffic system. IVHS requires the operation of vehicles and roadways as a combined system rather than as separate entities. This requirement raises important issues about the information that must reside with each element of the complete system (individual vehicles, individual locations along the roadway, and a central wayside location) and that must be communicated from element to element. The complexity of the road transportation system and the mixture of public and private sector interests virtually mandate that IVHS development and implementation will involve many different organizations. In order for these separate organizations to develop their respective portions of a combined system, standards must be established to govern the interfaces among those portions, particularly those involving the exchange of information.

Several communication standardization issues must be considered to meet the needs of the advanced vehicle control system (AVCS), the most advanced of the IVHS technologies under active consideration. These issues are being raised now so that the IVHS standards that are developed to meet relatively near-term needs will have sufficient flexibility and growth potential to meet the longer-term needs of AVCS.

IVHS developers must thoroughly understand the relative merits of diverse technical approaches and the inherent limitations of alternative technologies before applying the constraints inherent in standardization. If standards are imposed before this level of technical maturity is reached, the risk is high that the most promising solutions will be precluded by the standards, and the technology could be "dead ended." Of course, the costs of such a mistake would not become apparent until some years after the standards were imposed. At that time, either activity in the application field would decrease or all participants would be forced to undergo a costly and time-consuming retrenchment, in effect discarding their previous work and starting over under a new set of standards with a greater growth potential. Judicious timing of the decision to standardize can avoid such problems.

Institute of Transportation Studies, University of California, Berkeley, Calif. 94720.

INFORMATION NEEDS OF AVCS

All IVHS applications depend on timely availability of information, which must be supplied from one vehicle to another, from the roadway to the vehicle, or from the vehicle to the roadway. Definition of *communication* needs must follow definition of the *information* that must be supplied from one system element to another. That information may be transferred by means other than communication, such as direct sensing, so it is important to define the most appropriate means for transferring each category of information.

The communication needs of AVCS differ from the communication needs of the other IVHS functions in one fundamental way. Because virtually all AVCS communication is safety-critical, an AVCS communication failure would be likely to produce an accident with significant potential for property damage and injury. Therefore, AVCS communication links must incorporate redundancy in any of several forms (e.g., parallel communication links, multiple transmissions of data, or encoding schemes). The phenomena that AVCS systems address (vehicle dynamics) occur on time scales of fractions of a second; in contrast, phenomena that the other IVHS functions address are more likely to occur on scales of multiple seconds or minutes. Therefore, AVCS information changes more rapidly and, in turn, updates must be communicated more frequently. On the other hand, the quantity of information that must be communicated for each update may be substantially less than for the advanced traveler information system (ATIS), for example.

Because AVCS is not a single system, but rather a group of technologies, it has been subdivided into relatively homogeneous subfunctions. Mobility 2000 defined three levels of AVCS functionality: AVCS I, II, and III. The requirements of these subfunctions are cumulative rather than distinct, so that AVCS III incorporates all of the requirements of the two lower levels. AVCS I technology enhances the safety of a driver's responses to the road environment by offering per-

ceptual enhancements, warnings, and control or stability augmentation. AVCS II adds the capability for vehicles to operate under automatic lateral and longitudinal control on individual links in a network, to enhance both safety and link capacity. AVCS III extends these capabilities to comprehensive networks on freeways, so that vehicles can operate under full automatic control from the freeway entrance ramp to the exit ramp. Communication issues relevant to each of these three levels of AVCS operation must be addressed in the development of IVHS communications standards.

COMMUNICATION ISSUES RELEVANT TO AVCS I

Virtually all AVCS I functions are centered in the individual vehicles and can be applied anywhere in the road system, including freeways, rural roads, arterials, and local streets, without requiring infrastructure modifications. At early stages in the evolution of AVCS I, relatively few vehicles will be equipped with AVCS I technology. Therefore, encounters between vehicles will usually involve only the direct sensing of unequipped vehicles by equipped vehicles. As the AVCS I market penetration grows, the potential for interactions between equipped vehicles will increase and communication between those vehicles will become more of an issue.

The communication links relevant to AVCS I appear to be almost entirely vehicle-to-vehicle. Communications between vehicles and the roadway that could support the AVCS I warning functions could be incorporated within the longer time-scale functionality of ATIS or ATMS. These communications would be broadcast warnings of accidents or incidents ahead so that drivers could slow down or adjust their routes. Existing freeway traffic management systems communicate this information to drivers using changeable message signs, for example.

The information communicated from one vehicle to another in an AVCS I system would generally be a warning of a potentially unsafe condition, such as

- Watch out, I'm in your blind spot!
- I'm going to change lanes in front of you.
- I've just had a flat tire (or some other failure).
- I've just slammed on my brakes!

The information content of these messages is not large. Codes could be used to indicate the specific condition, identify the vehicle sending the message, and provide enough information (lane, direction, and milepost location) for the other vehicles to locate the transmitting vehicle. Although the message would be of very limited length (just a few bytes), it would have to be provided rapidly (perhaps within 100 ms), and with priority over other, less safety-critical information. The message would also have to be provided repeatedly within a short time so that if the first transmission fails, subsequent transmissions would have a high probability of success. The number and frequency of repetitions that would be needed to ensure adequate safety cannot be specified without substantial analysis and testing.

Numerous issues must be considered in defining the communication standards that would have to be applied even for this relatively simple case, including message length, repetition rate, range, selectivity, and reliability. Clearly, such issues require analysis, design, and testing. It is difficult to see how communication standards could be defined intelligently without resolving these issues.

COMMUNICATION ISSUES RELEVANT TO AVCS II

The extension to AVCS II functionality significantly increases the communication requirements. Two additional types of vehicle-to-vehicle communication are likely to be needed, as well as communication between the vehicle and the wayside computer systems. The AVCS II function assumes that suitably equipped vehicles would be operated in platoons on suitably equipped facilities, with very close longitudinal spacings between vehicles within the platoons. This close-formation operation would require rapid communication of information between the vehicles within a platoon in order to ensure stable platoon dynamics. Additional vehicle-to-vehicle communications would be needed to enable vehicles to safely enter and leave platoons.

Research on platoon dynamics indicates that each vehicle in a platoon would need continuing and timely information about the movements of its predecessor and the platoon leader (with updates perhaps as frequently as every 20 ms). Safe operation of platoons would also require that warnings of emergency conditions be "immediately" communicated from the failed vehicle to all the other vehicles in the platoon. Because of the very close spacings between vehicles in platoons (about 1 m), this requirement is likely to be more stringent in terms of response speed and reliability than for the analogous function in AVCS I. However, research has not yet been done to define how fast is fast enough.

Merging of vehicles into moving platoons and separation of vehicles from platoons introduce additional vehicle-to-vehicle communication needs. The quantity of information required for these purposes is not large compared to that required for the longitudinal control within the platoons, but the information generally comes from further away and from a direction where there is significant potential for interference and loss of line of sight (adjacent lane, substantially ahead of or behind the receiving vehicle).

The communications between vehicle and wayside for AVCS II introduce an element not present in AVCS I. These communications, primarily one-to-many communications of command information from the wayside to the vehicles, are generally meant to apply locally rather than globally. Therefore, wide-area broadcasting would not be an appropriate medium for this information, and a more selective medium would be needed. Many-to-one communication of information from the individual vehicles to the wayside control computers may also be needed. The amount of information to be received from each vehicle is not large, but the number of vehicles could be very large in a major metropolitan region.

It may be more appropriate to communicate the warnings about vehicle problems directly from vehicle to vehicle, without involving the wayside. Hopefully, the system may operate without requiring each vehicle to communicate its state information to the wayside on a regular basis. It may be ade-

quate for the wayside system to know each vehicle's position at relatively infrequent intervals (several seconds or even minutes), or it may not even be necessary for the wayside system to know any more than aggregate vehicle flow information. Considerable system engineering analysis and simulation is needed to establish the necessity of communicating each of these types of information from each vehicle to the wayside.

The types of general issues that need to be addressed to understand AVCS II communication needs include tradeoffs between sensing and communication, needed repetition rates for different messages, relative roles for vehicle-to-vehicle and vehicle-wayside communications, assignment of multiple priority levels to different messages, and spacings between wayside communication devices.

COMMUNICATION ISSUES RELEVANT TO AVCS III

The communication needs of AVCS III operations will include all of those for AVCS I and II, an additional set of vehicle-wayside communications associated with the system management functions of AVCS III, and a new element of wayside communications between distributed and central computing facilities. These new communication needs do not have the strong safety implications of their predecessors, but are in a sense more related to ATIS communications.

The new messages for AVCS III are all one-to-one communications of limited amounts of information. Individually each message would not impose a significant communication burden. However, in a large metropolitan system (e.g., Los Angeles), the number of vehicles involved could make this a very large communication burden. This factor leads to the consideration of highly distributed wayside control computers, each of which would have to communicate with only a limited number of vehicles. There would then be a significant amount of communication among these wayside computers and between each of them and the central coordinating computers. The system design implications of different configurations for these computer and communication systems are extremely complicated and have not yet been addressed at even the most rudimentary level. The appropriate distribution of wayside and vehicle control functionality, the depth of hierarchy in the system structure, and the implications for both computational and communication burdens will be understood only through substantial system engineering effort.

These questions address the central issues in system-level design and system management for an automated freeway system. These are probably the most technically difficult questions in the IVHS field, and will therefore require years of research to answer. Unfortunately, the range of possible answers, viewed at this stage in IVHS development, is so broad that it does not appear reasonable even to define order of magnitude bounds on the ensuing communication needs.

DIRECTION OF IVHS COMMUNICATION STANDARDS

Given the current state of development of the various IVHS technologies, what can be done now to move toward standardization of IVHS communications? It is clearly in the interest of the vendors of any of the IVHS components or systems to have standards developed as rapidly as possible to simplify product development and marketing tasks. On the other hand, so little is understood about the large-scale system implication of *any* of the IVHS technologies that it appears to be premature to define comprehensive communication standards at this time.

The pressure for standardization is not likely to relax in the face of these shortcomings in current knowledge, and indeed standardization should not have to wait until *all* technical uncertainties are resolved. The interesting challenge then is to try to develop standards frameworks with sufficient flexibility and growth potential to accommodate all reasonable future needs. It would be the height of folly to get locked into a set of IVHS standards that could meet the needs of IVHS applications only 5 or 10 years, rather than considering from the start the long-term evolutionary potential of IVHS and making sure that potential is not artificially constrained by insufficiently progressive standards.

It may be possible to embark on the road toward IVHS communication standardization once there is basic agreement about the choice of physical medium (e.g., radio, optics), the general network topology, and rough estimates of data traffic. At that point, it may be possible to address standardization of packet and frame formats (addressing conventions, error-correction and detection capabilities), media access protocols, and some higher-level protocols (such as routing). However, we are not yet even close to determining the underlying issues, such as the physical medium and network topology that would be most appropriate for any of the IVHS functions.

ACKNOWLEDGMENTS

This paper was prepared in cooperation with the state of California, Business Transportation and Housing Agency, Department of Transportation. The advice and recommendations of colleagues Pravin Varaiya, Jean Walrand, and Anthony Hitchcock of the Program on Advanced Technology for the Highway (PATH) are sincerely appreciated.

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

Publication of this paper sponsored by Committee on Communications.

Guidelines for Use of Leading and Lagging Left-Turn Signal Phasing

JOSEPH E. HUMMER, ROBERT E. MONTGOMERY, AND KUMARES C. SINHA

The use of optimum phase sequences at signalized intersections could save motorists many hours of delay and could result in fewer accidents. However, very little factual information has been available to guide engineers in choosing between the various signal phasing alternatives. To close that gap, leading and lagging signal sequences were evaluated in Indiana using a survey of licensed drivers, an examination of traffic conflicts, an analysis of accident records, and a simulation model of traffic flow. The guidelines developed as a result of these activities generally reflect the advantages documented for lagging sequences over leading sequences in a variety of situations. Lagging sequences are recommended for, among other situations, intersections serving heavy pedestrian volumes, diamond interchanges or one-way pairs, and intersections with fixed-time signals. However, when implementing lagging sequences, caution is recommended to prevent situations in which a vehicle could become "trapped" in an intersection as the green phase elapses.

Left turns at intersections have long been a source of concern for traffic engineers. In recent years, greater traffic volumes at many intersections and fiscal and right-of-way constraints on construction have led traffic engineers to design and implement increasingly sophisticated signal schemes to allow vehicles to turn left safely and efficiently. The permissive scheme is the most common type of signal scheme accommodating left turns in the United States. In this scheme, vehicles may turn left when receiving the green-ball signal and when sufficient gaps appear in the opposing traffic stream, which also has a green-ball signal. In another very common signal scheme, the protected scheme, vehicles may turn left only when receiving a green-arrow signal, which affords them exclusive right-of-way through the intersection. In most applications, the protected signal is given to vehicles turning left from a particular street before the green ball is given to the through movement on the same street (i.e., protected-leading). Most other common signal schemes to accommodate left-turning vehicles involve a variation on or combination of permissive and protected schemes, including:

- Protected-lagging, in which the green arrow is given to left-turning vehicles after the through movements have been serviced;
- Protected-permissive, in which protected left turns are made first in the cycle and a green-ball signal allows permissive left turns later in the cycle; and
- Permissive-protected, in which permissive left turns are allowed first in the cycle and protected left turns are accommodated later in the cycle.

J. E. Hummer, Department of Civil Engineering, University of North Carolina, Charlotte, N.C. 28223. R. E. Montgomery and K. C. Sinha, Purdue University, West Lafayette, Ind. 47907.

Protected-leading and protected-permissive are referred to as "leading" schemes, and protected-lagging and permissive-protected are known as "lagging" schemes.

Research has been conducted on a number of questions involving the common left-turn schemes. However, the question of the effects of leading and lagging schemes has received little attention from researchers. Many localities and practitioners, faced with the choice of lead or lag, base their decisions on tradition, hearsay, or feeling, not factual evidence. The intent of the research reported here was to examine the relative merits of leading and lagging phasing schemes and to develop appropriate guidelines that would assist decisions on lead and lag.

Finding an answer to the leading and lagging sequence question would have many potential benefits. If the guidelines save 1 second of delay per vehicle at 200 typical intersections, about 1 million hours per year would be saved. Such a reduction in vehicle delay would also save fuel and decrease pollution. Additional benefits could accrue to operating agencies and to taxpayers if construction projects to add intersection capacity are delayed or scaled down because of changes in signal sequence. Also, although the number of accidents involving left-turning vehicles per intersection is relatively small, the guidelines would potentially result in accident savings.

PURPOSE AND SCOPE

The primary purpose of the research was to produce guidelines for the use of leading and lagging left-turn signal sequences. A secondary purpose of the research was to advance the body of knowledge regarding left-turn signal schemes in general. General information on left-turn signal schemes would be useful in compiling a comprehensive set of guidelines on left-turn phases.

The scope of the research was limited in a number of ways. First, attention was given primarily to only the five common left-turn schemes described. Second, data collection activities were confined to Indiana. Third, with one exception, the research was concentrated on intersection types that are relatively common in Indiana. Intersections with five or more approaches, dual left-turn lanes, offset approaches, or a great deal of channelization are rare in Indiana, so the limited resources of the project were not expended on them. Although they are not common in Indiana, diamond interchanges where both ramp terminals had signals with left-turn arrows were included for study because an increasing number of those interchanges are being signalized.

The major areas of potential concern relative to leading and lagging and other left-turn issues explored in this research

include motorist preferences and understanding, safety, and delay. These areas were addressed during the review of relevant published research findings. Data on motorist preferences and understanding were gathered through a survey at the 1988 Indiana State Fair. Safety was explored using a field study of traffic conflicts and an analysis of accident data at a sample of intersections. A detailed microscopic simulation model of arterial street networks was the primary tool used to study delay. Safety-related variables were also analyzed using a series of simulation runs. The results from all of these different work elements were used to develop guidelines for the use of leading and lagging left-turn signal phasing. A detailed description of the methods, dates, and results of these work elements is provided elsewhere (*1*).

LITERATURE REVIEW

The literature on left-turn phasing, especially the left-turn phase sequence, was reviewed and provided information on delay, safety, and motorist preferences. For delay, no clear trend emerged between leading and lagging schemes at isolated intersections. However, it was clear that a policy that allows the choice of lead or lag at individual approaches in a coordinated system with the aim of maximizing the through bandwidth decreases delay (*2–4*).

Concern for the safety of drivers and passengers in vehicles that become "trapped" in an intersection while waiting to make a left turn has been consistent in the literature (*2,5–7*). Trapping may occur to a vehicle making a left turn on an approach with a permissive signal where the opposite approach has a lagging signal. When the permissive signal goes to yellow and then to red (to provide the lagging green-arrow signal for the left-turning traffic in the opposite direction), the signal for opposing through traffic remains green. A vehicle turning left with the permissive signal will not be able to complete its turn at the end of the cycle as at a normal permissive intersection. At best, the vehicle will be able to back up to the stop bar. If other vehicles in the left-turn queue have moved up behind it, the lead vehicle will not be able to back up to the stop bar and will be trapped in the middle of the intersection. At worst, the driver of the left-turning vehicle will not recognize that the opposing traffic still has a green signal and will try to turn, expecting the opposing traffic to stop as usual. Intersections where one approach has a permissive left turn of some kind and the opposing approach has a lagging sequence must be checked for the possibility of trapping. Trapping can be mitigated by eliminating the permissive turn (making it protected-only or prohibiting the turn), by eliminating the lagging sequence, by ensuring that the opposing approaches both have lagging sequences with left-turn phases that begin simultaneously, or by using other phasing measures. The literature revealed several reasons why lagging sequences might lead to fewer accidents than leading sequences at certain types of intersections where trapping conditions are not present (*5*). Data to evaluate the relative safety of the signal sequences were sparse, however.

The only study reviewed that examined motorist preferences for lead or lag showed a great deal of support for the lagging sequence (*8*). The sparse data available on the question of motorist confusion showed few such problems when drivers face a change in signal sequences or a variety of sequences in close proximity (*8–10*).

The plentiful literature on the tradeoffs between permissive, protected, and either protected-permissive or permissive-protected signals was also reviewed. The literature documented the well-known general trend that accidents increase and delay decreases as the level of left-turn protection decreases. Protected signals were recommended in the literature for intersections with high-speed approaches, restricted sight distances, or three or more opposing through-lanes. Warrants for the installation of some type of left-turn protection instead of permissive signals are available. Directional separation left-turn signals, where each intersection approach has the exclusive right-of-way in turn, are another option available to engineers at certain intersections.

MOTORIST SURVEY

A 4-day survey of Indiana drivers was conducted at the 1988 Indiana State Fair. The survey provided many useful results on the relative understanding of various left-turn signal and sign alternatives. The survey also provided data on the preferences of motorists for various left-turn signal alternatives, including the leading and lagging sequence alternative. Survey data were collected during short interviews conducted by transportation graduate students. Respondents received three fair amusement coupons (worth $0.45 each) for completing the interview.

Over 400 valid responses were received. Despite the fact that the survey was conducted in one place over a 4-day span, responses were received from a wide variety of people. The error rate computed for the nine understanding questions, and the lack of association between preferences expressed and particular interviewers or survey days, showed that the survey script, displays, and format were reasonable and that the data were not biased in any substantive way. However, applications of the survey data outside this project must be made carefully, keeping in mind the context of the survey (i.e., the tendencies of Indiana drivers and highways in 1988).

The leading sequence was preferred by 248 respondents, and the lagging sequence was preferred by 59 respondents; 95 respondents expressed no preference for either signal sequence. The difference between leading and lagging was found to be significant using a confidence interval at the 0.05 level, but the relatively high number of respondents with no preference indicates that the overall preference may not have been as strong as the confidence interval would indicate. Table 1, which summarizes the reasons given by respondents for their preferences, shows that more respondents preferred the leading sequence because it was more like normal (i.e., more common). Many other respondents credited the leading sequence with causing less delay and being safer. Table 2 shows the relationships between the preference for leading or lagging sequence and various independent variables from the survey. The preference for leading and lagging sequence was somewhat related to the age of the respondent, although the main contributor to the high chi-square value in this case was the tendency of younger drivers to have no preference more often. The variable for urban or rural county of residence was found to be related to the choice of leading or lagging sequence,

TABLE 1 REASONS FOR PREFERENCES FOR LEADING AND LAGGING SIGNAL SEQUENCES

Reason given \ Signal sequence preferred	Number of respondents*	
	Leading	Lagging
Safer	61	11
Less delay	65	17
Less confusion	27	11
More like normal	73	10
Unsure or other	39	11

* Some respondents provided more than one reason for their preference.

TABLE 2 RELATIONSHIPS BETWEEN PREFERENCES FOR LEADING OR LAGGING SEQUENCES AND VARIOUS INDEPENDENT VARIABLES

Variable	Chi-square value	Reason for significant or nearly significant relationship
Age	.054	Younger drivers had no preference more often
Sex	.126	--
Urban or rural county of residence	.002	Rural residents preferred lagging more often
Annual miles driven	.056	Those driving less preferred lagging more often
Number of errors on nine understanding questions	.526	--

with people from rural counties expressing a preference more often for the lagging sequence. The variable for annual miles driven was also somewhat related to the preference for leading or lagging signals, with people driving the least opting for the lagging sequence more often.

Several results from the motorist survey that did not pertain to the leading and lagging issue were also notable. The protected signal was far better understood than the permissive signal, which was in turn better understood than the protected-permissive signal. The "left turn yield on green ●" sign proved more confusing than the other protected-permissive sign conditions tested (the no-sign condition and the "left turn on green or arrow" sign). There was little to distinguish the protected sign conditions tested (no sign, "left turn signal" sign, and "left turn on arrow only" sign) on the basis of motorist understanding. Finally, the protected signal was the most preferred signal because most respondents associated it with less confusion, and the permissive signal was the least preferred signal.

TRAFFIC CONFLICTS

The relative safety afforded by leading and lagging signal sequences has not been well documented. To help overcome that gap, a traffic conflict study was conducted at six intersections in Indianapolis. Traffic conflicts are events involving the interaction of two or more road users where one or both users take evasive action such as braking or weaving to avoid a collision. Traffic conflict data have been shown to be correlated with accident data in many traffic situations; because traffic conflict data can be collected in a relatively short period of time, they are often used as a proxy for accident data (*11*).

Three pairs of intersections were identified for the traffic conflict study. Each pair consisted of an intersection with a permissive-protected signal and an intersection with a protected-permissive signal. In most respects except the signal type, the intersections were similar between members of a pair. All six intersections studied were intersections between a two-way street and a one-way street with fixed-time signals in Indianapolis. The intersections included a "downtown" pair with many pedestrians and low vehicle speeds, an "urban" pair with few pedestrians and 30- to 35-mph speed limits, and a "suburban diamond" (i.e., diamond freeway interchange) pair with no pedestrians and 40-mph speed limits. Data were gathered manually on all conflicts and unusual maneuvers that were witnessed by observers on two sides of a test intersection.

Table 3 shows the results of the conflict study for the four types of conflicts and unusual maneuvers that were most related to left-turning vehicles, including

- A left-turning vehicle interacting with an oncoming through vehicle ("left and oncoming");
- A left-turning vehicle interacting with a pedestrian crossing the approach onto which the vehicle is turning ("left and pedestrian");
- A left-turning vehicle hesitating or starting and then stopping suddenly when presented with a green-ball signal and no oncoming traffic or with a green-arrow signal ("indecision left"); and
- A left-turning vehicle crossing the stop bar and entering the intersection on a red-ball signal ("run red left").

Table 3 shows that numbers of conflicts sufficient for analysis were recorded during the periods of observation for almost every conflict type at each intersection. Table 3 also shows that the numbers of left-turning vehicles were very similar between members of the urban and suburban diamond pairs, and were quite different for members of the downtown pair. The conflict rates shown in Table 3 (conflicts per left-turning vehicle) were of reasonable magnitude, ranging from just under 4 percent to just under 0.4 percent.

The largest difference between leading and lagging sequences in Table 3 was for the left and pedestrian conflicts at the downtown pair, where the leading sequence was associated with three times as many conflicts and six times as great a conflict rate as the lagging sequence. In most cases at the leading site, these left and pedestrian conflicts happened when pedestrians stepped off the curb and into the approach to which left-turning vehicles were destined upon seeing a red signal for the cross-street (ignoring the "don't walk" signal). This result agrees with findings from the literature (5) and was considered in developing guidelines for left-turn signals.

Table 3 also shows that the lagging sequence intersection of the suburban diamond pair was associated with a significantly lower rate of run red left conflicts (at the 0.05 level) than the leading sequence intersection. Many times at the leading sequence intersection, three vehicles were observed

TABLE 3 LEFT-TURN CONFLICT RESULTS

Conflict type	Intersection	Number of conflicts	Number of left turns	Proportion of left turns in conflicts	Significant at 0.05?
Left and ped.	Dntn-lag	11	1828	.006	Yes
	Dntn-lead	33	892	.037	
Left and oncoming	Dntn-lag	23	1828	.013	Yes
	Dntn-lead	24	892	.027	
Indecision left	Dntn-lag	30	1828	.016	No
	Dntn-lead	13	892	.015	
Run red left	Dntn-lag	10	1828	.006	No
	Dntn-lead	4	892	.004	
Left and oncoming	Urb-lag	9	1073	.008	Yes
	Urb-lead	22	1022	.022	
Indecision left	Urb-lag	24	1073	.022	No
	Urb-lead	16	1022	.016	
Run red left	Urb-lag	9	1073	.008	No
	Urb-lead	7	1022	.007	
Left and oncoming	Sub-lag	17	1322	.013	No
	Sub-lead	16	1044	.015	
Indecision left	Sub-lag	48	1322	.036	Yes
	Sub-lead	18	1044	.017	
Run red left	Sub-lag	5	1322	.004	Yes
	Sub-lead	15	1044	.014	

making left turns after opposing traffic had begun to stop for the yellow-ball signal (i.e., three "sneakers"), with the third vehicle entering the intersection with the red-ball signal showing. There was generous supply of candidates for this behavior at the leading intersection because many vehicles wanting to make left turns joined the queue during the permissive phase of the cycle and were still in the queue as the permissive phase was ending. By contrast, at the lagging sequence intersection, the available supply of left-turning vehicles was almost always cleared on the green-arrow signal so fewer vehicles were available to run the red signal.

Another important result shown in Table 3 is that the lagging sequence was associated with significantly lower rates of left and oncoming conflicts (at the 0.05 level) than the leading sequence at the downtown and urban pairs of intersections. Two alternate explanations for these differences were available based on the data. First, the number of opposing vehicles recorded at the lagging intersection downtown was 6,947 versus 3,285 at the leading intersection downtown; 6,634 opposing vehicles were recorded at the lagging urban intersection versus 3,590 at the leading urban intersection. Thus, vehicles turning left at the lagging intersections may have had fewer opportunities to turn on the green-ball signal and, therefore, fewer opportunities to be involved in left and oncoming conflicts. This possibility was tested by comparing the conflict rates at the leading and lagging sequence intersections for 15-min periods with similar oncoming volumes. The tests showed that the lower oncoming volumes at the leading intersections may account for some but not all of the difference in conflict rates between leading and lagging signals. For the downtown pair during periods of similar oncoming volumes, the lagging sequence intersection had a significantly lower rate than the leading sequence intersection. For the urban pair during periods of similar oncoming volumes, the lagging intersection had a lower rate, but the difference was not significant.

The second explanation for the lower left and oncoming conflict rates at the lagging intersections in the urban and downtown pairs was the tendency at the leading intersections for left-turning vehicles to try to enter the intersection immediately after the yellow-arrow signal had ceased as if they still had the right-of-way. These "time stealers" then interacted with the more forthright of the oncoming vehicles, which had just received the green-ball signal. Examination of the descriptions of particular conflicts revealed that time stealers accounted for most of the difference in conflict rates between the leading and lagging downtown and urban intersections. There were a number of time stealers at the leading suburban diamond intersection as well, but the lagging intersection of that pair had an abundance of left and oncoming conflicts by indecisive left-turning vehicles and the two effects canceled each other in the final statistics.

Indecision conflicts accounted for the remaining significant difference between leading and lagging intersections shown in Table 3. The lagging intersection was associated with a higher rate of indecision conflicts than the leading intersection at all three intersection pairs, and the difference at the suburban diamond pair was significant at the 0.05 level. Examination of the data revealed that virtually all of the indecision conflicts, whether by a left-turning or other vehicle, occurred at the beginning of a signal phase. The number of signal cycles, rather than the number of vehicles observed, may have been the more appropriate available variable with which to compute a conflict rate. Therefore, the indecision conflict rates per signal cycle were computed; they confirm that it was the lagging sequence that was associated with higher indecision conflict rates, including significantly higher rates for the indecision left conflicts at the downtown and suburban diamond pairs.

Two basic reasons emerged to explain the generally higher rates of indecision conflicts at lagging sequence intersections. First, left-turning vehicles that received a lagging green arrow were hesitant to begin a turn until it was absolutely clear that oncoming traffic was going to stop. This was especially true at the suburban diamond location where the speeds of oncoming vehicles were relatively high. These high speeds sometimes led to false starts by left-turn vehicles, rapid decelerations by vehicles behind the left-turn queue leader, horn honking, and other unusual behavior. Second, drivers of left-turning and other vehicles often seemed surprised by a lagging signal sequence, and sometimes committed false or late starts upon receiving the right-of-way. Considering that there are very few lagging sequences in Indiana, some motorist surprise is understandable.

ACCIDENTS

For this project, accident data were used to help evaluate the relative safety of intersections with leading left-turn sequences and similar intersections with lagging signal sequences. Fourteen intersection approaches with lagging sequences (i.e., all Indiana intersections with lagging sequences for which data were available) were compared to 15 approaches with leading sequences. Almost all of the lagging sequence approaches and all of the leading sequence approaches were at intersections where a two-way street met a one-way street. All intersections studied had fixed-time signals, and most were in downtown areas. Indiana Department of Transportation accident records from 1985 through 1988 were used during the study, with traffic volume data from various sources to obtain accident rates for comparison. Only accidents involving a vehicle turning left from an approach with a left-turn signal of interest were analyzed.

Table 4 summarizes the reported accidents for the leading and lagging intersection sets. Accidents were more frequent and occurred at a greater rate at intersections with leading sequences, though the difference between leading and lagging sequences was not large for left-turn accidents per left-turn vehicle or left-turn accidents per total vehicle (i.e., all vehicles entering the intersection). The difference for the former was not significant at the 0.05 level; the difference for the latter was significant at the 0.05 level using the Z-test for proportions. Extreme caution should be used before basing left-turn sequence policy on such a small difference in accident rates between small samples of relatively homogeneous intersections.

The accident data in Table 4 were analyzed for relationships to several other accident variables. The variation of rates at leading and lagging sequence intersections with left-turn volume, with pavement and light conditions at the time of the accident, and with collision type were all investigated. In all three cases, no significant relationship was found. The severity of accidents in the leading and lagging intersection sets was

TABLE 4 LEAD AND LAG SET ACCIDENT DATA SUMMARY

Statistic	Lagging signals	Leading signals
Number of Indianapolis intersections	9	7
Number of intersections in other cities	5	8
Left turn accidents	44	69
Left turn volume, millions	56	74
Total intersection volume, millions	718	693
Accidents per million left turn vehicles	0.8	0.9
Accidents per million total vehicles	0.06	0.09

also investigated and was found to differ between the sets. Twenty-five accidents at the leading sequence intersections (35 percent) caused one or more reported personal injuries. In contrast, only three of the accidents at the lagging sequence intersections (7 percent) resulted in one or more reported personal injuries. This difference was found to be highly significant at the 0.05 level using a chi-square test.

Another general conclusion that could be drawn from Table 4 is that the number of left-turn accidents reported per intersection per year was relatively low regardless of the signal sequence. Over 4 years, 113 left-turn accidents were recorded at 29 intersection approaches, for a rate of just under one accident per approach per year. Because of a higher sample size and fewer uncontrolled factors, this conclusion has a much higher likelihood of being generally true than the conclusion discussed earlier regarding the difference between leading and lagging sequences. One of the consequences of the relatively low number of reported accidents per approach per year is that a large sample of intersections would be necessary in any future extensive evaluation of leading and lagging sequences or other left-turn alternatives using accidents. In addition, modest changes in the overall traffic safety picture of a region are all that can be expected from even the most widespread left-turn safety treatment programs if the number of accidents reported before the programs begins is low.

SIMULATIONS

The relationship of left-turn signal sequence to delay- and safety-related variables was investigated during this research using a series of experiments with the 1986 version of the NETSIM traffic-flow simulation model. NETSIM was chosen for this research because it is stochastic, microscopic, and supported by FHWA. NETSIM was also desirable because it can model an entire network of streets and intersections.

Five separate experiments were run with NETSIM, including experiments on intersections with four approaches, on intersections with three approaches, and on diamond interchanges. These experiments measured the utilization of the various signal phases by left-turn vehicles and used actual intersection data for inputs. Thirty-minute simulation runs of traffic flow near an intersection with a certain type of left-turn signal and other controlled variables were studied. Many factors were kept constant throughout the experiments to avoid bias. The intent in building models with NETSIM was to provide a fair test of leading and lagging sequences under conditions that were representative of those at intersections in Indiana. The Signal Operations Analysis Package (SOAP) was used to obtain signal-timing parameters throughout the experiments. A left-turn gap-acceptance distribution based on data collected for this project was used in NETSIM throughout the experiments. Comparisons of data collected for this project to NETSIM output, along with the long record of NETSIM in similar research and other recent validation efforts, demonstrated that the model produced reasonable results.

The five experiments were designed and run as factorials. Analysis of variance and Student-Newman-Keuls means tests were used to draw conclusions from the data. The type of left-turn signal was varied in each experiment. The volume of left-turn traffic, the volume of through traffic, and the type of progression on the major street was varied in all experiments except the actual intersection experiment. The desired approach speed and the type of signal equipment (i.e., fixed-time or actuated) were varied in the four-approach experiment, the desired approach speed was varied in the utilization of signal phases experiment, and the type of signal equipment was varied in the diamond interchange experiment. Only fixed-time signals were modeled during the three-approach and the utilization of signal phases experiments. Three different intersections and five different time periods (morning peak, midday, evening peak, overnight, and other hours) were used in the actual intersections experiment. Volume levels used in the experiments were based on peak-hour volume data from random samples of intersections in Indiana with left-turn signals. The volume levels used were generally moderate, causing nearly saturated conditions only when the combination of the highest-volume classes with protected signals was modeled.

Data summarizing the relationships between the delay-related measures of effectiveness and the various left-turn

signal types tested for each experiment are shown in Table 5. The largest experiment involved intersections with four approaches; it showed that protected-permissive signals caused slightly more delay, stopped delay, and stops than permissive-protected signals. No significant difference between protected-lagging and protected-leading signals was detected. The experiment on intersections with three approaches was highlighted by the fact that there was little difference between the protected-permissive and permissive-protected signals in delay or stopped delay, but the latter caused significantly fewer stops per vehicle. A variation on this experiment demonstrated the sensitivity of the lead and lag decision to the time in the signal cycle that the progression band arrived at the left-turn signal. The experiment on diamond interchanges documented the superiority of lagging over leading schemes in terms of delay and stops. The results for the delay-related measures of effectiveness for the utilization of signal phases experiment were very similar to the results for the three-approach experiment. The difference between leading and lagging for mean stops per vehicle was significant at the 0.05 level, but there was no significant difference between leading and lagging for the delay-related measures. Finally, the actual intersection experiment confirmed the relative efficiency of the lagging sequence for a limited range of intersections. During the experiments, all other main effects of factors (desired approach speed, signal type, progression class, left-turn volume, through volume, and left-turn signal type) and all interactions between any two of the factors were also investigated. The results are given in detail elsewhere (*1*).

Table 5 also shows the trend seen throughout the simulation experiments that permissive signals were associated with the least delay and the fewest stops, while protected signals were associated with the highest delay and the most stops. Only for the highest-volume levels during the diamond interchange experiment did the permissive signal produce more delay than a competitor signal and did the protected-lagging signal produce less delay than the protected-permissive signal. For all other combinations of volume levels and other variables tested, the rankings between types of left-turn signals on the basis of delay and stops remained unchanged. It should be noted that the measures of effectiveness in Table 5 were computed for all vehicles on the approaches to the intersection being simulated with left-turn signals, not just left-turn vehicles, and that delay and stop data for left-turn vehicles alone may present a different picture.

Table 6 shows results of the utilization of signal phases experiment. The lagging signal had significantly more left turns completed on

- The green-ball indication,
- The yellow-ball indication,
- Green indications, and
- Ball indications.

The leading signal had significantly more left turns on

- The yellow-arrow indication,
- The red indication,
- The last yellow indication before the red, and

TABLE 5 SUMMARY OF RELATIONSHIP BETWEEN MEASURES OF EFFECTIVENESS AND LEFT-TURN SIGNAL TYPES IN FIVE SIMULATION EXPERIMENTS

Experiment	Left turn signal	Mean delay, sec/veh	Mean stopped delay, sec/veh	Mean stops per vehicle
Four Approaches	Permissive	10.9	5.2	0.35
	Permissive-protected	13.5	7.4	0.43
	Protected-permissive	14.7	8.5	0.46
	Protected-lagging	19.4	12.8	0.54
	Protected-leading	19.9	13.3	0.56
Three Approaches	Permissive	7.2	4.0	0.27
	Permissive-protected	10.4	6.8	0.35
	Protected-permissive	10.4	6.8	0.36
Diamond Interchange	Permissive	11.9	7.0	0.30
	Permissive-protected	13.7	7.7	0.38
	Protected-permissive	17.3	10.5	0.45
	Protected-lagging	18.4	11.8	0.54
	Protected-leading	23.0	15.5	0.62
Utilization of signal phases	Permissive-protected	17.0	10.3	0.48
	Protected-permissive	16.9	10.4	0.49
Actual Intersections	Permissive-protected	12.4	No data	0.44
	Protected-permissive	16.5	No data	0.58

TABLE 6 SUMMARY OF ANOVA RESULTS ON UTILIZATION OF SIGNAL PHASES BY LEFT-TURN VEHICLES

Interval(s)	Mean value of percent of left turns on the interval(s)		Significance probability for signal type
	Permissive-protected	Protected-permissive	
Green ball	33	23	0.0001
Yellow ball	31	28	0.0150
Green arrow	25	20	0.0755
Yellow arrow	8	15	0.0008
Red	3	14	0.0001
Green (ball plus arrow)	58	44	0.0001
Yellow (ball plus arrow)	39	43	0.0945
Ball (green plus yellow)	64	51	0.0001
Arrow (green plus yellow)	32	35	0.1424
Last yellow before red	8	28	0.0001
Last yellow before red plus red	11	42	0.0001

- The last yellow indication before the red plus the red indication.

The magnitude of the difference noted above ranged from 3 percent to 31 percent in the case of the difference for the last yellow plus the red indications. There was no statistical difference between the signal levels for the percent of left turns on the green-arrow indications, yellow indications, or arrow indications.

The trend that emerged from Table 6 was that, for the conditions tested, lagging meant more turns on the green-ball and yellow-ball indications, while leading meant more turns near the end of the signal cycle. This trend helped explain the advantages lagging signals enjoyed in delay-related measures of effectiveness during various simulation experiments. The implications of this trend for safety are less obvious, however. The only well-established relationship between the utilization of various left-turn phases and safety documented in the literature review held that safety increased as the percent of left turns made on arrow indications increased. Because there was no difference in the percent of left turns made on the green-arrow indication or on arrow indications between leading and lagging, however, neither can be said to be safer based on this relationship.

Regarding the safety implications of the trend in the results noted above, there are two possible reasons that left turns which are made during the green-ball or yellow-ball indications at a lagging signal may be safer than turns at the end of a leading signal cycle. First, the leading turns at the end of the cycle could conflict with oncoming traffic and with cross-street traffic jumping into the intersection early, whereas the lagging turns on a ball indication in midcycle could conflict with cross-street drivers only when those drivers were making highly illegal maneuvers. Second, drivers contemplating left turns at the end of the leading cycle could feel more pressure to turn (or subject themselves and other drivers in the queue to lengthy delays) than drivers contemplating turns on a ball indication in the lagging cycle. More pressure to turn could result in an acceptance of greater risks. There are no data to substantiate these two reasons; therefore, a cautious outlook was assumed in incorporating this trend into the guidelines on leading and lagging sequences.

The magnitudes of all the differences summarized were documented and may be useful to engineers making traffic signal decisions. The results from the simulations should be used within the context in which they were produced. The limitations of the NETSIM model should be factored into any decision based on these results. Other important limitations of the experiments were biases against protected-permissive signals in the four-approach intersection experiment (no phase overlap at actuated signals) and in the diamond interchange experiment (no "four-phase" operation).

GUIDELINES

Based on these results, guidelines were developed on the use of leading and lagging phase sequences in Indiana. The guidelines are generally applicable at intersections similar to those which were tested during the research. The major features

that analysts should check before applying the guidelines include

- Three- or four-leg intersections on four-lane arterials;
- Intersection angles of approximately 90 degrees;
- Narrow or nonexistent medians;
- Single left-turn lanes;
- Adequate left-turn lane lengths (spillback is rare);
- Relatively unaggressive driver population (gap-acceptance distribution about 0.5 to 1.0 sec more relaxed than drivers in Washington, D.C., no left-turn "jumpers," a maximum of two left-turn "sneakers");
- Light to medium-heavy (but still unsaturated) volumes;
- Balanced flow between the directions on the street with the left-turn signals; and
- Simple two- or three-phase signal control at diamond interchanges.

If conditions at an intersection where a leading versus lagging decision is pending differ greatly from the above conditions, the guidelines should not be directly applied (although the research methods and results may still be of some use). Details on the limitations of the data collected are provided elsewhere (*1*).

The guidelines for choice of leading or lagging left-turn phase sequence when some form of left-turn phasing is warranted are as follows:

1. In coordinated signal systems, use should be made of any phasing sequence on a particular approach that will maximize the through bandwidth.
2. Lagging instead of leading phase sequences should be used at isolated signals serving heavy pedestrian traffic.
3. Lagging instead of leading phase sequences should be used at isolated diamond interchanges or one-way pairs.
4. Permissive-protected signals should be used instead of protected-permissive signals where there is a history of or a potential for left-turn and oncoming vehicle accidents but where protected-leading or protected-lagging signals are not feasible alternatives.
5. Permissive-protected signals should be used instead of protected-permissive signals at isolated intersections with four approaches if the signals are fixed-time or incapable of overlapping phases.
6. Intersections where one approach has permissive left turns and the opposing approach has a lagging sequence must be checked for the possibility of trapping. If trapping is possible the phasing should be changed to eliminate that possibility by eliminating the permissive turn (making it protected-only or prohibiting the turn), by eliminating the lagging sequence, by ensuring that the opposing approaches both have lagging sequences with left-turn phases that begin simultaneously, or by using some other phasing measure.
7. At intersections where the above guidelines do not fully answer the question of lead or lag, the existing phase sequence should not be changed or, if the signal or left-turn protected phase is new, the phase sequence which is most common at similar sites in the area should be used.

Figure 1 is a flow chart based on the guidelines to aid in making phase sequence decisions at individual intersections.

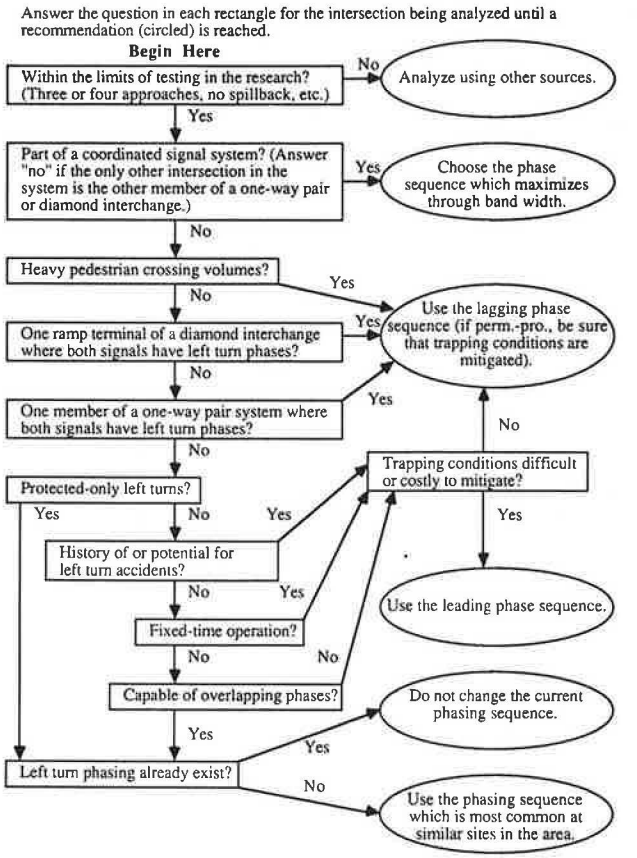


FIGURE 1 Flowchart for decisions on phasing sequence of individual intersections.

Two points must be kept in mind regarding the above guidelines and Figure 1. First, although the guidelines suggest that the signal sequence at a particular intersection in a coordinated system should be chosen to maximize the bandwidth (Point 1), uniformity of signal sequence along an arterial or in a given area may be desirable. When more data are available that show that a variety of signal sequences along an arterial or in a given area does not pose a hazard, policies that encourage more flexible signal sequence decisions may be warranted. Second, the guideline that encourages permissive-protected over protected-permissive signals when left and oncoming accidents have occurred or could occur (Point 4) is based on conflict, simulation, and other data pertaining to the end of the signal cycle (i.e., during and immediately after the yellow-ball indication for the protected-permissive signal). If there is a history of or potential for left and oncoming accidents during other parts of the signal cycle, this guideline does not apply, and other sources should be used to make decisions on the signal sequence in that case.

The guidelines have been developed with caution, and changes in phase sequence are called for only in situations where another phase sequence has been proven clearly superior. This cautious approach is appropriate because of the litigious climate surrounding traffic control decisions and the likelihood that accidents may increase immediately after a change in traffic control, such as from lead to lag. If future testing shows that the immediate negative impacts of changes in signal sequence are small, a more active role in changing intersections with the leading phase sequence to the lagging phase sequence should be assumed.

FUTURE WORK

Several other aspects of the leading and lagging issue deserve attention. Foremost on the agenda of future work should be a before-and-after field test of the guidelines developed during this research using both safety- and delay-related measures of effectiveness. A continuous effort over a period of several years is needed to conduct a proper evaluation.

Another area deserving future effort is the simulation of the use of the various signal phases. This portion of the research yielded interesting results, but the cumbersome data collection method limited the amount of data that could be collected. In addition, the question of whether it is better policy to encourage left turns on the green-ball signal or at the end of the signal cycle should be explored. A comprehensive examination of the utilization of signal phases—including alterations to NETSIM or some other traffic simulation model, a thorough validation of the improved model, an experiment comparing phasing alternatives, and a field or accident data collection effort sufficient to convert the simulation results into an estimate of accident reductions—would be a step forward for the traffic community.

Another useful extension of this study would be a series of similar simulation experiments with more varied volume levels. Modeling volumes typical of saturated conditions, unbalanced flows, or the middle of the night may yield some interesting data which could easily extend the scope of the guidelines for leading and lagging left-turn signal phasing.

ACKNOWLEDGMENTS

The research presented in this paper was performed as part of a project titled "Evaluation of Leading Versus Lagging Left-Turn Signal Phasing and All-Red Clearance Intervals" by the School of Civil Engineering of Purdue University for FHWA and the Indiana Department of Transportation. Carl Tuttle was the project advisor from the Indiana Department of Transportation and Ed Ratulowski was the project advisor from FHWA.

REFERENCES

1. J. E. Hummer. *An Evaluation of Leading Versus Lagging Left Turn Signal Phasing*. Ph.D. thesis. Purdue University, West Lafayette, Ind., Aug. 1989.
2. B. W. McKay. Lead and Lag Left Turn Signals. *Traffic Engineering*, April 1966, pp. 50–57.
3. *Guidelines for Signalized Left Turn Treatments*. Report FHWA-IP-81-4, FHWA, U.S. Department of Transportation, 1981.
4. S. L. Cohen and J. R. Mekemson. Optimization of Left-Turn Phase Sequence on Signalized Arterials. In *Transportation Research Record 1021*, TRB, National Research Council, Washington, D.C., 1985, pp. 53–58.
5. H. E. Hawkins. A Comparison of Leading and Lagging Greens in Traffic Signal Sequence. *Proc., 33rd Annual Meeting of ITE*, Toronto, Ontario, Canada, 1963, pp. 238–242.
6. P. Basha and D. Anderson. Authorize Six Month Trial Period of Lagging Left Turn. Report to Mayor and City Council, City of Scottsdale, Ariz., Feb. 29, 1988.
7. Florida Section, ITE. Left Turn Phase Design in Florida. *ITE Journal*, Sept. 1982, pp. 28–35.
8. P. Basha. Lagging Left Turn Arrow Test Results. Memo to Thomas J. Wilson, Acting City Manager, City of Scottsdale, Ariz., Aug. 25, 1988.
9. C. J. Messer, R. H. Whitson, C. L. Dudek, and E. J. Romano. A Variable Sequence Multiphase Progression Optimization Program. In *Highway Research Record 445*, HRB, National Research Council, Washington, D.C., 1973.
10. Tucson's Lag Left Summary. Traffic Engineering Division, City of Tucson, Ariz., Undated.
11. M. Parker and C. V. Zegeer. *Traffic Conflict Techniques for Safety and Operations, Engineer's Guide*. Report FHWA-IP-88-026. FHWA, U.S. Department of Transportation, 1988.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Intergreen Interval Controversy: Toward a Common Framework

C. S. PAPACOSTAS AND NEAL H. KASAMOTO

The chronological development of the most commonly used intergreen interval formulas is traced and a disparity is disclosed between the interpretation presumed by ITE and that originally proposed by Gazis et al. A realistic example clearly shows that proper application of the speed-location diagram introduced by Gazis et al. can enhance the traffic engineer's judgment and can provide a consistent means of reporting research-related observations. The speed-location diagram of the intergreen interval problem must be adopted as a standard tool by traffic engineering practitioners and researchers.

The intergreen interval, that is, the total time period between conflicting green displays at signalized intersections, is the subject of intense debate among traffic engineers. The intergreen interval is commonly displayed either as a steady yellow interval or a combination of yellow followed by an all-red period. An atypical method was used in Pittsburgh, Pennsylvania, and in Ketchikan, Alaska, at least into the late 1960s (*1*,*2*). In that case, the intergreen interval was displayed as a sequence of a simultaneous green and yellow interval followed by a standard yellow interval.

As far back as 1929, Matson, though viewing the intergreen interval as merely an intersection clearance period, wrote that "there are many ways of indicating this caution or clearance period. . . . An understanding of the effects of the clearance period is essential in determining just what is needed. When a definite statement is made as to what amount of time shall be set aside for clearance periods in each cycle, the choice of how these periods shall be indicated rests with the public and its education" (*3*). Yet, after more than 60 years, no consensus has emerged relative to any of these requirements. In 1989, the ITE Technical Council Committee 4A-16, having conducted a review of the vast literature on the subject, proposed revisions to its recommended practice in which it acknowledged that "[d]ivergent and strongly held positions are common when engineers discuss vehicle change intervals. . . . Even among engineers who agree on the method, there are disagreements relative to application" (*4*).

This paper presents an independent review of the chronological development of intergreen interval design equations, discusses the major differences between them, and shows that there are two disparate interpretations of the common design equation that is based on the equations of motion: the interpretation implicit in the *ITE Handbook* and that proposed by Gazis et al. (*5*). This paper shows that the correct interpretation of the Gazis et al. proposal provides a general framework that can help unify what may first appear to be irreconcilable differences of opinion. When properly used, speed-location diagrams in the form suggested by Gazis et al. can enhance the engineer's judgment of intergreen timing at specific intersections and can also aid researchers in properly reporting and interpreting their empirical data.

Regarding terminology, several terms have been used in the literature to refer to the intergreen interval and its subdivisions. Some of these terms (e.g., "clearance interval") attempt to convey a description of purpose or function, but disagreement about these terms has caused unnecessary communication difficulties. ITE (*4*) currently uses the term "change interval," a sufficiently neutral term but one that has also been used to refer to only the yellow display (*6*). The term "intergreen interval," which refers to the total time between conflicting green phases, was borrowed from Hulscher (*7*). The notation used in this paper is partly the authors'. The use of subscripts to certain variables encountered in intergreen interval formulas is an attempt to emphasize the differences in interpretation given to these variables by different authors. A significant part of the controversy regarding the timing and display of the intergreen interval will be traced to these differences.

EVOLUTION OF DESIGN EQUATION

Matson Model

In 1929, Matson (*3*) proposed a formula for computing the needed "clearance interval" to allow vehicles crossing the stop line at the onset of this interval to clear the width of the intersection (W) before control is transferred to the cross street. He also used the terms "amber period" and "caution period" to describe the subject interval. Matson's primary concern was the proper timing and coordination of fixed-signal systems to accommodate the progression of traffic waves traversing urban streets. The required duration was simply taken to be equal to the intersection width divided by the "speed which is normal to the area traversed," that is,

$$T = W/V_n \tag{1}$$

In his short treatment of the subject, Matson described the elaborate procedures needed to establish the "approximate speeds which will be suitable for a signalized street."

C. S. Papacostas, Department of Civil Engineering, University of Hawaii at Manoa, 2540 Dole St., Honolulu, Hawaii 96822. N. H. Kasamoto, Hawaii Department of Transportation, 79 South Nimitz Highway, Honolulu, Hawaii 96813.

1950 ITE Handbook

The 1950 edition of the *Traffic Engineering Handbook* (*8*) used the terms "clearance interval" and "yellow signal indication" and suggested adding the "minimum driver stopping distance" (S_{min}) to the numerator of Equation 1, yielding

$$T = (W + S_{min})/V_n \quad (2)$$

The rationale for adding the stopping distance was that a vehicle traveling at the "normal intersection approach speed" (V_n) could either stop (if located farther than S_{min} from the stop line at the onset of yellow) or clear the intersection at a constant speed (if located closer than S_{min} from the stop line at the onset of yellow).

The following formula, attributed to Earl Reeder, then director of Traffic and Transportation for the city of Miami, Florida, was also presented:

$$T = 0.8 + 0.04V_n + 0.7W/V_n \quad (3)$$

where V_n is given in mph and T in seconds.

Equation 3 results from substituting in Equation 2 the traditional stopping distance formula based on an equivalent constant deceleration rate, that is,

$$S = tV + V^2/(2d) \quad (4)$$

where t is the perception-reaction time and d is the deceleration rate. Apparently, Reeder used a perception-reaction time of 0.75 sec (rounded up to 0.8) and a deceleration rate of 17 ft/sec^2 along with conversion factors allowing the specification of V_n in mph and W in feet. The basis for these values is found in another section of the handbook, "Stopping Distances Used for Design Purposes." The following statements are also found in the 1950 handbook (*8*, p. 69): "Deceleration considered undesirable but not alarming to passengers is 11 feet per second per second," and "comfortable deceleration is 8.5 to 9 feet per second per second." Thus, the stopping distance implicit in the Reeder formula represents emergency rather than comfortable conditions. Moreover, it is not concerned with the deceleration rate that would be attainable on *wet* roadway surfaces, which is the condition governing design in various aspects of highway and traffic engineering. Parenthetically, the traditional design expression of deceleration in terms of kinetics is given by:

$$d = g(f \pm G) \quad (5)$$

where

g = acceleration due to gravity,
f = equivalent coefficient of friction representative of the overall speed change, and
G = roadway gradient.

A friction coefficient of about 0.3 (with some variation related to initial speed) is generally suggested as an appropriate value in calculating safe stopping distances on wet pavements. For a level or nearly level roadway, this value of f leads to an equivalent constant deceleration rate of about 10 ft/sec^2, which happens to equal the widely reported value of comfortable deceleration on dry pavements. The purely kinetic Equation 5 should be preferable to the mixed kinematic-kinetic formula suggested by Parsonson and Santiago (*9*) and adopted by ITE (*4*) because it makes explicit the effect of friction on safe operation. The choice of a high design value for deceleration by Reeder and others was apparently motivated by a desire to keep the duration of the change interval low in order to satisfy those practitioners who "frown on the idea of using yellow intervals in excess of 3 to 5 seconds" (*8*, p. 226).

Substitution of Equation 4 into Equation 2 yields the following general formula:

$$T = t + V_n/(2d_e) + W/V_n \quad (6)$$

where d_e is emergency deceleration.

The 1950 handbook also raised the possibility of deducting from the calculated change interval duration the time required by the leading stopped cross-street vehicle to accelerate from its stop-line position to the point of conflict with the clearing stream. The rationale for this deduction was also discussed much later (in 1977) by Williams (*10*), who, nevertheless, warned that the "time deduction for cross-flow acceleration needs to be applied with caution, and a value of zero should be used if light jumping is possible." In 1981, Parsonson and Santiago (*9*) also pointed out that "the concept pertains to stopped traffic starting up on the green, and not the vehicles approaching the intersection at speed when their signal goes green."

Matson et al.

A restatement of Equation 6 appeared in the Matson et al. (*11*) discussion of the needed yellow light period, except that they prescribed using the comfortable rather than emergency stopping distance. In an obscure theoretical derivation, they computed and compared the required time to stop (y_1) and the required time to clear (y_2) at constant vehicle speed a distance equal the stopping distance plus the intersection width. Cast in different notation than theirs, these two times are:

$$y_1 = t + V/d^* \quad (7)$$

$$y_2 = t + V/(2d^*) + W/V \quad (8)$$

where d^* is comfortable deceleration rate.

From the general comparison of y_1 and y_2, Matson et al. concluded that "time to stop becomes the critical value in determination of yellow light at higher speeds, though time to clear may be the critical value at lower speeds." However, they gave no explanation as to why they felt that the time to stop, as interpreted above, should be used as a criterion for setting the intergreen interval period. As an ITE committee pointed out much later, "once a driver decides to stop, the displayed signal indication becomes meaningless" (*12*). Nevertheless, the concept entered the consciousness of many traffic engineers and, without a doubt, has slanted their understanding and interpretation of the problem.

Gazis et al.

Gazis et al. (*5*) took a fresh look at what they called "the problem of the amber signal light" and formulated an analytical model to describe the predicament of a driver approaching a signalized intersection at the onset of yellow. They essentially expressed the uniform-acceleration equations describing the two possible maneuvers available to the driver, that is, either decelerating to a stop or attempting to clear the intersection, accelerating if necessary. Equation 9 gives the minimum stopping distance from which a vehicle traveling at an initial speed (V_o) can come to a comfortable stop, that is, at a *comfortable* deceleration rate (d^*):

$$X_s = t_s V_o + V_o^2/(2d^*) \tag{9}$$

where

X_s = minimum stopping distance,
t_s = perception-reaction time associated with the decision to stop,
V_o = vehicle speed at the onset of yellow, and
d^* = comfortable deceleration.

The parabola of Equation 9 is independent of the duration of the intergreen interval.

The maximum distance (X_a) from which a vehicle traveling at an initial speed (V_o) can *just* clear the intersection of width W during the intergreen interval of duration T, accelerating if necessary, is given by

$$X_a = V_o T + (1/2)a(T - t_a)^2 - (W + L) \tag{10}$$

where

X_a = maximum clearing distance,
a = equivalent constant acceleration rate,
T = duration of change interval,
t_a = perception-reaction time associated with the decision to clear the intersection,
W = intersection width, and
L = vehicle length.

Gazis et al. reasoned that if, for whatever reason, a vehicle cannot accelerate beyond its approach speed, Equation 10 becomes

$$X_0 = V_o T - (W + L) \tag{11}$$

This linear equation has an X_0-intercept of minus $(W + L)$ and a slope equal to T.

Reasoning that a vehicle approaching at the speed limit (V_1) should not be expected to accelerate in order to clear the intersection, Gazis et al. proposed that the speed limit be used for design purposes. Under this assumption, they proposed that the minimum duration of the change interval be given by the solution of Equations 9 and 11 after setting $X_S = X_0$, that is,

$$T_{min} = t_S + V_{des}/(2d^*) + (W + L)/V_{des} \tag{12}$$

where $V_{des} = V_1$ = the "design" speed used to calculate T_{min}.

For purposes of discussion, Figure 1 plots X_S and X_0 as functions of individual-vehicle approach speed V_o (i.e., Equations 9 and 11). When the two lines intersect, as in the case shown, the speed-position space is divided into five regions as follows:

A: A vehicle cannot clear at constant speed but can stop comfortably,
B: A vehicle cannot stop comfortably but can clear at constant speed,
C: A vehicle has the option to execute either maneuver,
D: A fast-moving vehicle can execute neither maneuver, and
E: A slow-moving vehicle can execute neither maneuver.

Gazis et al. introduced the term "dilemma zone" to describe the conditions encompassed by Regions D and E. A dilemma zone is said to exist if, for a given approach speed, $X_S > X_0$, as illustrated in connection with V_3 and V_4 in Figure 1. The length of the dilemma zone is given by the corresponding differences $(X_S - X_0)$ as shown. A vehicle traveling at a speed associated with a dilemma zone, however, will experience the problem only if it happens to be located within the dilemma zone at the onset of the intergreen interval. An approaching vehicle at the same speed, V_3 or V_4, would either be able to stop or be able to clear the intersection without accelerating if located in Regions A or B, respectively. Figure 2 shows a situation where the X_S and X_0 lines are tangent to each other, that is, where the dilemma zone region is eliminated for only one value of speed. Figure 3 shows a situation where the X_S

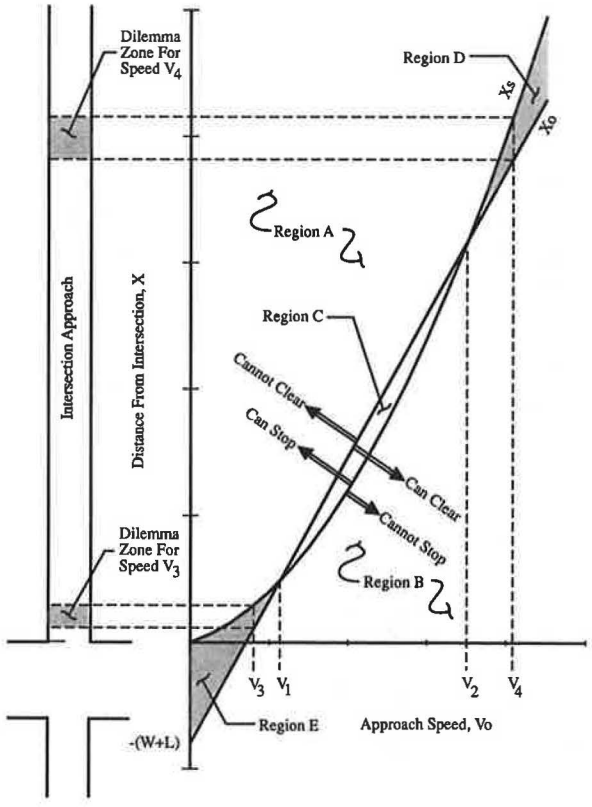


FIGURE 1 Case of intersecting X_0 and X_S plots.

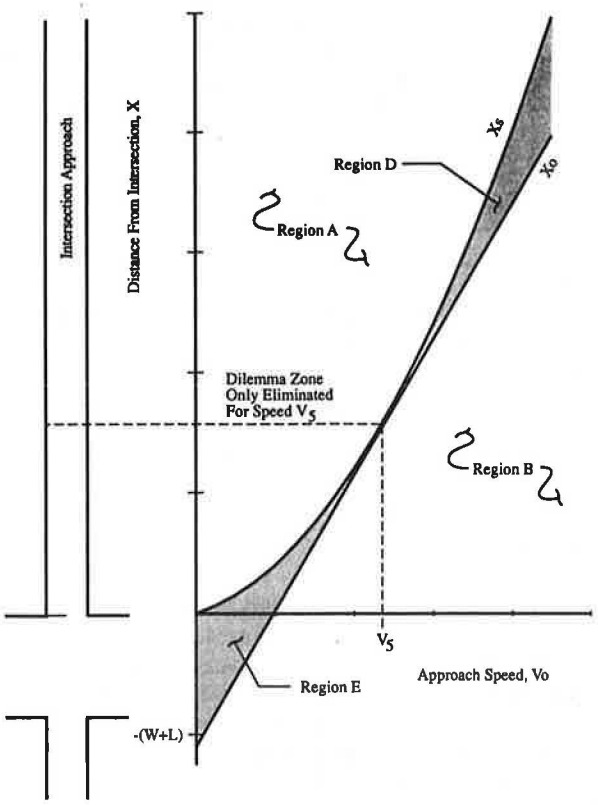


FIGURE 2 Case of tangent X_0 and X_S plots.

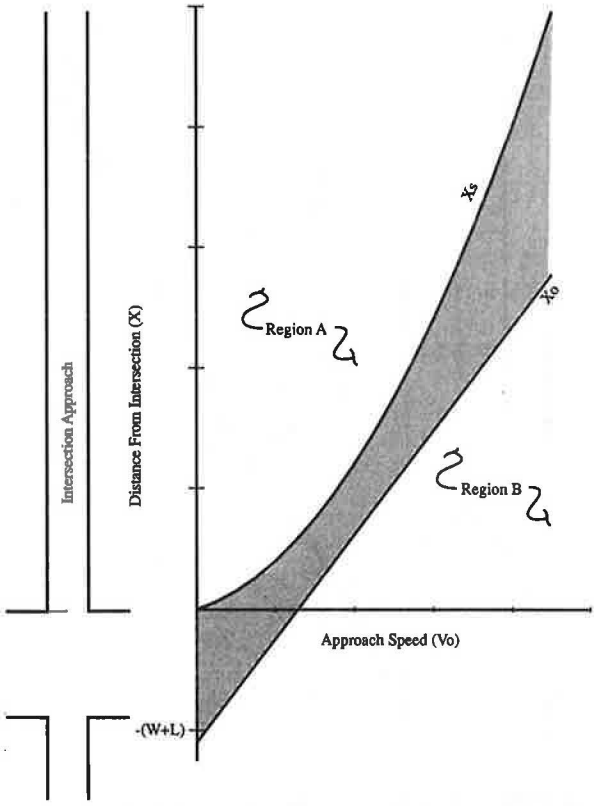


FIGURE 3 Case of nonintersecting X_0 and X_S plots.

and X_0 lines do not intersect and a dilemma region is associated with all speeds. This situation arises when T is set at a value smaller than the T_{min} value obtained by Equation 12. Again, even though dilemma zones may be encountered at all approaching speeds, not all approaching vehicles would in fact be located within the dilemma region at the onset of the intergreen interval.

It was clearly not the intent of Gazis et al. to suggest that the T_{min} value calculated by Equation 12 would eliminate the dilemma zone problem for all approaching speeds. In fact, they discussed the implications that the intergreen interval duration determined by Equation 12 would have on vehicles approaching at speeds other than V_{des}. They also examined the case of accelerating vehicles according to a nonconstant acceleration-speed relationship of the form $a = A - BV$. The presence of dilemma and option regions with and without acceleration are shown in Figure 4 for a case similar to that described by Figure 3. In general, if acceleration is possible, option zones tend to appear over a longer range of individual-vehicle approach speeds.

The presence of option zones (i.e., Region C) did not appear problematic to Gazis et al. However, in 1981, Bissell and Warren (*13*) pointed out that excessively long option zones may contribute to rear-end collisions when the driver of a leading vehicle chooses to stop and the driver of a following vehicle in the option zone decides to go.

Finally, it is important to note on the typical speed-location diagram the presence of a triangular area below the speed axis corresponding to low speeds. This area describes the situation where a vehicle already in the intersection area at the onset of the intergreen interval would not be able to clear the intersection at constant speed. A dilemma zone region also exists to the right of this triangle for slow-moving vehicles

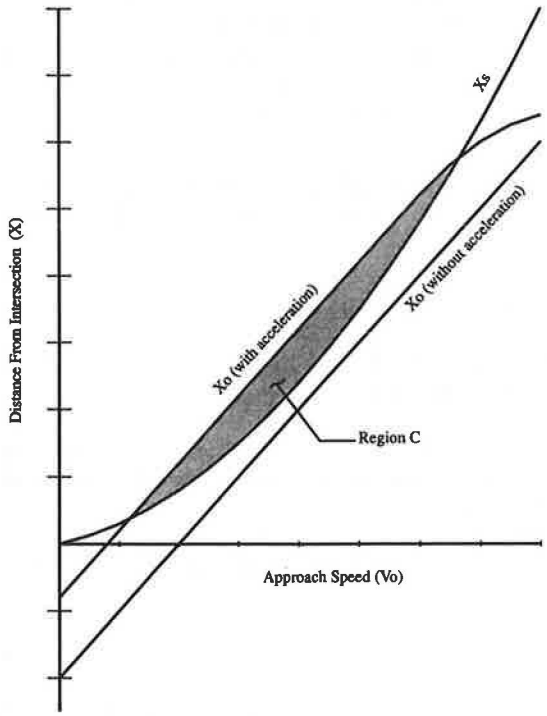


FIGURE 4 Effect of acceleration, when possible.

located very close to the stopping line at the onset of the intergreen interval. This region can be reduced but cannot be completely eliminated, no matter how long the duration of the intergreen interval is. The likelihood that approaching vehicles would actually experience the described conditions would be higher during periods of congested flow when speeds are low and acceleration is restricted by vehicles ahead. Unfortunately Gazis et al. failed to adequately emphasize this situation.

1965 ITE Handbook

The 1965 edition of the Traffic Engineering Handbook (*14*) presented two equations for the determination of the yellow interval as follows:

$$y_1 = t + V_a/(2d) \quad (13)$$

$$y_2 = t + V_a/(2d) + (W + L)/V_a \quad (14)$$

where V_a is specified generically as the "approach" speed. The handbook stated that Equation 13 yields the "minimum time to stop," whereas Equation 14 (attributed to Gazis et al.) yields "the minimum time to stop or clear the intersection." The 1965 handbook suggested that (*14*)

> The yellow clearance interval used should exceed the values of y_1 for the approach speed selected. . . . Where y_2 exceeds the value selected for the yellow interval and where hazardous conflict is likely, an all-red clearance interval is frequently used between the yellow interval and the green interval for opposing traffic.

The use of these two equations was apparently motivated by the work of Matson et al. (*11*), as reflected in the pair of Equations 7 and 8. Inexplicably, however, Equation 13 is misspecified because it does not yield the time to stop. If needed, the proper equation for this time to stop would be Equation 7. Nonetheless, this error has persisted over the years. As late as 1986, Wortman and Fox (*15*) interpreted Equation 13 (and, therefore, the first two terms on Equation 14) as representing "the time required for the driver to come to a safe stop." Lin's work (*16*) was also predicated on the same misconception. His explanation of Equation 14 is as follows:

> The sum of the first two terms in this equation represents the time required for a driver to come to a stop after the yellow interval begins. The last term of the equation is the time required to cross the intersection.

This unfortunate error has undoubtedly caused considerable confusion regarding the interpretation of Equation 14. It appears that some of the critics of the kinematic model have unknowingly leveled their criticism on their misinterpretation of the model rather than on the model itself.

1976 ITE Handbook

The 1976 edition of the *Transportation and Traffic Engineering Handbook* (*17*) retained Equations 13 and 14 and attempted to discuss the dilemma zone concept but attributed it to Olson and Rothery (*18*) rather than Gazis et al. (*5*). In this connection it stated that:

> An incorrect choice for the length of yellow period . . . can lead to the creation of a *dilemma zone*. This is an area close to an intersection in which a vehicle can neither stop safely nor can clear the intersection before the beginning of the red interval without speeding.

Equation 14 is referred to as yielding the "non-dilemma yellow period." The terseness of the dilemma zone description in the handbook and the fact that the citation of Gazis et al. (who had fully presented the concept) was dropped may have caused additional confusion to some users of the handbook. Specifically, some users, unfamiliar with the Gazis formulation, could have been left with the erroneous impression that the choice of a yellow interval of y_2 is meant to eliminate a dilemma zone for all approaching vehicles. As Figures 1 to 3 show, this is not the proper interpretation of Equation 14.

Additional difficulties may have been introduced by the way in which the 1976 handbook describes, as quoted above, the dilemma zone and its relation to the length of the intergreen interval. As Figure 1 shows, dilemma zones can exist close to the intersection for vehicles approaching at slow speeds (for example, V_3, Figure 1), and farther away from the intersection for fast-moving vehicles (for example V_4, Figure 1), even in instances where the length of the intergreen interval is set at y_2 according to Equation 14. Moreover, Figure 2 shows a situation where dilemma zones exist for some vehicles approaching at all speeds other than that used to calculate y_2. As mentioned before and contrary to the above quote, the possibility that slow-moving vehicles could encounter dilemma zones very close to the stop line cannot be eliminated altogether.

Another source of confusion is the 1976 handbook's implication that individual drivers located in a dilemma zone close to the intersection can clear the intersection by *speeding*. It is true, and Gazis et al. addressed this question, that under certain (but not all) conditions, drivers in a dilemma zone at the onset of the intergreen interval can clear the intersection by *accelerating* (see Figure 4). Exceeding the speed limit (i.e., speeding) is not a prerequisite to clearing the intersection under all such circumstances. The likelihood of drivers having to exceed the speed limit in order to clear the intersection before the beginning of the red interval is higher when the traffic is light and vehicles are approaching at relatively high speeds.

1982 ITE Handbook

Under the heading "yellow change and clearance intervals," the 1982 edition of the *Transportation and Traffic Engineering Handbook* (*19*) reintroduced a reference to Gazis et al. and explained their rationale relating to the dilemma zone problem. In an apparent attempt to rectify the misspecification of Equation 13 described above, the handbook begins with the stopping distance equation, that is,

$$S = tV + V^2/(2d) \quad (15)$$

It divides both sides by V so that the left side becomes identical

to that of Equation 13. The handbook calls the resulting term S/V the "minimum clearance time" needed for a "driver to proceed *into* the intersection." Thus, what in earlier editions was considered to be the "time to stop" is now given a different interpretation, that is, the minimum "clearance time" required by a vehicle traveling at a constant speed (V) to cover the stopping distance (S) that would be traversed if the vehicle were to come to a stop from its initial speed (V). In the context of a design formula where S is set to a particular value of S_{min}, this so-called "clearance time" applies only to a nonaccelerating vehicle traveling at the speed used for design purposes (V_{des}). Whether such a vehicle would in fact reach the stop line after S_{min}/V_{des} seconds depends on its initial location at the onset of the intergreen interval.

The value obtained via Equation 14 is described in the handbook as the minimum clearance time that would permit a driver to proceed *through* the intersection. As with the preceding case, the handbook does not make clear the fact that this condition would be satisfied only by vehicles approaching at the speed used in Equation 14 to calculate the intergreen interval requirement that happen to be located at or closer than the comfortable stopping distance implied by that speed.

True Gazis Contribution

Despite references in the ITE handbooks to the work by Gazis et al., the ITE conception of the dilemma zone problem has remained faithful to Matson's original idea of the clearance interval, that is, the time required by vehicles traveling at a single "control" speed to traverse the comfortable stopping distance (S_{min}) corresponding to that speed plus the width of the cross street. According to this restricted conception, vehicles that happen to be located closer than S_{min} at the onset of the intergreen interval would be able to clear the intersection before the cross street receives a green signal, whereas vehicles located farther away would be able to stop comfortably. In other words, ITE's failure to explicitly assess the implications of setting the duration of the intergreen interval, according to Equation 14, on vehicles that approach the intersection at speeds other than that used for design has led ITE to strongly imply that the duration given by Equation 14 can totally eliminate the dilemma zone problem. In terms of their mathematical form, the basic ITE formula (Equation 14) and the Gazis formula (Equation 12) are identical. However, the greatest contribution of Gazis et al. to the understanding of the problem is not that they came up with a formula that had been around for a long time. Rather, their main and, unfortunately, least appreciated contribution lies in the fact that they presented the larger framework, the speed-location diagram shown in Figures 1 through 4, which must *always* be used in judging the appropriateness of the calculated intergreen interval. Thus, given the speed-location diagram of Figure 2, only an imprudent engineer would accept the intergreen duration shown for a design speed that happens to be equal to V_5 merely because it satisfies Equations 12 and 14. Without reference to the speed-location diagram, the presence of a dilemma zone for all speeds other than V_5 would not be readily evident. Similarly, the value of the intergreen interval corresponding to design speeds V_1 and V_2 in Figure 1 also satisfies Equations 12 and 14. However, the dilemma and option implications are distinctly different in the two cases. In other words, satisfaction of Equations 12 or 14 is a necessary but not sufficient reason to accept the calculated value of the intergreen interval requirement at a given intersection. The "solution" to Equation 12 or 14 can at best be viewed as an initial estimate of the intergreen interval requirement, subject to adjustments based on the resulting speed-location diagram implied by this initial estimate.

IMPORTANCE OF SPEED-LOCATION FRAMEWORK: AN ILLUSTRATION

The fundamental importance of the speed-location framework to guide the interpretation of experimental results is illustrated by using the data reported by Stimpson et al. (*6*). Their research attempted to determine whether changing the time duration of yellow signals (referred to as change interval) should affect the frequency of potential conflicts. They defined a potential conflict to exist whenever "the last-to-cross vehicle spent at least 0.2 seconds in the intersection past red onset." To accomplish their objective, they selected two suburban intersections, one in Bethesda, Maryland, the other in Atlanta, Georgia. They carefully selected the experimental sites to ensure certain conditions, including average approach speed near 30 mph, "short" yellow durations (less than 5 sec), reasonably isolated intersections with pretimed signals, four-legged intersections with negligible grade, and good pavement surfaces. The existing yellow durations were 4.7 and 4.3 sec for the Maryland and the Georgia intersections, respectively. For each intersection, they compared the percentage of potential intersection conflicts when the existing yellow was present against the percentage observed when the yellow signal was extended. The percentage of potential conflicts was defined as the ratio of last-to-cross "decision vehicles" (see below) that spent at least 0.2 sec in the intersection past red onset to the total number of decision vehicles that were last to cross.

The yellow at the Maryland location was extended from 4.7 to 6.0 sec, which was the maximum duration acceptable to the responsible traffic engineer. In an attempt to ensure comparability of the results obtained at the two sites, they extended the yellow at the Georgia intersection to "produce a percentage increase of similar magnitude" (i.e., from about 4.3 to about 5.6 sec, with minor variations). At each intersection, before-and-after data were collected separately for peak and off-peak conditions on dry pavements. Observations relating to wet pavement conditions were also collected at the Maryland site. The data were collected via lapse photography and corresponded to vehicles that occupied a "catch zone" 2 sec prior to the onset of yellow. The catch zone was selected so that it "included the dilemma zone at most approach speeds" on the following basis (*6*):

> The upstream extremity of the catch zone was chosen as the point from which a car with an initial speed of 10 mph in excess of the local average speed could come to a full stop at the traffic signal using an average deceleration of 0.25 g (8 ft/sec^2). The downstream extremity was chosen at a point from which a vehicle traveling 10 mph below the average at yellow onset could just clear the cross street prior to red onset. At the

> Maryland site the catch zone extended from 65 feet to 320 feet and at the Georgia site from 25 feet to 320 feet.

The recording and data reduction procedures were summarized as follows (*6*):

> Filming commenced at least two seconds prior to yellow signal onset and continued until all vehicles initially in the catch zone either stopped or cleared the intersection. . . . For the purpose of this study, a vehicle was called a decision vehicle if, in a particular approach lane, it was 1) the first vehicle to stop, or 2) the last vehicle to cross the intersection. Data collection continued until about 150 decision vehicles were obtained at each site under each experimental condition.

A reduction in the percentage of potential conflicts was observed when the yellow signal was extended at each of the two locations and for all experimental conditions investigated. The results corresponding to peak and off-peak conditions on dry pavements are shown in Table 1. Of special interest here is the comparison of the results between the two sites (*6*):

> The results . . . show that potential conflict percentages differed between the two sites both with the initial and extended yellow durations. These differences undoubtedly reflect differences between the two sites in terms of geometry, approach speed and traffic volume . . . but there is not at present quantitative relationships that would predict potential conflict frequency in terms of these, and possible other, factors.

We agree with Stimpson et al. as to the possible factors that gave rise to the observed differences between the two sites. We contend, however, that the speed-location diagrams reflecting the two yellow durations at the two intersections studied could contribute to an explanation of their findings. Such speed-location diagrams for the Maryland and Georgia sites are shown in Figures 5 and 6. The X_0 curves shown were based on a perception-reaction time of 1 sec, a deceleration rate of 10 ft/sec^2, and vehicle length of 20 ft. Also shown on each figure are the "catch zones" within which experimental data were collected as described. It was not possible to show similar ranges of observed speeds because only the average approach speeds are given by Stimpson et al.; these averages are included in the two figures.

Even a cursory inspection of the two figures is sufficient to pinpoint the prevailing differences at the two intersections and to provide a reasonable explanation of the experimental findings: Given the initial yellow durations at the two inter-

TABLE 1 SUMMARY OF RESULTS (*6*)

a) Maryland Intersection

	Before (T = 4.7 sec.)	After (T = 6 sec.)
Peak	19%	2%
Offpeak	15%	1%

b) Georgia Intersection

	Before (T = 4.4 sec.)	After (T = 5.6 sec.)	After (T = 5.8 sec.)
Peak	90%	21%	(*)
Offpeak	63%	(*)	19%

(*) Not Applicable

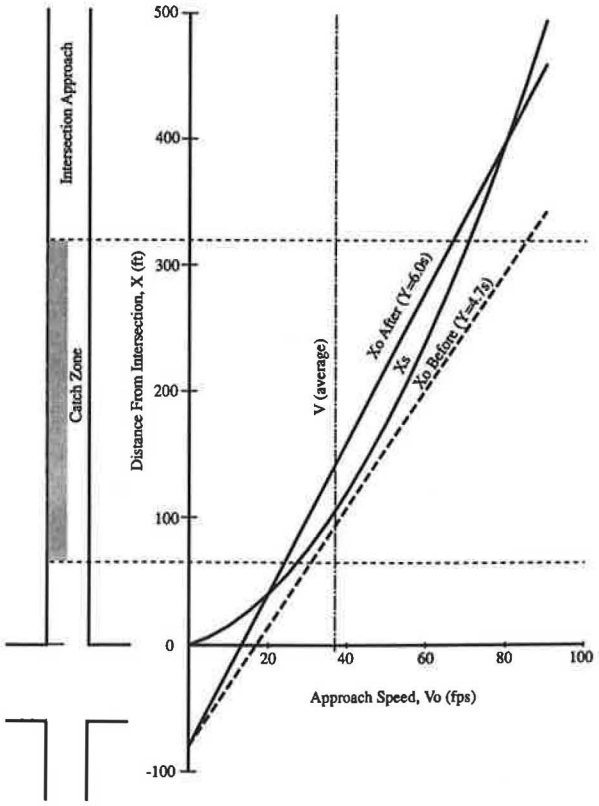


FIGURE 5 Speed-location diagram of Maryland site (*6*).

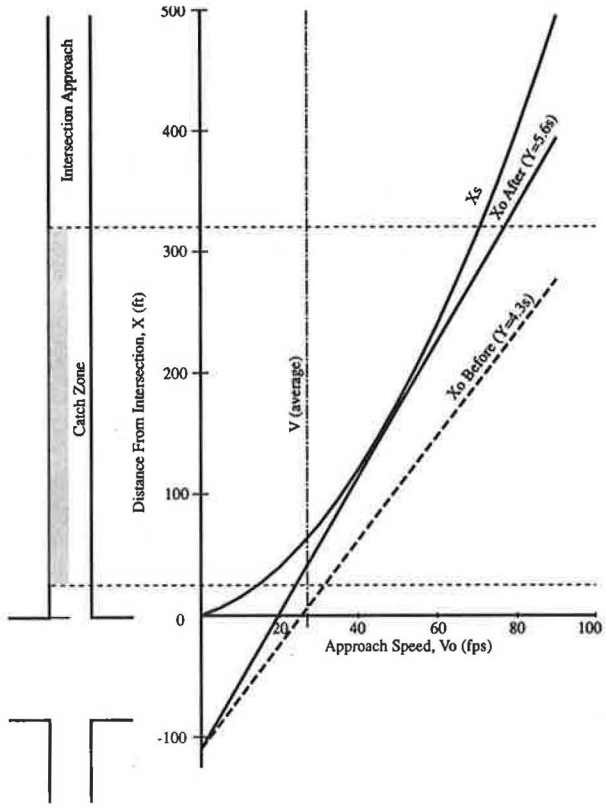


FIGURE 6 Speed-location diagram of Georgia site (*6*).

sections, it is eminently obvious that dilemma zones of significant lengths existed at the Georgia site along the entire length of the catch zone. This intersection's 90-ft width (as reflected by the X_0-intercept) plays a significant role in the presence of dilemma zones. By contrast, at the Maryland site, dilemma zones of any significant length (about 20 ft or more) appear only on the high-speed end of the graph. Thus, the two simple diagrams are sufficient in this case to present the traffic engineer with a clear warning that the Georgia site was likely to experience a higher percentage of potential conflicts than the Maryland site.

The two diagrams are also consistent with the experimental results obtained after the extension of the yellow durations. In the Maryland case, no dilemma region remained over the length of the catch zone after resetting the yellow light; this is consistent with the finding that the percentage of potential conflicts dropped to essentially zero. By contrast, the diagram corresponding to the Georgia site after the signal extension shows that dilemma regions were present at both ends of the catch zone. This finding is consistent with the observed persistence of vehicles in potential conflict at the Georgia site and not at the Maryland site.

Despite the meticulousness with which the researchers attempted to establish their experimental design, they inadvertently failed to account for significant differences at the two intersections that could have been readily disclosed through the use of speed-location diagrams. In other words, in this case, ensuring an equivalent extension of the two signals was clearly not sufficient to render the two intersections comparable. We have discovered similar inadvertent flaws in several experimental designs reported in the literature.

ADDITIONAL COMMENTS

To some degree, the percentages of potential conflicts reported by Stimpson et al. were dependent on the specific choice of catch zones to which their observations were confined. The Maryland site catch zone practically excluded the dilemma region adjacent to the stop line which arises at low approach speeds; the Georgia site catch zone included part of that region. Without a doubt, the observed percentages would have been different had the two catch zones been extended to the stop-line location, including among the decision vehicles those facing the near-side dilemma zones. The observed percentages, of course, would also depend on the actual distribution of observed decision vehicles on the speed-location plane, information that is currently unavailable.

Experimental data can be presented in a convenient and meaningful way by showing, on speed-location diagrams such as those in Figures 5 and 6, the initial speeds and positions of the observed vehicles. Distinct symbols may be used to associate to each observed vehicle its subsequent action (e.g., whether it stopped, whether it cleared the intersection during yellow, or whether it cleared the intersection on red). At a minimum, such practice is capable of transmitting needed information about the intersection under study and its signal characteristics, the sample of vehicles used, and the specific actions taken by the observed drivers. The X_0 and X_S curves can provide guidance in assessing the actions of observed drivers, for example, whether a driver that chose to clear the intersection could have stopped comfortably. Most authors report the duration of the intergreen intervals prevailing during their experimental sessions but typically fail to provide sufficient information to even discern whether dilemma and/or option regions were present at the sites investigated. Moreover, information relating to the distribution of vehicles in the experimental samples used between the various regions of the speed-location diagram is typically absent. Sample selection procedures so differ between researchers that their conclusions are often impossible to compare. For example, Olson and Rothery (*18*) recorded free-flowing vehicles over catch zones not extending to the stop line and consciously disregarded vehicles traveling at considerably lower speeds. As described above, Stimpson et al. (*6*) considered all vehicles occupying similar catch zones at the onset of yellow. Williams (*10*) reports that 816 close-decision vehicles were recorded at a single intersection but gives no further description of his methods; Chang et al. (*20*) sampled vehicles approaching faster than 20 mph. Moreover, with the possible exception of May (*1*,*2*), who has reported his observed sample on speed-location diagrams tailored to the specific characteristics of the intersections studied, researchers tend to merely report some statistical descriptor of their sample speeds along with the signalization existing during their experimental sections. Typically, sufficient information to plot the corresponding speed-location diagrams is unavailable in research reports. Sampling inconsistencies are probably a main source of the conflicting conclusions reported by various researchers, particularly those who attempt to generalize data applicable to restricted scopes and those who attempt to discover behaviorally sound models through blind regression analyses using data from incompatible sources. Without the kind of site-specific information that can be depicted by speed-location diagrams, the conclusions drawn by researchers must remain suspect, particularly when, as in the case of the preceding illustration, the researchers attribute nonconforming findings to unknown factors.

FURTHER REFINEMENTS

Further refinements to the basic speed-location diagram described here are possible. For example, an extended diagram can aid in the depiction of the implications of legal requirements, particularly those that are implied by the permissive yellow rules. It can also clarify the role and implications of competing proposals regarding the division of the intergreen interval into yellow and all-red components, including the current ITE-proposed recommended practice (*4*,*12*). However, length limitations preclude a full discussion of these important questions here.

CONCLUSIONS

This paper traces the chronological development of the most commonly used intergreen interval formula and identifies the major differences of interpretation that are prevalent among traffic engineers and researchers. By far, the most critical difference lies in the disparity between the interpretation presumed by ITE and that originally proposed by Gazis et al. Although both interpretations are based on the same equa-

tions of motion, the former strongly implies that, given appropriate design parameters, the mere solution of the design equation is sufficient to yield an intergreen interval duration which eliminates the dilemma zone problem. This paper clearly shows that this is not the case and that, properly used, the speed-location framework can be an invaluable tool that is capable of enhancing the traffic engineer's judgment and evaluation of initial estimates of intergreen timings at specific intersections. It can also provide a consistent means of reporting research-related data in a manner that can aid the interpretation, understanding, and comparison of formerly incompatible research findings. It is therefore strongly recommended that the speed-location diagram be adopted as a standard tool by traffic engineering practitioners and researchers. Given existing computer technology and graphics software, incorporating the speed-location diagram in practice would be relatively easy.

ACKNOWLEDGMENT

The authors wish to thank Kay Kasamoto for her editorial help.

REFERENCES

1. A. D. May. *Study of Clearance Interval at Traffic Signals*. Institute of Transportation and Traffic Engineering, University of California, Berkeley, 1967.
2. A. D. May. Clearance Intervals at Traffic Signals. In *Highway Research Record 221*, HRB, National Research Council, Washington, D.C., 1968, pp. 47–71.
3. T. M. Matson. The Principles of Traffic Signal Timing. *Trans., 18th Annual Safety Congress*, National Safety Council, Vol. 3, 1929, pp. 109–139.
4. ITE Technical Committee 4A-16. Proposed Recommended Practice—Determining Vehicle Signal Change Intervals: Part I. *ITE Journal*, July 1989, pp. 29–32.
5. D. Gazis, R. Herman, and A. Maradudin. The Problem of the Amber Signal in Traffic Flow. *Operations Research*, Vol. 8, No. 1, Jan.–Feb. 1960.
6. W. A. Stimpson, P. A. Zador, and P. J. Tarnoff. The Influence of Time Duration of Yellow Traffic Signals on Driver Response. *ITE Journal*, Nov. 1980, pp. 22–29.
7. F. R. Hulscher. The Problem of Stopping Drivers After the Termination of the Green Signal at Traffic Lights. *Traffic Engineering and Control*, Vol. 25, No. 3, March 1984, pp. 110–116.
8. H. K. Evans (ed.). *Traffic Engineering Handbook*. ITE, New Haven, Conn., 1950.
9. P. S. Parsonson and A. Santiago. Traffic Signal Change Interval Must Be Improved. *Public Works*, Sept. 1981, pp. 110–113.
10. W. L. Williams. Driver Behavior During the Yellow Interval. In *Transportation Research Record 644*, TRB, National Research Council, Washington, D.C., 1977, pp. 75–78.
11. T. M. Matson, W. S. Smith, and F. W. Hurd. *Traffic Engineering*. McGraw-Hill Book Company, Inc., New York, Toronto, London, 1955.
12. ITE Technical Committee 4A-16. Proposed Recommended Practice—Determining Vehicle Signal Change Intervals: Part II, Literature Review and Committee Deliberations. ITE, Washington, D.C., c. 1989.
13. H. H. Bissell and D. L. Warren. The Yellow Signal is not a Clearance Interval. *ITE Journal*, Feb. 1981, pp. 14–17.
14. J. E. Baerwald (ed.). *Traffic Engineering Handbook*. 3rd ed. ITE, Washington, D.C., 1965.
15. R. H. Wortman and T. C. Fox. A Reassessment of the Traffic Signal Change Interval. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 62–68.
16. F. B. Lin. Timing Design of Signal Change Intervals. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 46–51.
17. J. E. Baerwald. *Transportation and Traffic Engineering Handbook*. ITE, New Jersey, 1976.
18. P. L. Olson and R. Rothery. Driver Response to Amber Phase of Traffic Signals. In *Highway Research Bulletin 330*, HRB, National Research Council, Washington, D.C., 1962, pp. 40–51.
19. W. S. Homburger. *Transportation and Traffic Engineering Handbook*. ITE, New Jersey, 1982.
20. M. S. Chang, C. J. Messer, and A. J. Santiago. Timing Traffic Signal Change Intervals Based on Driver Behavior. In *Transportation Research Record 1027*, TRB, National Research Council, Washington, D.C., 1985, pp. 20–30.

DISCUSSION

FENG-BOR LIN
Civil and Environmental Engineering, Clarkson University, Potsdam, N.Y. 13699-5710

The authors should be commended for their attempt to resolve the controversy surrounding the timing of the intergreen interval. Their paper is based on the argument that the dilemma zone concept as advanced by Gazis et al. (*1*) must always be used in evaluating a given intergreen interval. This bias has limited the scope of their literature review and discussions.

Much of the controversy in timing the intergreen interval stems from several equations suggested by ITE over the years. In 1985, for example, ITE proposed that the following equation be used to determine the yellow interval (*2*):

$$Y = t + V/[2\ (a + gG)] \tag{1}$$

where

Y = length of yellow interval,
t = driver perception/reaction time,
V = vehicle approach speed,
a = deceleration rate,
G = grade of approach lane, and
g = gravitational acceleration.

Although the root of this equation can be traced back to the dilemma zone concept, ITE (*2*) has also indicated that the primary measure of effectiveness for the yellow interval is the percentage of vehicles entering the intersection after the yellow interval expires. This measure of effectiveness reflects the need to reduce the potential of right-angle collisions. In other words, the real intention of Equation 1 is to ensure that the yellow interval is long enough to allow most drivers who are faced with a yellow light to come to a stop rather than continue to enter the intersection after the red onset. Unfortunately, because they are derived from dilemma considerations, Equation 1 and other similar equations are not compatible with this intended timing requirement. This incompatibility is one reason for the different interpretations of such equations; it is also a weakness of ITE's equations. Several studies of driver behavior (*3–5*) have consistently shown that the yellow interval needed to prevent a high per-

centage of drivers from entering on red is independent of vehicle approach speed. The underlying reason for this phenomenon is probably drivers' willingness to tolerate (or apply) greater deceleration rates at higher approach speeds. In any case, the findings of these studies have prompted suggestions to use a uniform yellow interval. The authors have ignored this side of the controversy.

Because the yellow interval needed to prevent most drivers from entering on red is independent of vehicle approach speed, the consideration of dilemma situations is really not that important. For the same reason, the use of the speed-location diagram as illustrated by the authors becomes an unnecessary exercise. For example, the experimental results reported in Table 1 can be logically explained in terms of yellow interval demand, which refers to the length of the yellow interval that is needed in a change interval to prevent drivers from entering on red. A typical cumulative distribution of yellow interval demand is shown in Figure 7. This distribution is based on observed driver behavior at intersections that are controlled with pretimed signals (*5*). It should be noted that such a distribution is independent of vehicle approach speed. Assuming that the drivers at the Maryland and the Georgia sites exhibited the same behavior as that shown in Figure 7, this figure can be used to predict the impact of the yellow interval on the conflict potential at these sites. For example, the Maryland site has a clearance distance of about 75 ft. If vehicles moved across this intersection at an average speed of 30 mph, then it would take an average of 1.7 sec to clear the intersection. With a 4.7-sec yellow interval during off-peak hours, this clearance time requires a driver to enter the intersection within the first 3.2 sec (e.g., 4.7 − 1.7 + 0.2) of the yellow interval in order to avoid occupying the intersection for more than 0.2 sec after the red onset. Figure 7 shows that in 55 percent of the change intervals, drivers will continue to enter the intersection after 3.2 sec of the yellow interval has elapsed. In other words, late entries can be expected to exist in 45 percent of the change intervals. Similar estimates can be obtained for the various conditions reported by the authors in Table 1. These estimates and their relationships to the respective conflict potentials are shown in Figure 8. Due to the lack of actual data on clearance time, the estimated percent-

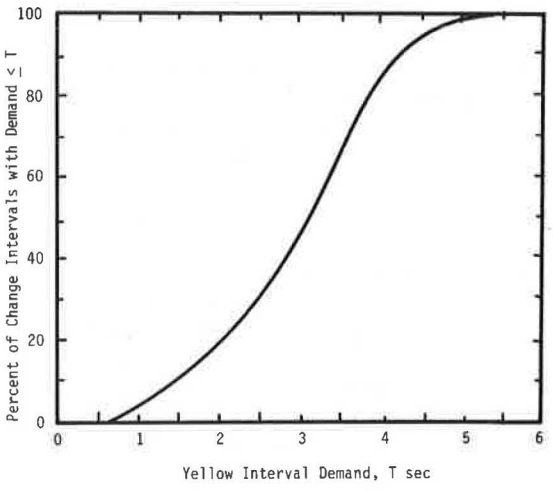


FIGURE 7 Cumulative distribution of yellow interval demand.

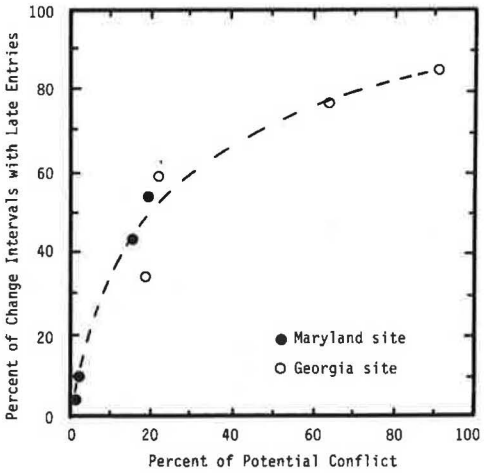


FIGURE 8 Relationship between conflict potential and percentage of change intervals with late entries.

age of change intervals with late entries are based on a clearance speed of 30 mph for off-peak hours and 25 mph for peak hours. This figure clearly shows that the impact of changing yellow interval can be reasonably predicted without knowing whether there is a dilemma.

In light of the predominant concern about right-angle collisions, I believe that the dilemma zone concept is elusive and its importance overstated. A situation that is a dilemma to one driver is not necessarily a problem to another.

REFERENCES

1. D. Grazis, R. Herman, and A. Maradudin. The Problem of the Amber Light in Traffic Flow. *Operations Research*, Vol. 8, No. 1, Jan.–Feb. 1960.
2. ITE Technical Committee 4A-16. Proposed Recommended Practice: Determining Vehicle Changed Intervals. *ITE Journal*, May 1985, pp. 61–64.
3. M. S. Chang, C. J. Messer, and A. J. Santiago. Timing Traffic Signal Change Intervals Based on Driver Behavior. In *Transportation Research Record 1027*, TRB, National Research Council, Washington, D.C., 1985, pp. 20–30.
4. R. H. Wortman and T. C. Fox. A Reassessment of the Traffic Signal Change Interval. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 62–68.
5. F. B. Lin and S. Vijaykumar. Timing Design of Signal Change Interval. *Traffic Engineering and Control*, Vol. 29, No. 10, Oct. 1988, pp. 531–536.

AUTHORS' CLOSURE

We thank Lin for his discussion and the opportunity to respond to his comments. Contrary to his expectation, our paper did not attempt to resolve the controversy but rather to propose a common framework that would help avoid the prevalent problem we discovered in our literature search relating to differences in definitions, methods of measurement, sampling, and so on. Among the advantages of adopting the speed-location diagram as the basic framework is that it can explicitly present aspects of the problem that have been persistently

considered by almost all, if not all, researchers in one way or another. These aspects include the speed and location of observed vehicles at the onset of the intergreen interval, the width of the intersection, the actions taken by observed drivers and their consequences, and the equations of motion governing the stop-or-go decision. Other methods of presenting observed data either omit some of these factors or consider them implicitly.

It is unfortunate that Lin failed to make a distinction between ITE's very restricted interpretation of the problem and the larger framework afforded by the full speed-location diagram, and that he dismissed the usefulness of the latter based on the shortcomings of the former, a practice against which we clearly warned. Of particular concern is the fact that the ITE formula is based on a design situation that involves a vehicle that, at the onset of the intergreen interval, happens to be traveling at the selected design speed and happens to be located behind the stop line at a distance exactly equal to the stopping distance corresponding to that speed. ITE's assumption, which is evident in Lin's Equation 1, is that if the signal were to be timed for this design situation, the conditions faced by vehicles approaching at other speed and location combinations would be covered. Our paper showed that this is not the case and that the speed-location diagram can provide valuable guidelines which can aid the timing engineer's judgment. More definite timing guidelines must await further study of experimental data in the context of the speed-location diagram.

We have also shown that the presentation of experimental data on the speed-location diagram can help researchers (particularly timing engineers) to systematically interpret their results by showing the intersection width, the characteristics of the signal, the speed and location of the subject vehicles at the onset of yellow, the drivers' decision to stop or go, and the consequences of these decisions. Our review has revealed that a large part of the conflicting conclusions reported in the literature can be traced to differences in sampling which can be made explicit through the use of the speed-location diagram. This point is discussed in the "Additional Comments" section of our paper, which we urge Lin to study more carefully.

Lin criticizes our paper for not discussing the potential effectiveness of adopting a constant yellow interval. Even though our work in this area has included the question of the division of the intergreen interval into yellow and all-red and other important issues, it was not possible to present our findings on these matters in a single paper because of length limitations. Our failure to emphasize this point may have contributed to Lin's primary concern with the yellow interval. A constant yellow would impart a degree of certainty to drivers who are not familiar with the intersection; what may be considered to be a reasonable action by a familiar driver may not always be expected to be so in the case of the unfamiliar driver who has no idea as to the timing or operational characteristics of an intersection visited for the first time. Whether a constant yellow, however, would cause other difficulties, such as excessive all-red intervals at some intersections, requires additional research.

Lin argues that ITE's proposal of using his Equation 1 to calculate the required yellow is inappropriate because several studies have "shown that the yellow interval needed to prevent a high percentage of drivers from entering on red is independent of vehicle approach speed." We agree that the ITE proposal is inappropriate, but our objection to it lies in the fact that the ITE formula is based on using a single approach speed rather than on examining the resulting conditions over the full range of speed and location conditions. As explained later, Lin's model also suffers from this shortcoming.

The studies cited by Lin prefer the use of time to reach the stop line as a superior criterion for the timing of the yellow interval. As usually defined, this time depends on both the vehicle's speed and its location at the onset of yellow, rather than speed alone. The ability of the speed-location diagram to show both vehicle speed and location at the onset of yellow is precisely one of the reasons that we propose its adoption. Moreover, as usually defined, the time to reach the stop line can be explicitly shown as the slope of a straight line drawn from the origin of the speed-location diagram to the point representing the initial conditions corresponding to a particular vehicle in the sample. Alternatively, a straight line from the origin with a particular slope would divide the speed-location space into two regions representing, respectively, the conditions that allow vehicles to enter the intersection within the time represented by the slope and those that do not. Bissel and Warren (*1*), in fact, used this concept with a slope that was equal to the duration of yellow. Our studies have shown that superposition of such a line on the speed-location diagram described in the paper can further enhance its usefulness. We have termed a diagram that includes this line and a representation of the division of the intergreen interval into yellow and all-red (see below) as the "expanded speed-location diagram," which we intend to describe more fully in a subsequent paper.

Figure 9 shows such an expanded speed-location diagram for the conditions prevailing at one of the sites studied by Lin (*2*). The dashed line has a slope that is equal to the duration of the existing yellow. Vehicles whose speed and location at the onset of yellow place them above this line cannot reach the stop line (i.e., cannot enter the intersection) at constant speed before expiration of the yellow interval, whereas vehicles that plot below the dashed line can. The diagram also shows two clearing lines: one based on the existing intergreen interval consisting of the sum of the yellow and all-red subdivisions (T = 5.9 sec), the other on the yellow portion only (Y = 3.0 sec). The mean approach speed was reported to be 30.6 mph with a 15th percentile of 24.2 mph and a 95th percentile of 35.8, but no information is provided regarding how speeds were measured, the particular catch zone used, or the distribution of the vehicles in the sample on the diagram. Of 11 sites studied, this intersection approach was observed to have the longest maximum change interval requirement of 8.9 sec, that is, at least one observed vehicle took this long from the onset of yellow to clear the intersection. Figure 9 shows that severe dilemma regions existed at this intersection approach; it is also noteworthy that the existing timing was shorter than that recommended by ITE. In other words, the speed-location diagram can be drawn irrespective of the method used in setting the timing and it should not be viewed as wed to the ITE method as Lin implied.

Had Lin chosen to report his sample of vehicles on the speed-location diagram, a more systematic investigation of

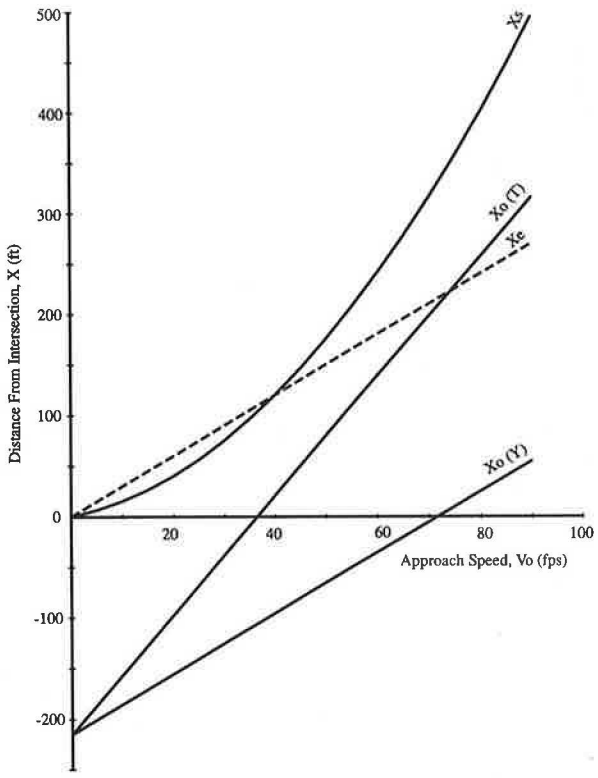


FIGURE 9 Expanded speed-location diagram for site studied by Lin.

the observed drivers could be possible. Instead, he chose to combine the data obtained at the 11 sites with widely varying widths and signal characteristics in order to examine "driver's aggregate needs" and to perform regression analyses in the hope of discerning a universally applicable relationship between the change interval requirement and the average approach speed. A year later, Lin et al. (*3*) used a similar procedure with additional data to arrive at a different change interval design equation. They also proposed the use of a yellow interval requirement distribution, similar to that shown by Figure 7 in Lin's discussion, to be used in timing the duration of yellow. This concept was further discussed by Lin and Vijaykumar (*4*). An examination of the yellow interval requirement distributions reported in these two papers (*3*,*4*) for several of the sites studied by Lin and his co-workers reveals a great variability between them. For example, the maximum observed yellow interval requirement (shown as approximately equal to 5.5 sec in Lin's discussion) actually ranged from a low of 3.3 sec at one site to a high of 6.4 sec at another (*4*). Thus Lin's claim that such a typical distribution exists cannot be substantiated at this time. In addition, in his illustration on how to apply his model, he calculated and subtracted from the intergreen interval a crossing time based on a single assumed clearing speed. This is reminiscent of the ITE interpretation of crossing time based on a single design speed, which is inappropriate. Moreover, the significant discrepancies evident on his Figure 8 between his calculated "percent of change intervals with late entries" and the "percent of potential conflicts" observed by Stimpson et al. (*5*) provide sufficient reason to question the validity of Lin's typical distribution of yellow interval requirements.

Finally, it is instructive to examine an explanation offered by Lin and Vijaykumar (*4*) of their finding that their observed maximum change interval demands were, on average, 53 percent longer than the actual settings of the change interval. They stated that "the drivers either could not or would not comply with the traffic regulation that requires one to clear the intersection by the time the change interval expired." In this connection, a paper cited in Lin's discussion observed that "the presence of a police vehicle at the site significantly reduced the percentage of vehicles entering on red" (*6*). In other words, Lin's preferred model cannot distinguish between observed noncomplying actions that are under the driver's control from those that are not. The speed-location representation can be more helpful in this respect.

In conclusion, consideration of Lin's discussion strengthens the case for adopting the speed-location diagram as a tool for evaluating research results. Lin and his co-workers are among the few researchers who have actually conducted field experiments. For this they should be commended. We encourage them, however, to report their data in the context of the speed-location framework so that progress can be made toward the resolution of the controversy.

REFERENCES

1. H. H. Bissell and D. L. Warren. The Yellow Signal Is Not a Clearance Interval. *ITE Journal*, Feb. 1981, pp. 14–17.
2. F. B. Lin. Timing Design of Signal Change Intervals. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 46–51.
3. F. B. Lin, D. Cooke, and S. Vijaykumar. Utilization and Timing of Signal Clearance Interval. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987, pp. 86–95.
4. F. B. Lin and S. Vijaykumar. The Timing Design of Signal Change Interval. *Traffic Engineering and Control*, Vol. 29, No. 10, 1988, pp. 531–536.
5. W. A. Stimpson, P. A. Zador, and P. J. Tarnoff. The Problem of the Amber Signal in Traffic Flow. *Operations Research*, Vol. 8, No. 1, Jan.–Feb. 1960.
6. R. H. Wortman and T. C. Fox. A Reassessment of the Traffic Signal Change Interval. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 62–68.

The authors are solely responsible for the contents of this paper.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Comparison of Left-Turn Accident Rates for Different Types of Left-Turn Phasing

JONATHAN UPCHURCH

Left-turn accident rates are compared for five types of left-turn phasing: permissive; leading exclusive/permissive; lagging exclusive/permissive; leading exclusive; and lagging exclusive. Two different study designs were used to compare left-turn accident rates; both a simple comparison design and a simple before-and-after design were used. Left-turn accident rate (number of left-turn accidents per million left-turning vehicles) is used to compare the relative safety of the different types of left-turn phasing. Left-turn accident rates are shown for each type of left-turn phasing and are further subdivided by whether there are two or three lanes of opposing traffic, by left-turn volume, and by opposing volume. The before-and-after data are categorized according to the types of left-turn phasing in the before and after periods. Observations and conclusions are made about the effect of volume, number of lanes of opposing traffic, and type of left-turn phasing on the accident rate.

During the past several years there has been a substantial interest in different types of left-turn signal phasing. Research has centered on both the operational and safety characteristics associated with different types of left-turn phasing. Agent (*1–3*), Beaudry (*4*), Mohle and Rorabaugh (*5*), the Florida Section of ITE (*6*), Cottrell and Allen (*7*), Machemehl and Mechler (*8*), the Colorado-Wyoming Section of ITE (*9*), Warren (*10*), and Upchurch (*11*) each conducted research that looked at various operational and safety aspects of left-turn phasing.

The research project described here compared the accident rates associated with different types of left-turn phasing.

PREVIOUS WORK COMPARING ACCIDENT EXPERIENCE

None of the previous studies compared the accident experience of all five types of left-turn phasing, namely: permissive; leading exclusive/permissive; lagging exclusive/permissive; leading exclusive; and lagging exclusive. Seldom are all five of these types of phasing used in one jurisdiction. Generally speaking, many of the previous research efforts have been before-and-after types of studies that compared a change from one type of phasing to another. Warren (*10*), for example, compared intersections that were changed from leading exclusive to leading exclusive/permissive and intersections that were changed from permissive to leading exclusive/permissive. Little research has compared the accident experience of leading phasing versus lagging phasing.

Department of Civil Engineering, Arizona State University, Tempe, Ariz. 85287-5306.

RESEARCH APPROACH

This study compares the relative safety of different types of left-turn phasing by comparing the accident rate for left-turn accidents. Left-turn accidents are those in which the manner of collision reported on an accident report form involves a vehicle turning left. The accident rate is based on the number of left-turn accidents and the associated left-turn volume. The rate is expressed in terms of number of left-turn accidents per 1 million left-turning vehicles.

ACCIDENT STATISTICS FOR A SIMPLE COMPARISON

One method of comparing the relative safety of different types of left-turn phasing is to compare accident rates on approaches with one type of left-turn phasing with accident rates on approaches with a second type of left-turn phasing. This is called a simple comparison design.

Two large data bases were created for this project. The data bases included information, by approach, for 495 signalized intersections on roadways maintained by the Arizona Department of Transportation (ADOT) and 132 signalized intersections in six local jurisdictions in Arizona. The data provided the opportunity to develop accident statistics and enabled a comparison. Statistics were developed for different types of left-turn phasing, varying numbers of opposing lanes (two or three), varying ranges of left-turn volume, and varying ranges of opposing volume.

Each sample used in developing the accident statistics represents a single approach at an intersection. A total of 523 samples (intersection approaches) were included in developing the accident statistics. Approaches with two opposing lanes had 329 samples; approaches with three opposing lanes had 194 samples. All approaches used for this analysis had a separate left-turn lane.

For intersections on the state highway system, most samples represent a 4-year accident history (1983 through 1986). For intersections in local jurisdictions, samples range from a minimum of 7 months to a maximum of 48 months (all in the period from 1981 to 1989). The "mean" accident rate is a

weighted average which is weighted in proportion to the time period sampled on an approach.

Gross accident statistics are shown in Table 1. Statistics are shown for five types of left-turn phasing: permissive; leading exclusive/permissive; lagging exclusive/permissive; leading exclusive; and lagging exclusive. Separate statistics are presented for locations having two opposing lanes of traffic and locations having three opposing lanes of traffic. The mean left-turn accident rate is shown along with the sample size (N) on which that mean rate is based.

The sample size for lagging exclusive phasing is too small to rely on the average accident rates for comparison purposes. For the four remaining types of phasing, the following observations and conclusions can be made about the statistics that are not stratified by volume:

- Leading exclusive phasing has the lowest left-turn accident rate.
- When there are two opposing lanes, lagging exclusive/permissive has the worst accident rate.
- When there are three opposing lanes, leading exclusive/permissive has the worst accident rate.
- For two opposing lanes, the order of safety (from best to worst) is leading exclusive, permissive, leading exclusive/permissive, and lagging exclusive/permissive. However, there is a small difference in the accident rate among the last three types of phasing.
- For three opposing lanes, the order of safety (from best to worst) is leading exclusive, lagging exclusive/permissive, permissive, and leading exclusive/permissive.
- In three out of four cases, accident rates are higher with three opposing lanes. The exception is for lagging exclusive/permissive phasing (although the difference in rates is small).

Tables 2 and 3 show similar accident statistics for various ranges of left-turn volume (vehicles per day) and various ranges of opposing volume (vehicles per day).

Opposing volume is defined as the through and right-turn volume on the approach opposite the left-turn movement. Again, the sample size for lagging exclusive phasing is too small to rely on the average accident rates for comparison purposes. For the four remaining types of phasing, the following observations and conclusions can be made about the statistics that are stratified by volume:

- Several cases have a sample size of five or less. No interpretations are made for these cases because it would be risky to make comparisons with mean accident rates based on such a small sample size.
- Leading exclusive phasing has the lowest left-turn accident rate in almost every case. This is true in every left-turn volume range and every opposing volume range except one (19 out of 20 cases).
- When there are two lanes of opposing traffic, lagging exclusive/permissive tends to have the worst accident rate.
- When there are three lanes of opposing traffic, leading exclusive/permissive tends to have the worst accident rate.
- When there are two lanes of opposing traffic, the order of safety (from best to worst) tends to be leading exclusive, permissive, leading exclusive/permissive, and lagging exclusive/permissive. However, there is a small difference in the accident rate among the last three types of phasing.
- When there are three lanes of opposing traffic, the order of safety (from best to worst) tends to be leading exclusive, lagging exclusive/permissive, permissive, and leading exclusive/permissive.
- Generally, accident rates are higher for three opposing lanes of traffic than for two opposing lanes of traffic. This is true in 30 out of 40 cases (combinations of phasing and volume). Lagging exclusive/permissive tends to be an exception to this rule.
- Some trends are apparent in the accident rate as a function of volume:
  - —For all four types of phasing (permissive, leading exclusive/permissive, lagging exclusive/permissive, and leading exclusive), with *two* opposing lanes of traffic, the accident rate *decreases* as left-turn volume *increases*. Figures 1 to 3 plot left-turn accident rate as a function of left-turn volume for three of these conditions (permissive phasing, lagging exclusive/permissive, and leading exclusive). Note that the vertical scale is different on each of these three figures.
  - —For all four types of phasing (permissive, leading exclusive/permissive, lagging exclusive/permissive, and leading exclusive), with *two* opposing lanes of traffic, the accident rate *increases* as opposing volume *increases*.
  - —For *three* opposing lanes of traffic, only one trend is apparent in left-turn accident rate as a function of volume.

TABLE 1 STATISTICS ON LEFT-TURN ACCIDENT RATE

		Permissive	Leading Exclusive/ Permissive	Lagging Exclusive/ Permissive	Leading Exclusive	Lagging Exclusive
2 opposing lanes	M	2.62	2.71	3.02	1.02	2.09
	N	162	62	44	57	4
3 opposing lanes	M	3.83	4.54	2.65	1.33	0.55
	N	25	52	35	80	2

M = Mean Left Turn Accident Rate

N = Number of approaches in the sample

Left Turn Accident Rate is based upon the number of left turn accidents (Manner of Collision) and the associated left turn volume. The rate is in terms of accidents per million left turning vehicles. Each sample represents a single approach at an intersection.

TABLE 2 STATISTICS ON LEFT-TURN ACCIDENT RATE STRATIFIED BY LEFT-TURN VOLUME

Left Turn Vol.	Permissive		Leading Exclusive/ Permissive		Lagging Exclusive/ Permissive		Leading Exclusive		Lagging Exclusive	
	MEAN	N	MEAN	N	MEAN	N	MEAN	N	MEAN	N
2 Opposing Lanes										
0-1000	3.07	93	4	16	4.71	10	1.24	14	6.3	1
1000-2000	2.38	51	2.44	25	2.89	13	1.42	22	1.43	1
2000-3000	.87	13	2.43	16	2.66	9	.51	13	.62	1
3000-4000	1.62	3	2.87	3	2.19	7	.52	2	N.A.	0
>4000	.45	2	.84	2	1.21	5	.24	6	0	1
Cumu-lative	2.62	162	2.71	62	3.02	44	1.02	57	2.09	4
3 Opposing Lanes										
0-1000	4.21	8	4.33	17	1.11	7	1.37	12	1.66	1
1000-2000	3.51	12	5.94	8	4.34	12	1.09	23	0	1
2000-3000	4.06	5	3.98	11	2.87	6	1.26	26	N.A.	0
3000-4000	N.A.	0	3.98	11	2.03	6	.84	12	N.A.	0
>4000	N.A.	0	5.27	5	1.67	4	.92	7	N.A.	0
Cumu-lative	3.83	25	4.54	52	2.65	35	1.33	80	.55	2

Left Turn Accident Rate is based upon the number of left turn accidents (Manner of Collision) and the associated left turn volume. The rate is in terms of accidents per million left turning vehicles. Each sample represents a single approach at an intersection.

Left Turn Volume is the 24 hour left turn volume.
MEAN is the mean accident rate for the approaches in the sample.
N is the sample size.
Cumulative is the weighted average mean for all volumes.

TABLE 3 STATISTICS ON LEFT-TURN ACCIDENT RATE STRATIFIED BY OPPOSING VOLUME

Opposing Volume	Permissive		Leading Exclusive/ Permissive		Lagging Exclusive/ Permissive		Leading Exclusive		Lagging Exclusive	
	MEAN	N	MEAN	N	MEAN	N	MEAN	N	MEAN	N
2 Opposing Lanes										
0-5000	1.4	71	1.97	15	1.43	5	.23	9	0	1
5000-10000	1.98	58	2.92	21	3.26	15	.49	17	1.43	1
10000-15000	3.54	17	2.89	19	3.47	18	2.07	19	3.46	2
15000-20000	6.08	8	2.33	5	3.54	2	.64	6	N.A.	0
>20000	4.99	8	4.54	2	2.37	2	.69	6	N.A.	0
Cumu-lative	2.62	162	2.71	62	3.02	42	1.02	57	2.09	4
3 Opposing Lanes										
0-5000	3.28	5	3.91	2	N.A.	0	.25	3	1.66	1
5000-10000	2.05	7	4.78	10	2.57	8	1.01	12	N.A.	0
10000-15000	4.83	5	4.32	12	3.3	11	.98	22	.0	1
15000-20000	6.61	4	4.98	16	2.51	10	1.15	17	N.A.	0
>20000	2.78	4	4.07	12	1.88	6	1.45	26	N.A.	0
Cumu-lative	3.83	25	4.54	52	2.65	35	1.33	80	.55	2

Left Turn Accident Rate is based upon the number of left turn accidents (Manner of Collision) and the associated left turn volume.Rate is in terms of accidents per million left turning vehicles. Each sample represents a single approach at an intersection.

Opposing Volume is the 24 hour opposing volume (through and right turning vehicles on the opposite approach).
MEAN is the mean accident rate for the approaches in the sample.
N is the sample size.
Cumulative is the weighted average mean for all volumes.

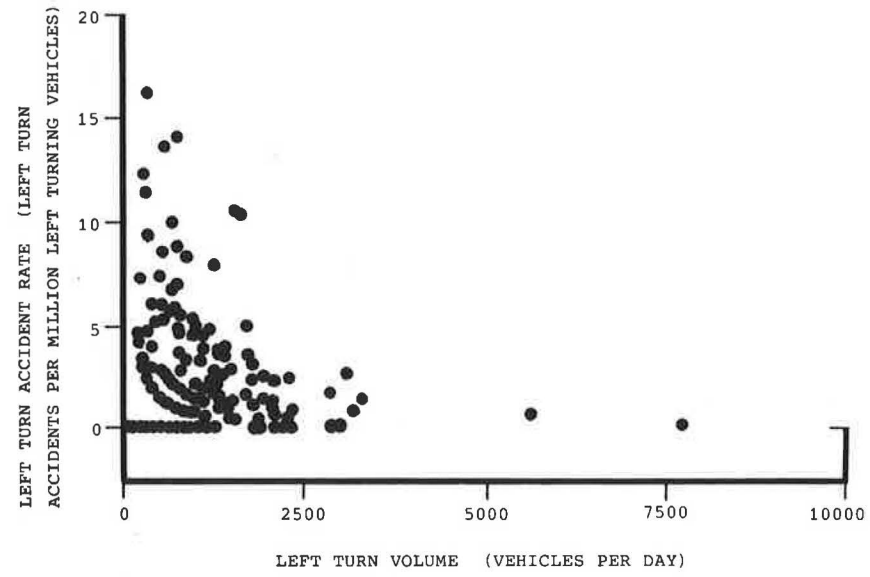


FIGURE 1 Left-turn accident rate plotted for permissive phasing.

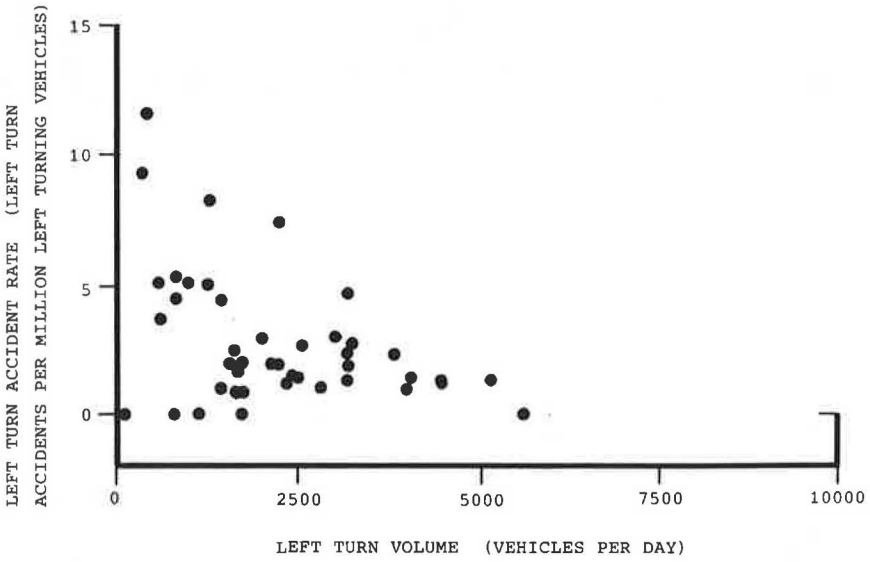


FIGURE 2 Left-turn accident rate plotted for lagging exclusive/permissive phasing.

For permissive left-turn phasing the accident rate *increases* as opposing volume *increases*.

The study team also looked at accident statistics for conditions that were stratified by both left-turn volume and opposing volume at the same time. This stratification would allow a traffic engineer to pick a range of left-turn volume, a range of opposing volume, a number of opposing lanes, and a type of phasing and determine the accident rate for those conditions. For example, the condition of left-turn volumes between 0 and 1,000 per day, opposing volume between 0 and 5,000 per day, two opposing lanes, and permissive left-turn phasing had a left-turn accident rate of 1.53 (based on a sample size of 44).

The availability of accident rate information of this form would be a tremendous asset to the traffic engineer. Unfortunately, stratifying conditions to this level of detail resulted in very small sample sizes for most cases. Eighty-eight percent of the cases had a sample size of five or less. Forty-two percent of the cases had a sample size of zero.

ACCIDENT STATISTICS FOR CONVERSIONS

A second means of comparing the relative safety of different types of left-turn phasing is to compare the accident experience before and after a location had been converted from one type of phasing to another. This is the simple before-and-after design. To make this type of comparison, additional information was obtained on conversions from one type of phasing to another for both ADOT roadway intersections and local jurisdiction intersections.

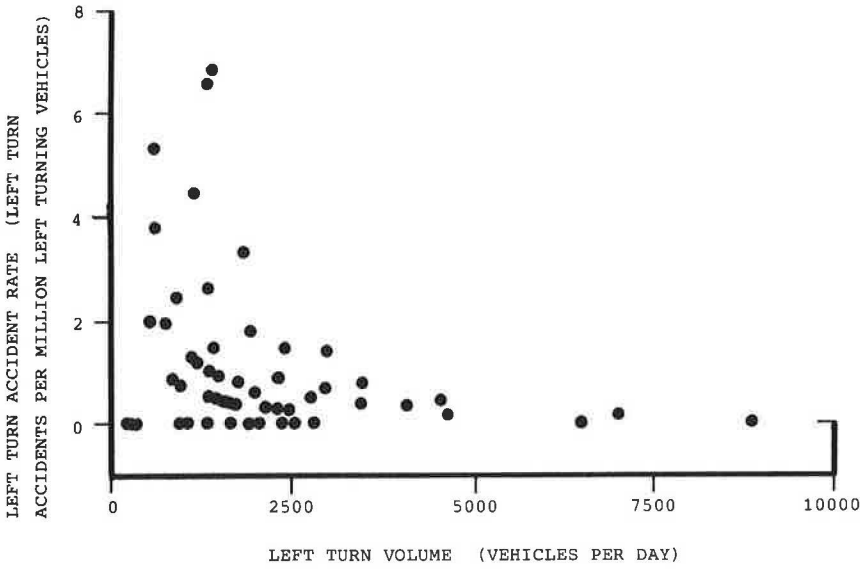


FIGURE 3 Left-turn accident rate plotted for leading exclusive phasing.

Six local jurisdictions provided information. The number of intersection approaches in each jurisdiction that were usable for the analysis were as follows:

Jurisdiction	*No. of Usable Approaches*
ADOT	15
Glendale	12
Maricopa County	0
Mesa	0
Pima County	3
Scottsdale	157
Tempe	7
Total	194

The local jurisdiction conversions used for a before-and-after analysis included some conversions that were made in 1984 and several that were made in late 1988 and early 1989. For each conversion, 4 years of before accident data and 4 years of after accident data were used where available. In many cases, such as the conversions done in late 1988 and early 1989, a shorter after period was available. In most of these cases, accident data through the end of 1989 were acquired.

With five different types of left-turn phasing, 20 different conversions could take place. The different types of conversions are not equally popular. For example, it is rare to convert from some more restrictive type of phasing to permissive phasing. Among the 194 approaches used in the statistical evaluation, the more popular types of conversion and their frequency are noted below:

Conversion	*Frequency*
Permissive to leading exclusive/permissive	20
Permissive to lagging exclusive/permissive	17
Leading exclusive/permissive to permissive	17
Leading exclusive/permissive to lagging exclusive/permissive	73
Leading exclusive to leading exclusive/permissive	25
Leading exclusive to lagging exclusive/permissive	15
Leading exclusive to lagging exclusive	22

These seven types of conversion accounted for 189 of the 194 intersection approaches. Eleven of the 20 possible types of conversions were never done.

Tables 4 and 5 show before and after accident statistics for the 194 approaches that were converted from one type of left-turn phasing to another. The left-turn accident rate is based on the number of left-turn accidents (manner of collision) and the associated left-turn volume. The rate is in terms of accidents per million left-turning vehicles. Each sample represents a single approach at an intersection. Data are shown for the before and after accident rates, the total number of months of data in the before period and in the after period, and the number of intersection approaches on which the statistics are based. Table 4 shows statistics for intersections with two opposing lanes of traffic; Table 5 shows statistics for intersections with three opposing lanes of traffic.

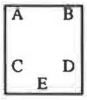
The following observations and conclusions can be made for the conversions made at approaches having two opposing lanes of traffic.

- Each before case and after case has at least 5½ approach years of data on which the statistics are based.
- The following conversions resulted in decreases in the left-turn accident rate:
  - —From permissive to leading exclusive/permissive,
  - —From permissive to lagging exclusive/permissive, and
  - —From leading exclusive/permissive to lagging exclusive/permissive.
- The following conversions resulted in increases in the left-turn accident rate:
  - —From leading exclusive/permissive to permissive,
  - —From leading exclusive to leading exclusive/permissive,
  - —From leading exclusive to lagging exclusive/permissive, and
  - —From leading exclusive to lagging exclusive.

TABLE 4 BEFORE AND AFTER LEFT-TURN ACCIDENT RATES FOR APPROACHES CONVERTED FROM ONE TYPE OF PHASING TO ANOTHER—TWO OPPOSING LANES

BEFORE TYPE OF PHASING	AFTER TYPE OF PHASING: Permissive	Leading E/P	Lagging E/P	Leading Exclusive	Lagging Exclusive
Permissive	4.77 3.49 608 425 17	5.44 4.16 359 131 9			
Leading E/P	2.07 2.66 462 340 14	3.10 2.25 1170 622 35			
Lagging E/P					
Leading Exclusive	0.93 3.11 144 70 3	0.38 1.57 220 67 6		1.46 1.91 346 144 10	
Lagging Exclusive					

KEY



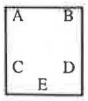
A = Before left turn accident rate
B = After left turn accident rate
C = Total number of months of before data
D = Total number of months of after data
E = Number of intersection approaches

Left Turn Accident Rate is based upon the number of left turn accidents (Manner of Collision) and the associated left turn volume. The rate is in terms of accidents per million left turning vehicles. Each sample represents a single approach at an intersection.

TABLE 5 BEFORE AND AFTER LEFT-TURN ACCIDENT RATES FOR APPROACHES CONVERTED FROM ONE TYPE OF PHASING TO ANOTHER—THREE OPPOSING LANES

BEFORE TYPE OF PHASING	AFTER TYPE OF PHASING: Permissive	Leading E/P	Lagging E/P	Leading Exclusive	Lagging Exclusive
Permissive	4.64 5.55 144 77 3	8.75 1.37 194 59 8	18.96 0.36 87 67 3		
Leading E/P	2.25 5.85 82 73 3	4.54 2.74 1181 831 38	7.08 0.75 12 68 2		
Lagging E/P					
Leading Exclusive	1.40 4.72 998 594 22	2.13 1.03 329 84 9		0.35 0.35 390 114 12	
Lagging Exclusive					

KEY



A = Before left turn accident rate
B = After left turn accident rate
C = Total number of months of before data
D = Total number of months of after data
E = Number of intersection approaches

Left Turn Accident Rate is based upon the number of left turn accidents (Manner of Collision) and the associated left turn volume. The rate is in terms of accidents per million left turning vehicles. Each sample represents a single approach at an intersection.

- The statistics for conversions from permissive to leading exclusive/permissive and from leading exclusive/permissive to permissive reinforce each other. Both statistics suggest that leading exclusive/permissive is safer than permissive.

The following observations and conclusions can be made for the conversions made at approaches having three opposing lanes of traffic.

- Each before case or after case has at least 5 approach years of data on which the statistics are based.
- The following conversions resulted in decreases in the left-turn accident rate:
  - —From permissive to lagging exclusive/permissive,
  - —From permissive to leading exclusive,
  - —From leading exclusive/permissive to lagging exclusive/permissive,
  - —From leading exclusive/permissive to leading exclusive, and
  - —From leading exclusive to lagging exclusive/permissive.
- The following conversions resulted in increases in the left-turn accident rate:
  - —From permissive to leading exclusive/permissive,
  - —From leading exclusive/permissive to permissive, and
  - —From leading exclusive to leading exclusive/permissive.
- The statistics for conversions from permissive to leading exclusive/permissive, and from leading exclusive/permissive to permissive contradict each other. The former statistic suggests that permissive phasing is safer. The latter statistic suggests that exclusive/permissive phasing is safer. It is possible that conditions at these two sets of intersections are different (traffic volumes, for example) and that these differences may account for the contradiction.
- The statistics for conversions from leading exclusive to leading exclusive/permissive, and from leading exclusive/permissive to leading exclusive reinforce each other. Both statistics suggest that leading exclusive is safer than leading exclusive/permissive.

The cases with two opposing lanes of traffic can be compared to those with three opposing lanes of traffic. In most cases the trends are the same. For example, a conversion from leading exclusive/permissive to permissive will result in an increased accident rate for approaches with two opposing lanes of traffic *and* for approaches with three opposing lanes of traffic.

In two cases, however, the trends are opposite. For two opposing lanes of traffic, a conversion from permissive to leading exclusive/permissive results in a decrease in accident rate. The opposite is true for three opposing lanes. This finding for three opposing lanes supports the view of some traffic engineers who are reluctant to use exclusive/permissive phasing with three opposing lanes because a larger gap is required, because it is more difficult for the driver to judge an acceptable gap, and because there is a greater chance that an oncoming vehicle in one lane will be masked out by a vehicle in another lane.

The other case in which trends are opposite is conversion from leading exclusive to lagging exclusive/permissive. For two opposing lanes, this conversion results in an increase in accidents. For three opposing lanes, it results in a decrease.

INTERPRETATION OF STUDY DESIGNS

It is important to understand some of the limitations of the accident rate information presented here. Both designs—the simple comparison and the before-and-after design—are simple. The intersections used to develop these accident statistics were *not* randomly selected. They are simply the intersections for which jurisdictions were able to provide all of the necessary data. Although efforts were made to make intersections as alike as possible (in type of phasing, number of opposing lanes, left-turn volume, opposing volume, and the existence of a separate left-turn lane), there may still be differences in intersection characteristics among the different groups.

Although these limitations are shortcomings of the study design, the strength of this study is that a very large sample size was involved in both designs. The simple comparison design included 523 intersection approaches, and the before-and-after design included 194 intersection approaches. The fact that all the necessary data, including turning-movement counts, were available at such a large number of intersections is an achievement, in comparison to other studies of left-turn signal phasing.

CONCLUSIONS

The choice of left-turn phasing type at a signalized intersection affects the left-turn accident rate. The accident rate is also influenced by the number of opposing lanes of traffic, left-turn volume, and the volume of opposing traffic. The traffic engineer should consider each of these factors when selecting the type of left-turn phasing to be installed at an intersection.

This information on relative safety will assist the traffic engineer in selecting the type of left-turn phasing to be used at a signalized intersection.

ACKNOWLEDGMENT

This research was sponsored by the Arizona Department of Transportation.

REFERENCES

1. K. R. Agent and R. C. Deen. *Warrants for Left-Turn Signal Phasing*. Division of Research, Bureau of Highways, Kentucky Department of Transportation, Lexington, Oct. 1978.
2. K. R. Agent and R. C. Deen. Warrants for Left Turn Signal Phasing. In *Transportation Research Record 737*, TRB, National Research Council, Washington, D.C. 1979, pp. 1–10.
3. K. R. Agent. *An Evaluation of Permissive Left Turn Phasing*. Division of Research, Bureau of Highways, Kentucky Department of Transportation, Lexington, April 1979.
4. P. M. Beaudry. *Guidelines for the Installation of Separate Left-Turn Signal Phasing*. Northwestern University Traffic Institute, Evanston, Ill., Aug. 1979.
5. R. M. Mohle and T. Rorabaugh. *Left Turn Phasing, Volume 4: A Study of Clearance Intervals, Flashing Operation, and Left-Turn Phasing at Traffic Signals*. Report FHWA-RD-78-49. FHWA, U.S. Department of Transportation, May 1980.
6. Florida Section, ITE. Left Turn Phase Design in Florida. *ITE Journal*, Sept. 1982, pp. 28–35.
7. B. H. Cottrell and G. R. Allen. *Guidelines for Exclusive/Per-*

missive Left-Turn Signal Phasing. Virginia Highway and Transportation Research Council, Charlottesville, July 1982.
8. R. B. Machemehl and A. M. Mechler, *Procedural Guide for Left Turn Analysis*. Center for Transportation Research, University of Texas, Austin, Nov. 1983.
9. Colorado-Wyoming Section, ITE. A Study of Use of Warrants for the Installation of Left-Turn Phasing at Signalized Intersections. ITE, Washington, D.C., March 1985.
10. D. L. Warren. Accident Analysis of Left-Turn Phasing. *Public Roads*, Vol. 48, No. 4, March 1985, pp. 121–127.
11. J. E. Upchurch. Guidelines for Selecting Type of Left Turn Phasing. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 30–38.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Evaluation of Delay Models for Motor Vehicles at Light Rail Crossings

RICHARD A. BERRY AND JAMES C. WILLIAMS

The application of theoretically and empirically based delay equations, developed for isolated traffic signal operation, to the problem of isolated at-grade light rail crossings was evaluated. Data were collected at 24 crossings of five transit authorities to validate a model for average stopped delay from among 21 candidates. The successively validated model was the pretimed delay equation from the 1985 *Highway Capacity Manual*. When a background cycle is imposed by traffic signals controlling the crossing, the delay model developed by Allsop was found to be suitable when applied with its lower limit parameters.

Because light rail transit (LRT) has minimal right-of-way requirements and, for the most part, crosses streets and highways at-grade, it is much less costly to build than conventional heavy rail transit, with its grade-separated operation. However, when streets are crossed at-grade, the quality of traffic service provided within the street network decreases because of an increase in stops and delay to motor vehicle traffic.

For most recently constructed LRT systems, the decision to operate at-grade has been a policy decision due, in part, to economics. In addition, there is a widespread perception among transit planners that at-grade crossing capacity should be based on the potential person-capacity of the crossing, thus inherently favoring unobstructed right-of-way for the light rail vehicle. Because there are currently no accepted methods to measure the traffic impact of at-grade light rail crossings, each transit property is required to develop its own approach, often without the benefit of the experience of other transit properties.

Little research has been done to objectively measure the magnitude of the traffic impacts. Research has been concentrated on the application of techniques developed for the evaluation of signalized street intersections (*1–3*), but has generally been unable to support the validity of their use at LRT crossings with field data, simulation, or statistical analysis (*1,3,4*).

In this work, theoretically and empirically based delay equations developed for isolated traffic signals are evaluated and compared with field data collected from five LRT systems currently operating in the United States.

OPERATIONAL OVERVIEW OF LIGHT RAIL TRANSIT

The degree of interaction between the light rail vehicles and motor vehicle traffic (and, thus, their mutual impacts) is briefly discussed in this section. Four general operational aspects of light rail transit affecting this interaction are

- Operating philosophy of the transit authority,
- Location of the at-grade crossing with respect to nearby signalized intersections,
- LRT scheduling, and
- Traffic control devices at the crossing.

Operating Philosophy

The operating philosophy of transit authorities for at-grade crossings can vary from requiring the trains to follow the rules of the road and wait for gaps in the traffic to cross streets, to having unconditional preemptive authority to cross streets on demand. The particular operating philosophy of a transit authority also affects the other aspects.

Crossing Location

At-grade LRT crossings are either isolated from signalized intersections or are adjacent to or within signalized intersections. In the latter cases, the timing of the traffic signal must consider the light rail movement through the intersection, either by allowing the LRT vehicle to preempt the traffic signal or by allowing the LRT vehicle to move through the intersection on a particular phase. This work, however, is principally concerned with isolated crossings. If a crossing is sufficiently removed from a signalized intersection, traffic impacts are limited only to those associated with crossing capacity and delay. It must be recognized, however, that where LRT crossings are close to traffic signals, queues generated by the crossing may interfere with the intersection and vice versa. Simulation experiments by Cline et al. (*4*) found that traffic signals within about 400 ft of LRT crossings may experience this problem.

Schedule

In measuring delay at LRT crossings, the effect of the operating schedule must be considered. LRT schedules are typically constructed in 30-sec to 1-min increments and LRT vehicle operators are often considered "on time" if their arrival at a time point is within the schedule increment. This scheduling ensures a variance in the arrival time of light rail vehicles at crossings. Sources of the variance include

R. A. Berry, DeShazo, Starek and Tang, Inc., 2317 Colonial, Mesquite, Tex. 75150. J. C. Williams, Department of Civil Engineering, University of Texas, Arlington, Tex. 76019-0308.

- Variations in station dwell time caused by surges in boarding and alighting, and the occasional boarding and alighting of handicapped passengers;
- Traffic congestion when operations are in mixed traffic;
- Variable time headways (scheduled) caused by changing ridership throughout the day; and
- Individual train operator and vehicle characteristics.

Traffic Control Devices

Two types of traffic control devices are used at LRT crossings: passive (such as crossbucks or stop signs) and active (flashing lights, with or without gates, and standard traffic signals). Active traffic control devices allow moderate disruption to motor vehicle traffic while providing a higher level of protection to the alert driver.

LITERATURE REVIEW

Much of the previous work on the impact of LRT crossings on public streets has concentrated on whether at-grade crossing capacity is adequate. This work has concentrated on two major measures of effectiveness: the volume-to-capacity (*v/c*) ratio and delay.

When dealing with isolated LRT crossings (i.e., crossings not directly associated with a signalized intersection), the *v/c* ratio is seldom of much concern. An at-grade LRT crossing with active traffic control devices is essentially an intersection with two-phase traffic signal control. Even on LRT lines with relatively frequent service, the fraction of time the cross street is blocked is typically low when compared with the fraction of time the street is "blocked" at nearby signalized intersections by the red interval. (For the purposes of this work, the time when the motor vehicle traffic may not cross the LRT guideway is considered to be the blockage time.) Multiphase signalized intersections are much more likely than at-grade LRT crossings to control the overall capacity of a street (*5*). Therefore, most techniques using the *v/c* ratio have been developed for LRT crossings adjacent to, or within, intersections, where one or more of the timed phases may be controlled by the LRT vehicle. Often an adjustment of the *v/c* ratio to account for the LRT crossing has been used as a first step in estimating the resulting delay on the cross street.

Stone and Wild (*1*), using capacity techniques from the 1965 *Highway Capacity Manual* (HCM) (*6*) and the work of May and Pratt (*7*) and Crommelin (*8*), developed a regression equation relating the intersection utilization factor (*v/c* ratio) to individual vehicle delay.

The San Diego Association of Governments (*9*) also used the 1965 HCM (*6*) to assign levels of service to LRT crossings. Load factors were approximated using field-measured stopped delay at LRT crossings, and were then compared to the cycle length of adjacent traffic signals.

Gibson et al. (*10*) assumed that nonconflicting traffic movements would be allowed to move with the LRT vehicles during preemption, and that all intersection traffic must be stopped only during the clearance time immediately prior to preemption. This assumption typically resulted in increases to the *v/c* ratio of less than 5 percent.

A method developed for the San Diego Trolley Bayside Line (*11*) estimated the percentage of the traffic signal cycle used by light rail operations for each movement, then increased the affected movements appropriately. Level of service was then calculated, as described in Circular 212 (*12*), using critical movements.

Grote (*2*) also used a critical movement analysis (*12*) to estimate the *v/c* ratio of the crossing (assuming one of the critical movements to be LRT). Individual vehicle delay was estimated using Stone and Wild's regression equation (*1*).

Radwan and Hwang (*3*) estimated the gain of passenger-seconds through intersections containing LRT using a modified Webster's delay equation and a probabilistic procedure for estimating train arrivals at the LRT crossing.

The NETSIM traffic simulation model was used by Cline et al. (*4*) to evaluate delay and queuing for isolated LRT crossings and those within intersections. Regression equations predicting motor vehicle delays were formulated from the results of the simulation effort.

EXPERIMENTAL DESIGN

Field studies were performed to collect empirical data from a range of transit properties in order to validate the measures of effectiveness proposed for traffic impact evaluation. Although this step is crucial to evaluating alternative methods of estimating the traffic impacts of at-grade LRT crossings, it has been omitted in most previous studies. Only Cline et al. (*4*) attempted validation, and then by using simulation methods, which, although using typical values for signalized intersections, were not validated for light rail applications.

Data Identification

Two sources were used to identify what traffic, roadway, and light rail data should be collected. The 1985 HCM (*13*) contains a detailed list of traffic and roadway factors that affect roadway capacity and traffic delay estimates. Gerlough and Huber (*14*) review a number of well-known models for estimating the motor vehicle delay at signalized intersections. The components of these models provided additional guidance on the selection of appropriate traffic and light rail data. Specific factors targeted for collection included

- Area type,
- Number of lanes,
- Lane widths,
- Grades,
- Volumes,
- Arrival type,
- Saturation flow rates,
- Cycle length,
- Effective green times, and
- Effective red times.

Site Selection

Principal site selection factors included

- The ability to collect traffic impact data at isolated LRT crossings (i.e., crossing without parallel movements of motor vehicle traffic); and
- The need to obtain data from a range of roadway, traffic, and light rail conditions.

Five transit authorities with isolated LRT crossings were selected. Two other transit authorities with only median or adjacent running alignments were visited.

The 14 crossings studied included a range of traffic control strategies, traffic volumes, approach lanes, light rail headways, train lengths, and light rail control criteria. The transit properties themselves included older properties with substantial operating experience and newer properties which had only recently begun rail operations.

Candidate Transit Properties

Seven transit properties were visited during the course of the two field studies. The seven transit properties were

1. Port Authority Transit (PAT) of Allegheny County, serving the Pittsburgh, Pennsylvania, metropolitan area;
2. Niagara Frontier Transportation Authority (NFTA), serving Buffalo, New York;
3. Greater Cleveland Regional Transportation Authority (GCRTA), serving the Cleveland, Ohio, metropolitan area;
4. San Francisco Municipal Railway (Muni), serving the San Francisco, California, metropolitan area;
5. Santa Clara County Transportation Agency (SCCTA), serving the San Jose, California, metropolitan area;
6. Sacramento Regional Transportation District (SRTD), serving Sacramento, California; and
7. San Diego Trolley, Inc. (SDTI), serving the San Diego, California, metropolitan area.

Data on the traffic impacts of isolated LRT crossings were successfully collected for every authority except GCRTA and SCCTA. These two authorities did not have any isolated LRT crossings.

Data Collection Procedures

A simple data collection procedure was used at each transit property visited. First, each at-grade LRT crossing was inspected to determine which crossings were suitable for the study. Criteria included traffic volumes, number of lanes, location and types of nearby traffic control devices, sight lines for data collection, and crossing warning devices. Second, a meeting was held with personnel from the transit property to select crossings and discuss light rail operations. Next, a street intersection in the vicinity of the crossing was selected for collecting saturation flow rate data typical of the area. Finally, the actual collection of the field data was performed both at the LRT crossings and at the street intersections. Upon completing collection of the field data, each site was measured and check lists were reviewed to ensure that all necessary data had been collected.

Actual data collection of at-grade LRT crossing operations in the field was performed via videotape of the selected LRT crossings. In total, approximately 53 hr of videotape were made of LRT crossing operations.

Saturation flow rate data were collected at signalized intersections near each crossing to serve as control data for typical traffic operations in the area. Intersections were chosen where the traffic characteristics were judged to be similar based on the cross-section of the intersection approach, the traffic volume, the traffic speed, and the intersection environment.

Data Analysis

Initial data analysis was performed by viewing the videotapes and manually removing information from them. The video camera's clock superimposed on each tape provided a time base for the data. Data collected from the videotape included

1. Traffic volumes in 15-sec and 1-min increments,
2. Train volumes,
3. Train headways,
4. Crossing protection equipment operating times,
5. Number of motor vehicle stops,
6. Individual motor vehicle stopped delay,
7. First car lost times,
8. Saturation flow rates, and
9. Queue lengths.

Further analysis of the data obtained from the videotapes was performed with the goal of estimating

1. Arrival distribution of the motor vehicle traffic,
2. Distribution of train headways,
3. Effective blockage time, or red time, incurred at the crossing,
4. Crossing's effective green time to cycle length ratio,
5. Crossing's capacity,
6. Crossing's *v/c* ratio,
7. Percent stops for the motor vehicle traffic,
8. Average individual stopped delay, and
9. Mean, 85th percentile, and 95th percentile queue lengths.

Many of the findings of the field studies have been reported by Berry and Williams (*15*).

During the planning of the field studies, three methods for estimating delay at signalized intersections were investigated. The first might be described as an input-output method. For this method, the elapsed travel time over a length of roadway is measured for each vehicle. The approach delay is the difference between the "normal" travel time and the travel time where delay due to LRT vehicle crossings is incurred. This method was rejected because of the difficulty in videotaping the necessary length of roadway and the need to define a "normal" travel time.

The second method investigated is widely used for intersection delay studies and results in an estimate of stopped delay. Described by Reilly et al. (*16*) and the 1985 HCM (*13*), this method used counts of stopped vehicles taken at regular intervals, such as 10 or 15 sec. The sum of the number of stopped vehicles is then multiplied by the interval between the stopped vehicle counts and divided by the total volume

during the study period. The one drawback is that the time interval should not be an integer factor of the traffic signal cycle length. This method was rejected because that constraint could not be ensured.

The third method investigated was used in this project. It consists of tracking the trajectories of individual vehicles on their approach to the crossing and noting the total time, if any, that they are stopped. Although this method of estimating stopped delay is more labor intensive than the other two, it should provide the most accurate and precise estimate of crossing delay to motor vehicles.

SUMMARY OF FINDINGS

Estimates of the average individual stopped delay for the different LRT crossing studies are shown in Tables 1 through 5. These estimates were calculated by summing the total amount of stopped delay observed over each discrete time period (one signal cycle or 5 min, depending on the method of crossing control; 15 min; or 1 hr) and dividing by the traffic volume observed during that time period.

The study sites have been classified by the type and operation of control at the crossing, and are shown in separate tables. Crossings with no control or flashing light units are listed in Table 1; those with flashing light units and gates are listed in Table 2. Two crossings were controlled by standard traffic signals which alternated right-of-way between the cross street and the LRT; there was no street parallel with the LRT. These signals operated on a background cycle and would periodically stop vehicles when no train was crossing. Furthermore, the LRT vehicle preempted signal operation upon its approach to the crossing. Therefore, delay to motor vehicles at the crossings can be separately tabulated: delay when stopped for LRT crossings (Table 3) and delay when stopped due to the normal cycling of the traffic signal (Table 4). Total delay (summing delay under both conditions) at these two crossings is shown in Table 5.

TABLE 1 LIGHT RAIL-RELATED STOPPED DELAY—NO CONTROL AND FLASHING LIGHT UNIT CONTROL

Site	Cross Street	Peak Hour Volume (vph)	No. of Peak Hour Trains (tph)	Crossing v/c Ratio	Stops on Approach (%)	Individual Stopped Delay		
						Range for 5 min period (sec/veh)	Range for 15 min period (sec/veh)	Hourly Average (sec/veh)
	Ocean Avenue							
1	AM Obs	261	16	0.22	4.8	0.0-1.4	0.0-0.9	0.3
2	PM Obs	262	12	0.29	8.0	0.0-2.3	0.0-1.4	0.6
	Potomac Avenue							
3	AM Obs	417	28	0.47	51.5	0.0-19.9	1.6-14.2	11.8
4	PM Obs	270	26	0.25	32.3	0.0-27.6	2.1-13.8	8.1
	Mt. Lebanon Blvd							
5	AM Obs	561	22	0.38	11.2	0.0-3.5	0.4-2.0	1.0
6	PM Obs	614	41	0.45	22.6	0.0-14.4	0.4-5.6	2.6

TABLE 2 LIGHT RAIL-RELATED STOPPED DELAY—FLASHING LIGHT UNITS WITH GATES

Site	Cross Street	Peak Hour Volume (vph)	No. of Peak Hour Trains (tph)	Crossing v/c Ratio	Stops on Approach (%)	Individual Stopped Delay		
						Range for 5 min period (sec/veh)	Range for 15 min period (sec/veh)	Hourly Average (sec/veh)
	65th Street							
7	AM Obs	1115	8	0.42	9.7	0.0-13.7	1.4-5.4	2.6
8	PM Obs	1054	8	0.38	9.2	0.0-19.8	0.3-12.5	4.1
	Alhambra Blvd							
9	AM Obs	379	8	0.28	9.1	0.0-5.1	0.1-3.0	1.3
10	PM Obs	456	8	0.36	12.3	0.0-11.2	0.0-4.7	2.7
	Alhambra Blvd							
11	AM Obs	290	8	0.25	8.7	0.0-7.4	0.1-3.2	1.5
	H Street							
12	AM Obs	616	8	0.25	13.0	0.0-23.1	1.9-8.7	4.6
	Dairy Mart Road							
13	AM Obs	320	8	0.07	8.1	0.0-6.7	0.7-2.0	1.6
14	PM Obs	425	8	0.09	10.9	0.0-11.3	0.5-6.4	2.8

TABLE 3 LIGHT RAIL-RELATED STOPPED DELAY—TRAFFIC SIGNAL CONTROL

Site	Cross Street	Peak Hour Volume (vph)	No. of Peak Hour Trains (tph)*	Crossing v/c Ratio	Stops on Approach (%)	Individual Stopped Delay		
						Range for one cycle (sec/veh)	Range for 15 min period (sec/veh)	Hourly Average (sec/veh)
	Church Street							
15	AM Obs	381	11/9	0.12	34.9	6.7-43.9	12.5-19.0	14.0
16	PM Obs	672	10/9	0.14	31.3	8.0-37.4	18.8-28.8	24.6
	Chippewa Street							
17	AM Obs	793	10/9	0.33	33.7	0.9-39.5	10.6-25.3	15.4
18	PM Obs	504	9/10	0.22	25.6	2.0-32.5	6.5-13.8	9.0
	Church Street							
19	AM Obs	776	11/9	0.17	22.8	0.0-24.8	6.0-9.3	8.0
20	PM Obs	721	9/9	0.22	28.4	2.8-24.7	7.2-12.1	8.3

* Inbound/Outbound

TABLE 4 BACKGROUND CYCLE STOPPED DELAY—TRAFFIC SIGNAL CONTROL

Site	Cross Street	Peak Hour Volume (vph)	No. of Peak Hour Trains (tph)	Crossing v/c Ratio	Stops on Approach (%)	Individual Stopped Delay		
						Range for one cycle (sec/veh)	Range for 15 min period (sec/veh)	Hourly Average (sec/veh)
	Church Street							
15	AM Obs	208	29	0.12	60.6	4.0-25.8	13.2-17.6	14.4
16	PM Obs	378	27	0.24	73.5	3.1-29.1	13.9-19.0	16.2
	Chippewa Street							
17	AM Obs	448	38	0.30	70.8	0.0-29.0	7.6-13.8	10.1
18	PM Obs	326	41	0.23	52.1	0.0-23.0	6.2-7.4	6.8
	Church Street							
19	AM Obs	478	43	0.16	56.3	0.0-22.0	6.3-9.7	7.4
20	PM Obs	432	40	0.16	48.6	1.1-22.7	5.4-8.4	6.7

TABLE 5 COMPOSITE STOPPED DELAY—TRAFFIC SIGNAL CONTROL

Site	Cross Street	Peak Hour Volume (vph)	No. of Peak Hour Trains (tph)	Crossing v/c Ratio	Stops on Approach (%)	Individual Stopped Delay		
						Range for one cycle (sec/veh)	Range for 15 min period (sec/veh)	Hourly Average (sec/veh)
	Church Street							
15	AM Obs	381	20	0.56	68.0	4.0-46.4	18.6-22.5	20.9
16	PM Obs	672	19	0.28	72.6	5.5-37.4	16.5-22.4	19.7
	Chippewa Street							
17	AM Obs	793	19	0.54	73.6	0.0-50.8	14.6-27.5	20.4
18	PM Obs	504	19	0.38	59.3	0.1-32.5	10.9-16.2	13.0
	Church Street							
19	AM Obs	776	20	0.26		0.2-54.3	10.1-14.7	12.7
20	PM Obs	721	18	0.40	57.6	N.A.	N.A.	N.A.

Crossings that are listed twice (Alhambra Boulevard in Table 2 and Church Street in Tables 3–5) indicate observations made over 2 days, and are shown separately for each day.

PERFORMANCE OF DELAY MODELS

Part of the focus of this project was to determine if any of the many existing delay models developed for the analysis of street intersections was suitable for estimating the delays to motor vehicle traffic caused by isolated at-grade LRT crossings. This project did not have as a goal the development of a new delay model of either theoretical or empirical form. Eleven different delay models were applied during the validation process. These models reflect a broad range of the theoretical and empirical aspects of intersection delay. Each modeler's original work contains a detailed discussion of its construction.

Because the analysis of the field data resulted in calculated average individual vehicular stopped delay, the values of models that estimate approach delay also have been adjusted to reflect average stopped delay. As stated above, direct measurement of approach delay was rejected because of the difficulties of videotaping the necessary length of roadway and defining normal, undelayed travel times. We recognize that the relationship between approach delay and stopped delay is not a static factor. However, little guidance can be found in the literature on the dynamic relationship between approach and stopped delay. The magnitude of the adjustment of approach delay is provided by Reilly et al. (*16*) and the 1985 HCM (*13*). They recommend multiplying values for individual average stopped delay by a static factor of 1.3 in order to estimate the individual average approach delay. Sadegh and Radwan (*17*) support this factor in their study of the 1985 HCM delay model. In comparing it with Webster's model (*18*), they note that the first term of each, the uniform delay term, is similar in most respects. Webster's model estimates average individual approach delay, and the 1985 HCM model estimates average individual stopped delay. The primary difference between them is found in the coefficients. The coefficient of the 1985 HCM model is 76 percent of the coefficient of Webster's model. This translates into a factor of approximately 1.32. The adjustment factor used in this effort is

$$\text{Avg. ind. stopped delay} = 0.76 * (\text{avg. ind. approach delay})$$

The delay models included in the validation process are listed below. Except as noted, the delay models included in the validation are taken from Gerlough and Huber (*14*).

- May;
- Allsop, both lower and upper limits;
- Wardrop;
- Webster;
- Allsop's approximation of Webster's model;
- Miller;
- Hutchinson;
- Texas Transportation Institute (*4*), with and without the modified intercept term;
- Stone and Wild (*1*);
- 1985 HCM (*13*), using both the pretimed (A) and actuated (B) progression factors; and
- NCHRP Project 3-28(2) (*19*), both uniform and overflow delay equations.

The May, Allsop, Wardrop, and NCHRP uniform delay models all assume uniform vehicle arrivals in the intersection. Webster, Allsop's approximation of Webster, Miller, Hutchinson, and the HCM model all contain terms to estimate the delay caused by random vehicle arrivals in addition to uniform arrivals. The two Texas Transportation Institute models and the Stone and Wild model are the result of regression analyses correlating the relationship between the *v/c* ratio of an intersection approach and the delay expected on that approach.

In addition to the foregoing models, the following six model fragments were also included:

- Uniform arrival component of the Webster model,
- Random arrival component of the Webster model,
- Uniform arrival component of the HCM model,
- Random arrival component of the HCM model,
- Uniform arrival component of the Hutchinson model, and
- Random arrival component of the Hutchinson model.

These variants comprise the individual components of 3 of the 11 primary models. They were included without regard to the assumptions and boundary conditions associated with their parent models. Two factors influenced the decision to include these variants. First, except for the two Texas Transportation Institute models, the application of the at-grade LRT crossing problem to each of the other nine delay models selected is outside the bounds used when they were validated. Second, it was not known if any of the selected models would be successfully correlated with the field data. Hence, it was felt that the examination of the delay models should include as many as possible. It must be recognized, however, that the random components of the three delay models may be inappropriate because they estimate the delay over and above the uniform delay. Consequently, they should be applicable only in conjunction with their uniform delay components and where the *v/c* ratio approaches one. This is an unlikely case, however, because a LRT crossing, with only two phases, will not, in most cases, have a high *v/c* ratio. For equal approach conditions, the crossing capacity will normally exceed the capacity of up- and downstream signalized intersections. The data reflect this; hence, the models were tested at low *v/c* ratio ranges.

All of the models except the 1985 HCM and NCHRP models estimate the average individual vehicular delay. The 1985 HCM and NCHRP models estimate average individual vehicular stopped delay. The estimates of the other models have been reduced by a factor of 0.76 to account for the difference between approach delay and stopped delay.

MODEL EVALUATION

Using 15-min analysis periods, the data collection effort resulted in four to eight data points per site. Because of the small number of data points, rigorous statistical testing has

not been performed. However, it is still possible to draw general conclusions about the appropriateness of applying each model from the trends it exhibits.

Three indicators have been used to rank the performance of the delay models:

1. The coefficient of determination (R^2) resulting from the observed and predicted values;
2. The mean difference between the observed and predicted values; and
3. The variance of the difference between the observed and predicted values.

Figures 1, 2, and 3 list the five best models for each site as determined by indicators one, two, and three, respectively. The site numbers are defined in Tables 1–5, and the delay model symbols are defined in Table 6. The information in these figures for the two sites controlled by standard traffic signals (Church and Chippewa streets) included delay accrued only when traffic is stopped for light rail vehicle crossings. As mentioned, these signals operated on a background cycle and would stop motor vehicle traffic periodically when no light rail vehicle was crossing, thus creating additional delay. The results using total, or composite, delay are shown in Figures 4, 5, and 6.

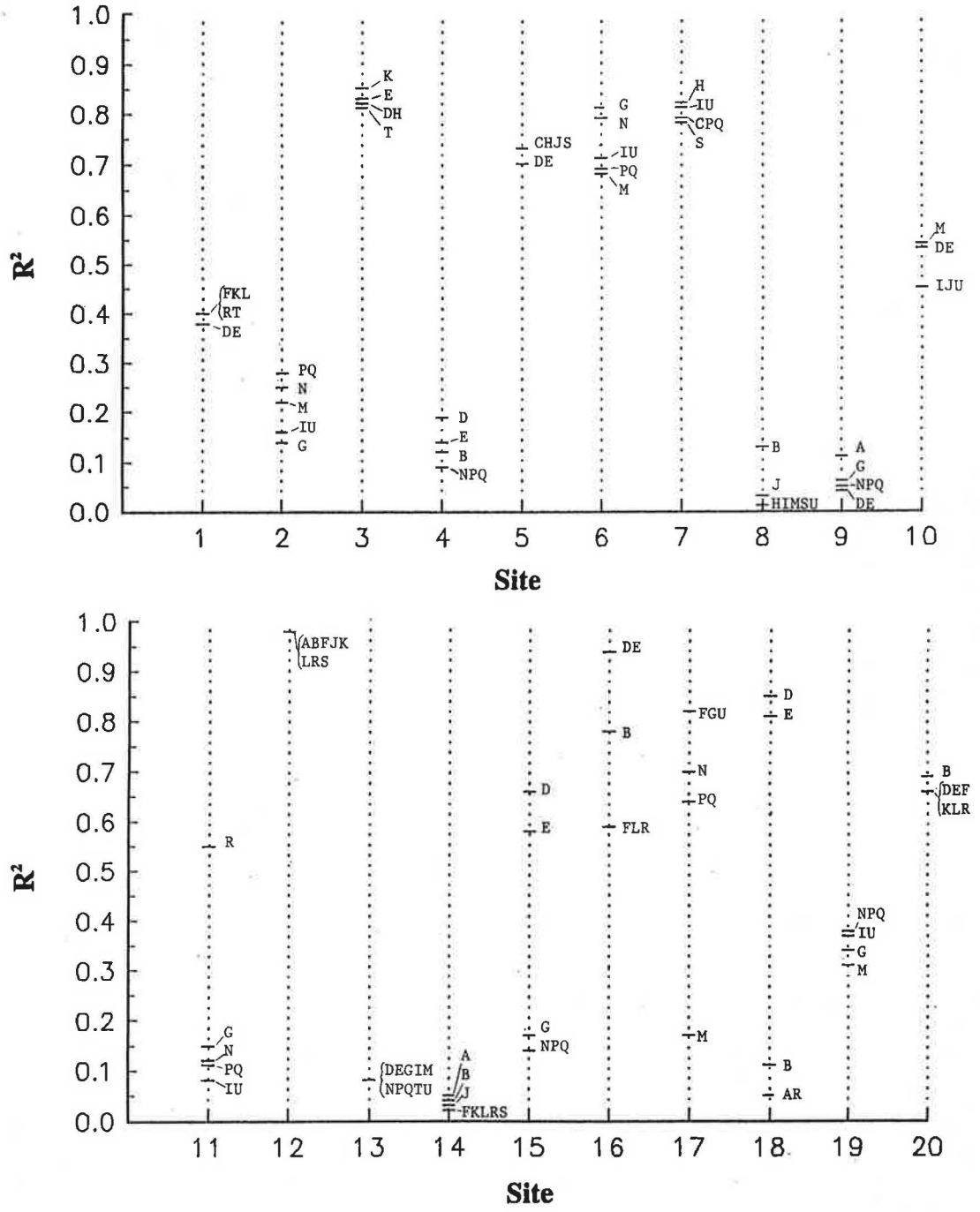


FIGURE 1 Regression analysis.

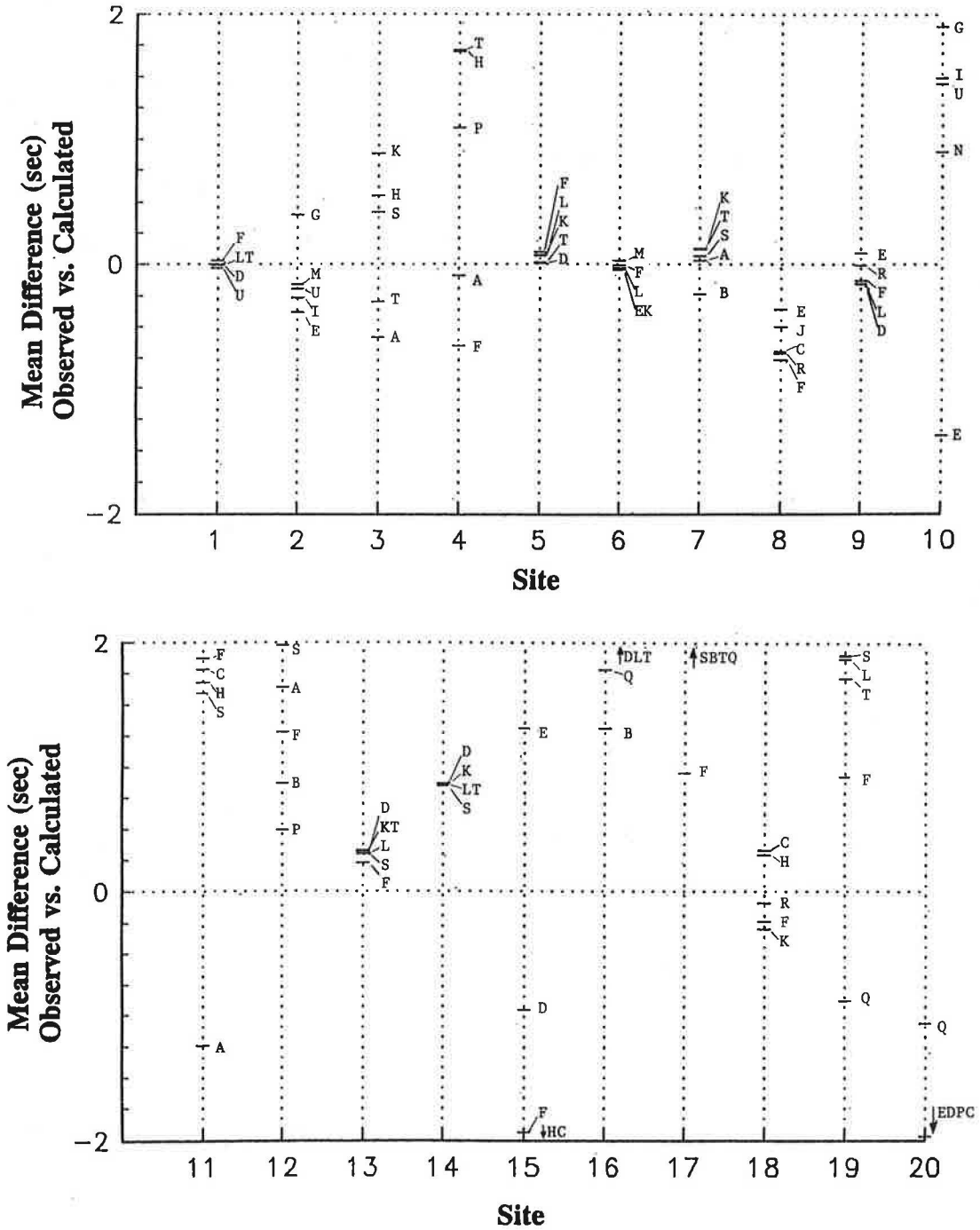


FIGURE 2 Mean difference between observed and calculated delay times.

For the purpose of overall model selection, a ranking procedure was used for each indicator. The best equation was given a score of 22, and the worst equation was given a score of one. Limits were placed on the rankings, and values outside the limits were given a score of zero. For the regression analysis, a value of zero was assigned if the R^2 value was found to be less than 0.50. For the analysis of the mean and standard deviation of the differences, a value of zero was assigned if the mean or standard deviation exceeded 5.0 sec. Five seconds represents the difference between Levels of Service (LOS) A and B and one half of the difference between LOS B and C in the 1985 HCM. It is an appropriate range in light of the low delays and resultant levels of service observed during the field studies.

As evident in Table 6, a wide range of R^2 values results from the regression analysis. The morning observation at H Street on the SDTI South Line has the best series of R^2 values; the evening observation at 65th Street on the SRTD Butter-

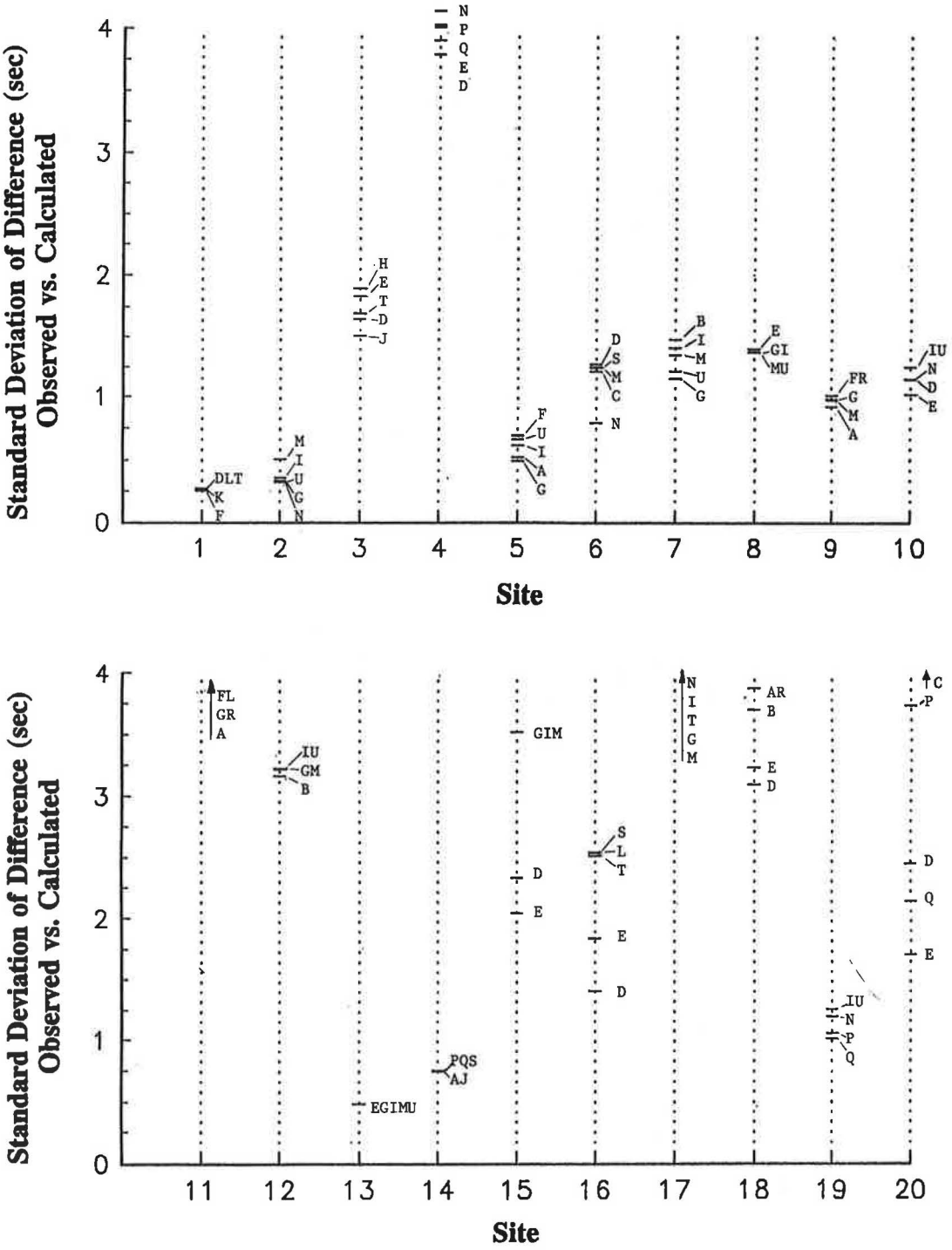


FIGURE 3 Standard deviation of differences between observed and calculated delay.

field Line has the worst series of R^2 values. Overall the actuated HCM equation (HCM B) has the best ranking among the delay models, with the pretimed HCM equation (HCM A) a close second. Note also that the regression based models did not appear to perform better than the theoretical models.

Similar results occurred for the regression analysis of the composite traffic signal operations. The primary difference between the general application and the composite application is that the best model, the actuated HCM equation (HCM B), was not a clear leader. The NCHRP Overflow delay model was within one point of the HCM equation. The remaining seven models with scores are all grouped together without clear preference among them.

Examination of the mean difference between the observed and calculated delay values shows that among the five best models, there is little variation, usually less than 1 sec. The

TABLE 6 MODEL CODES USED IN FIGURES 1–6

Model Code	Model	
A	Allsop	lower limit
B		upper limit
C		approx. of Webster's model
D	HCM	pretimed coefficients
E		actuated coefficients
F		uniform delay component
G		random delay component
H	Hutchinson	
I		random delay component
J	Miller	A (v/c < 0.5)
K		B (v/c > 0.5)
L	NCHRP	uniform delay
M		overflow delay
N	Stone and Wild	
P	TTI	modified coefficient
		unmodified coefficient
Q	Wardrop	
R	Webster	
S		uniform delay component
T		random delay component

Note: May's model and the uniform arrival component of Hutchinson's model are identical to the uniform arrival component of the Highway Capacity Manual delay model.

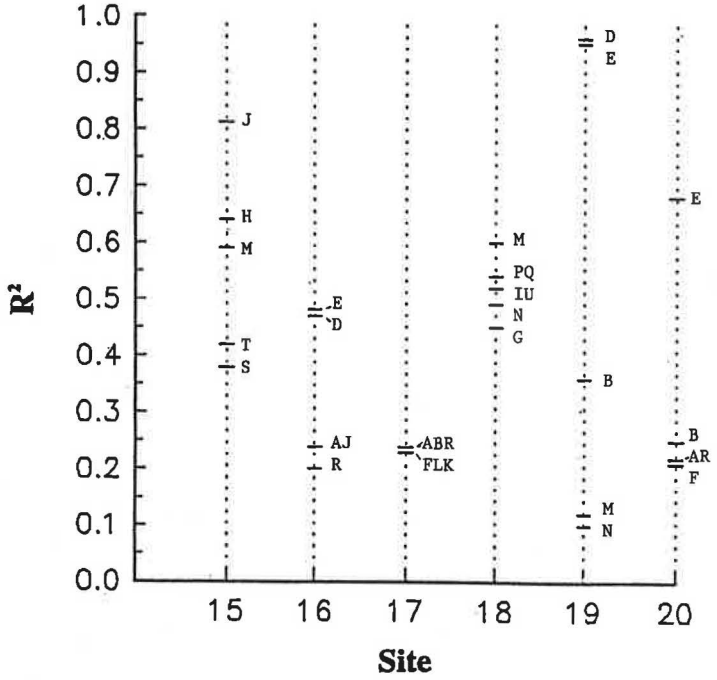


FIGURE 4 Regression analysis of sites with traffic signal control—composite analysis.

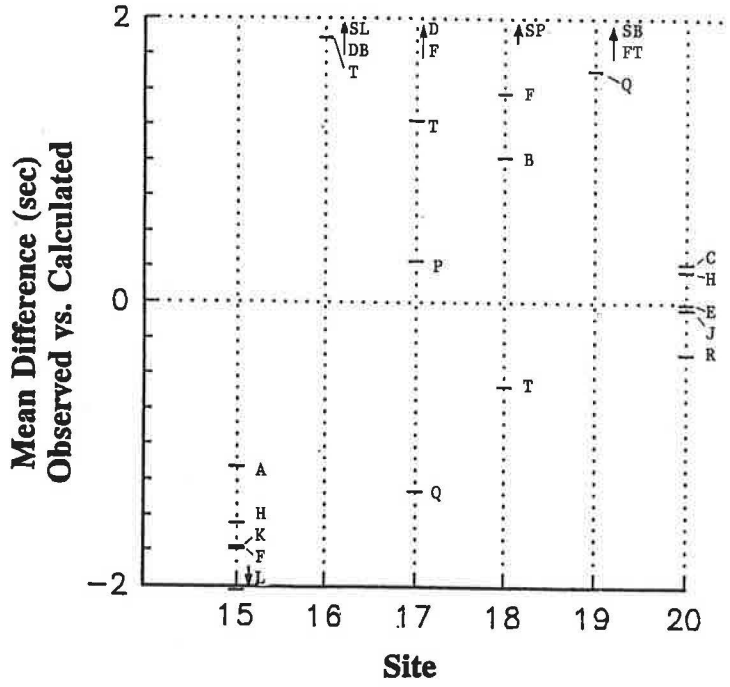


FIGURE 5 Mean difference between observed and calculated delay times for sites with traffic signal control—composite analysis.

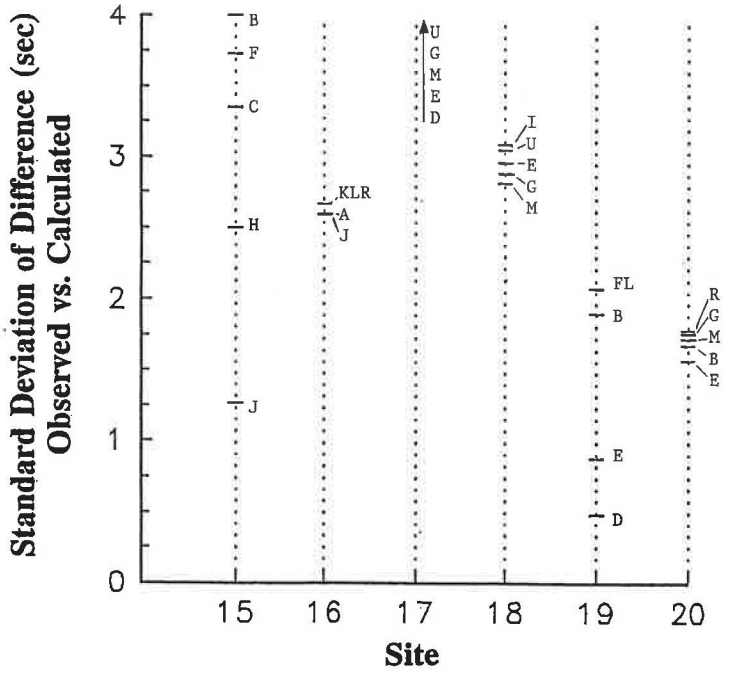


FIGURE 6 Standard deviation of differences between observed and calculated delay for sites with traffic signal control—composite analysis.

difference between the best model and the worst model is greatest at Church Street (No. 2) on the Transit Mall in Buffalo, New York, during the evening peak period, at 5.79 sec. Overall, the best model was the uniform delay component of Webster, with the pretimed HCM equation (HCM A) a close second. Again, the regression-based models did not perform as well as the theoretical models in the rankings, and, again, the results of the composite analysis were similar.

The importance of the mean difference is illustrated when estimating the level of service of the crossing. If the mean difference for the delay equation being used is great, the resulting level of service estimate may be in error. For example, if the mean difference for the HCM pretimed equation is 10 sec and the value guiding the estimate is 15 sec of stopped delay, then the true level of service is between LOS A (15 sec − 10 sec = 5 sec) and LOS C (15 sec + 10 sec = 25 sec).

The standard deviation of the differences is important in estimating possible variation and its influence on the resultant level of service estimate. Examination of the standard deviations of the differences between the observed and calculated delay values shows that among the five best models, there is, again, little variation. The difference between the best model and the worst model is greatest at Church Street (No. 2) during the evening peak period, at 2.44 sec. Overall, the actuated HCM equation (HCM B) again has the best ranking among the delay models, with the pretimed HCM equation (HCM A) again a close second, and, once again, the regression-based models did not perform better than the theoretical models. The results of the composite analysis were similar to the general application.

CONCLUSIONS

This study evaluated the application of theoretically and empirically based delay equations, developed for isolated traffic signal operation, to the problem of isolated at-grade LRT crossings. The key finding of the study is that the pretimed HCM equation is a suitable model for isolated LRT crossing evaluation.

At crossings where control is provided by traffic signals operating with background cycles, Allsop's equation was found to be a suitable model when applied with its lower limit parameters.

In addition to these findings, it was found that

- The results from the exclusive use of the random delay component of multiple component delay models were not better than those from use of the complete delay model, or only the uniform component of the model, and
- The delay models based on regression analysis did not perform better than the theoretical delay models.

REFERENCES

1. T. J. Stone and W. A. Wild. Design Considerations for LRT in Existing Medians: Developing Warrants for Priority Treatments. In *Special Report 195: Light Rail Transit: Planning, Design, and Implementation*. TRB, National Research Council, Washington, D.C., 1982.
2. W. Grote. *Impacts of Light Rail Transit Operations on Urban Intersection Delay and Capacity*. Master's thesis. University of Utah, 1984.
3. A. E. Radwan and K. P. Hwang. Preferential Control Warrants of Light Rail Transit Movements. In *Transportation Research Record 1010*, TRB, National Research Council, Washington, D.C., 1985.
4. J. C. Cline, T. Urbanik, and B. Rymer. Delay at Light Rail Transit Grade Crossings. In *Texas Transportation Research Report 339-10*, Texas State Department of Highways and Public Transportation, 1987.
5. R. A. Berry. Assessing the Operation of At-Grade Light Rail Transit Crossings. Presented at Traffic Congestion 88: Issues and Answers, ITE, 1988.
6. *Special Report 87: Highway Capacity Manual*. HRB, National Research Council, Washington, D.C., 1965.
7. A. D. May, Jr. and D. Pratt. A Simulation Study of Load Factor at Signalized Intersections. *Traffic Engineering*, 1968.
8. R. W. Crommelin. Employing Intersection Capacity Utilization Values to Estimate Overall Level of Service. *Traffic Engineering*, 1974.
9. San Diego Association of Governments. *San Diego Trolley: The First Three Years*. Office of Planning Assistance, UMTA, U.S. Department of Transportation, 1984.
10. P. A. Gibson, B. B. Lin, and R. Robenhymer. Traffic Impacts of Light Rail Transit. In *Special Report 195: Light Rail Transit: Planning, Design, and Implementation*. TRB, National Research Council, Washington, D.C., 1982.
11. Gannett-Fleming/Schimpeler Corradino. Traffic and Circulation Technical Report. *Bayside LRT Line Preliminary Engineering and Environmental Impact Report Study*, Metropolitan Transit Development Board, San Diego, Calif., 1987.
12. *Transportation Research Circular 212: Interim Materials on Highway Capacity*. TRB, National Research Council, Washington, D.C., 1980.
13. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
14. D. L. Gerlough and M. J. Huber. *Special Report 165: Traffic Flow Theory—A Monograph*. TRB, National Research Council, Washington, D.C., 1975.
15. R. A. Berry and J. C. Williams. Traffic Characteristics of At-Grade Light Rail Crossings. *Compendium of Technical Papers*, 59th Annual Meeting of the ITE, 1989.
16. W. R. Reilly, C. C. Gardner, and J. H. Kell. *A Technique for Measurement of Delay at Intersections*. Report RD-76-135/137, FHWA, U.S. Department of Transportation, 1976.
17. A. Sadegh and A. E. Radwan. Comparative Assessment of 1985 HCM Delay Model. *Journal of Transportation Engineering*, ASCE, Vol. 114, No. 2, 1988.
18. F. V. Webster. *Traffic Signal Settings*. Road Research Technical Paper 39, Great Britain Road Research Laboratory, 1958.
19. JHK & Associates. NCHRP Signalized Intersection Capacity Method. NCHRP Project 3-28(2), TRB, National Research Council, Washington, D.C., 1982.

Publication of this paper sponsored by Committee on Traffic Control Devices.

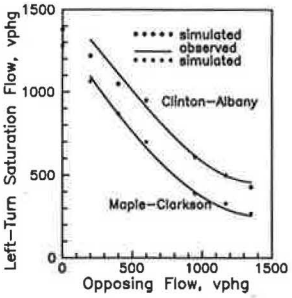


FIGURE 2 Observed and simulated saturation flows of opposed left turns at two intersections.

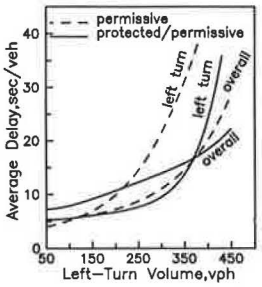


FIGURE 3 Example characteristics of full-actuated signal operations involving permissive phasing and protected/permissive phasing.

which concern opposed left turns from a shared lane, further show that the model is a reasonable tool for analysis.

The saturation flows shown in Figure 2 are similar to those reported in an earlier study (*8*). Nevertheless, they reveal that the saturation flows of opposed left turns can vary significantly from one intersection to another. The differences between the saturation flows as shown in Figure 2 are primarily attributable to the fact that the Clinton-Albany intersection has a much larger storage area than the Maple-Clarkson intersection. The Clinton-Albany intersection allows an average of 2.7 vehicles to complete the turns after a signal change interval begins, compared with an average of only 1.3 vehicles at the Maple-Clarkson intersection (*9*).

The simulation analysis performed in this study is based on the following conditions:

- Left-turn drivers have a critical gap of 5 sec.
- There is a 15 percent chance that the first left-turn vehicle in a queue will turn in front of the first opposing vehicle immediately after the green light is turned on.
- Vehicles approach the intersection randomly at an average speed of 30 mph. All vehicles are passenger cars.
- The saturation flows of straight-through, unopposed left-turn, and right-turn movements are 1,700, 1,500, and 1,350 vphg, respectively.

When the opposing volume is very heavy, left turns can be made only in the first few seconds after the green onset or after the change interval begins. The simulated number of such turns varies from one cycle to another; the average is approximately two vehicles per cycle.

PHASING PLAN SELECTION

Figure 3 gives an insight into the relative performance characteristics of a control with permissive left turns and another with protected/permissive left turns. This comparison is based on an opposing flow of 500 vph and a cross-traffic pattern that has a critical lane flow of 500 vph. The figure shows that permissive left turns can bring about shorter left-turn delays and overall delays when the left-turn volume is small. As the left-turn volume increases, protected/permissive left-turn phasing can easily provide a better service to the left-turn vehicles, although permissive left-turn phasing may still yield shorter overall delays. When the left-turn volume is sufficiently high, it becomes possible for protected/permissive left-turn phasing to reduce not only the left-turn delays but also the overall delays. It is also apparent from the figure that delays are more sensitive to the left-turn volume when permissive left-turn phasing is used. In contrast, protected/permissive left-turn phasing can provide more equitable services to the vehicles in every lane. As a result, the delays resulting from such a phasing arrangement are less sensitive to the left-turn volume. In terms of overall delays, the control efficiencies may deteriorate by more than 30 percent when an improper phasing arrangement is implemented.

The choice between permissive and protected/permissive phasings naturally requires a trade-off between maximizing overall control efficiency and avoiding excessive delays and queue lengths in any traffic lane. What constitutes unacceptable delays and queue lengths can vary from one signal installation to another. Field observations made at the Market Street–Sandstone intersection in Potsdam, New York, indicate that an opposed left-turn flow of 230 vph suffering an average stopped delay of more than 40 sec per vehicle often incurs a queue length in excess of 400 ft. In principle, the queue length in any lane should not be allowed to extend to the upstream intersection. For most intersections, this perhaps implies a maximum allowable queue length of about 500 to 700 ft. Whenever possible, stopped delays should also be kept under 40 sec per vehicle to prevent the level of service from degenerating into a Level E or, possibly, a Level F as defined in the 1985 *Highway Capacity Manual* (*10*).

In this study, a phasing plan is considered superior to a competing plan if the following conditions are satisfied:

1. The stopped delay in every lane is less than 40 sec per vehicle;
2. The queue length in every lane is less than 500 ft; and
3. The overall delay is smaller than that produced by the competing plan.

Under very heavy flow conditions, neither permissive phasing nor protected/permissive phasing may be able to maintain acceptable delays and queue lengths. In such a case, the phasing arrangement that can provide more equitable services by

eliminating extremely long delays and queue lengths is considered to be better.

Based on these criteria, the two options of left-turn phasing are compared under a variety of conditions. In this comparison, the signal timing settings for a given condition are adjusted to achieve near-optimal signal operations for both permissive phasing and protected/permissive phasing. The delays and queue lengths derived from computer simulation for such operations are then used to determine the preferred phasing strategy. The stopped delay of each simulated vehicle is the total stopped delay accumulated from the moment the vehicle is generated at a location 600 ft upstream of the stop line until it clears the intersection.

The results of this simulation analysis are shown in Figure 4 for cases involving one opposing lane and in Figure 5 for those involving two opposing lanes. In these figures, the combinations of left-turn volume and opposing volume that allow permissive phasing and protected/permissive phasing to achieve comparable signal operations are represented by one curve for a specified level of cross traffic. The combinations of left-turn volume and opposing volume above such a curve represent conditions favorable to the implementation of protected/permissive phasing; those combinations below the curve should preferably be treated with permissive left-turn phasing.

In Figures 4 and 5, low, moderate, and heavy cross-traffic levels are represented by critical movement volumes, denoted as Q_c, of 100, 600, and 900 vph, respectively. In fact, such volumes alone really cannot give an accurate picture of the impact of the cross traffic. From the viewpoint of left-turn phasing, the most important element related to the cross traffic is the amount of green time consumed by the cross traffic, not the critical movement volume. It should be noted that

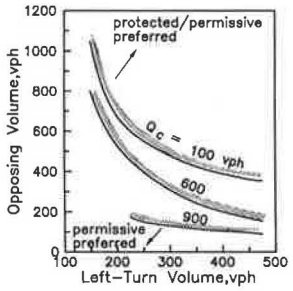


FIGURE 4 Preferred left-turn phasings (one opposing lane).

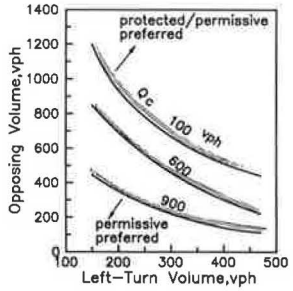


FIGURE 5 Preferred left-turn phasings (two opposing lanes).

the average cross-street green times related to the above critical volumes fall into the following ranges when permissive phasing is in place: 6 to 12 sec for Q_c = 100 vph; 18 to 25 sec for Q_c = 600 vph; and 30 to 45 sec for Q_c = 900 vph.

Figures 4 and 5 show that the choice between permissive phasing and protected/permissive phasing can be significantly affected by left-turn volume, opposing volume, number of opposing lanes, and the level of cross traffic. The effective length of the left-turn bay has less obvious impact on such a choice, probably because both permissive left turns and protected/permissive left turns can be adversely affected by a left-turn bay of insufficient length. Nevertheless, as shown in the next section, providing sufficiently long left-turn bays can greatly enhance the ability of protected/permissive phasing to improve signal operations.

Within the levels of the cross traffic considered in this study, Figures 4 and 5 show that the provision of a separate left-turn phase is difficult to justify when the left-turn volume is less than 140 vph (an equivalent of not more than 15 sec of green time for the cross street). As the left-turn volume increases, it becomes easier for protected/permissive phasing to produce better signal operations. It is also obvious from these figures that the use of the product of left-turn volume and opposing volume (*2*) to guide phasing decisions is not suitable for full-actuated control. Direct application of Figures 4 and 5 would allow more intelligent decisions. It should be noted, however, that these figures are derived from comparing signal operations that have good signal timing settings. How improper timing settings may affect the relationships presented in the figures has not been investigated.

A hidden feature of Figures 4 and 5 should also be pointed out. Generally, protected/permissive phasing would be most effective in improving signal operations when the capacity of an intersection cannot adequately accommodate permissive left turns but is sufficient to accommodate protected/permissive left turns. The capacity of an intersection is considered to be inadequate if near-optimal timing settings cannot prevent stopped delays and queue length in every lane from exceeding 40 sec per vehicle and 500 ft, respectively. Under very heavy flow conditions both permissive phasing and protected/permissive phasing may be unable to maintain acceptable signal operations. In such a case, protected/permissive phasing may only be able to mitigate the severity of congestions by redistributing overcongestions among several lanes. Under this circumstance, protected/permissive phasing may not have a clear-cut advantage over permissive phasing. Consequently, the phasing selection should be made more cautiously. To assist in the selection of phasing plans, the capacity constraints imposed on permissive phasing and protected/permissive phasing are shown in Figures 6 and 7. These figures are applicable to intersections where full-length exclusive left-turn lanes are available; they are also applicable to intersections where left-turn bays are not blocked.

LENGTH OF LEFT-TURN BAY

If the length of a left-turn bay is not long enough, the vehicles using the bay and its adjacent lane may block each other. When such an operating condition exists, the effectiveness of protected/permissive phasing in improving signal operations

Left-Turn Signal Phasing for Full-Actuated Signal Control

FENG-BOR LIN

Many warrants and guidelines exist concerning left-turn phasing at signalized intersections. However, there is still a lack of understanding about the left-turn phasing requirements for full-actuated signal control. Using computer simulation, a knowledge base is developed to assist in the choice between permissive phasing and protected/permissive phasing for this type of signal control. The important factors concerning such a choice include left-turn volume, opposing volume, the number of opposing lanes, length of left-turn bay, and the volume of cross traffic. The collective impact of these factors on left-turn phasing cannot be adequately assessed through the use of simple rules of thumb. The lengths of left-turn bays that allow protected/permissive phasing to be effectively used are also identified.

Left-turn movements are a major source of traffic conflicts at signalized intersections. The existence of such movements aggravates traffic delays and safety problems; it also complicates the selection of left-turn phasing plans for the optimization of signal operations. Many of factors have been used as criteria for left-turn signal phasing. Accident experiences, left-turn and opposing traffic volumes, delays, gap acceptance, traffic conflicts, and intersection capacity are some examples of such factors. Agent and Dean (*1*) presented a very informative review of the application of these various factors in developing warrants for left-turn phasing.

Several guidelines for left-turn phasing are set forth in the *Traffic Control Devices Handbook* (*2*). In terms of traffic volume, these guidelines suggest that separate left-turn phasing be considered when the product of left-turn and opposing volumes during peak hours exceeds 100,000 on a four-lane street or 50,000 on a two-lane street, provided that the left-turn volume is more than two vehicles per cycle during the peak-hour period. In terms of delay, the guidelines suggest that separate left-turn phasing be considered when the following conditions are met: left-turn delay is at least 2.0 vehicle-hours in a peak hour on a critical approach; left-turn volume is greater than two per cycle during the peak hour; and average delay per left-turning vehicle is more than 35 sec. More recently, several researchers (*3–5*) proposed additional warrants and guidelines for left-turn phasing.

Despite the existence of a number of guidelines, there is still a lack of understanding about the left-turn phasing requirements for full-actuated signal control (*6*). Full-actuated control is a primary means for isolated control of individual intersections. The performance characteristics of this type of control are governed by timing settings, detector configuration, phasing arrangement, geometric design of intersection, and prevailing traffic conditions. Changing the phasing arrangement for the traffic on one street can create a chain reaction in the operation of every phase. The dynamic nature of full-actuated signal operations makes the selection of proper phasing arrangements difficult.

This study determines, for full-actuated control that relies on long inductive loop detectors for presence detection of vehicles, how left-turn phasing should be selected to make the signal operations as efficient as possible. The analysis is based on data derived from a microscopic simulation model. The simulation analysis concerns the choice between permissive left-turn phasing and protected/permissive left-turn phasing.

Department of Civil and Environmental Engineering, Damon Hall, Clarkson University, Potsdam, N.Y. 13676.

SCOPE OF STUDY

From the perspective of the efficiency of signal operations, a number of geometric design, traffic, and signal timing factors can affect the phasing decisions for left-turn movements. The geometric design and traffic factors considered in this study are depicted in Figure 1. These factors include the effective length (L) of left-turn bay; left-turn volume (Q_L); straight-through volume (Q_s) in the lane adjacent to the left-turn bay; opposing volume (Q_o) and the number of opposing lanes; and the cross-traffic volume, such as Q_{c1} and Q_{c2}.

The effective length of a left-turn bay refers to the length within which stopped vehicles will not block the vehicular movements in the adjacent lane. The minimum length of a left-turn bay is assumed to be 50 ft. In the absence of a left-turn bay, the left lane of an intersection approach is assumed to be for the exclusive use of the left-turn vehicles.

The opposing volume (Q_o) is a primary factor affecting the need for separate left-turn phasing. The impact of this volume depends on the number of opposing lanes involved. An opposing volume concentrated in one lane has a more severe detrimental impact than when the same volume is distributed over several lanes. This study considers only the left-turn movements that are faced with either one or two opposing lanes.

The cross traffic influences the amount of green time available to the left-turn vehicles. This available green time in turn affects the delays and congestions associated with a signal operation. Because many simulation runs are needed to analyze a specific combination of the factors involved, it is impractical to examine the impact of the cross traffic by allowing the traffic volume in each cross-street lane to vary independently. As an alternative, fixed cross-traffic patterns, representing low, moderate, and heavy traffic movements on the cross street, are used for the analysis.

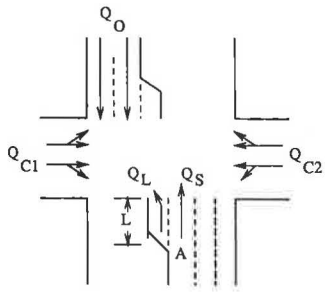


FIGURE 1 Major geometric and traffic factors affecting opposed left-turn movements.

Regarding signal timing, the analysis performed in this study allows each traffic lane to have a 65-ft detector loop. This loop length is reduced for a left-turn bay that has an effective length of less than 65 ft. The minimum green of each phase is set at 0 sec, while the vehicle interval and the maximum green are varied according to the flow pattern being analyzed.

By analyzing the interactions among the various governing factors and the resulting signal operations, this study attempts to establish guidelines concerning the choice between permissive phasing and protected/permissive phasing. This study also examines how protected/permissive phasing can be made more effective through the choice of proper effective lengths for left-turn bays.

When a left-turn bay exists, the adjacent approach lane can be used by both left-turn and straight-through vehicles. The queues formed by these vehicles may extend upstream of the diverging point denoted as A in Figure 1. In such a case, other arriving vehicles will be blocked even if unused storage spaces are downstream of the diverging point. Whatever the left-turn phasing arrangement, this blockage of traffic can greatly reduce the capacity of an intersection and lead to frequent lane changes. The severity of the impact of such blockage depends in part on the left-turn volume (Q_L) and its interacting straight-through volume (Q_s). This study establishes a knowledge base to facilitate the selection of effective bay length in support of the implementation of protected/permissive left-turn phasing.

TOOL FOR ANALYSIS

The simulation model used in this study was developed at Clarkson University. This microscopic model simulates the signal operations at isolated intersections. The model has two major components. One component is a flow processor that generates vehicles and moves them downstream through the intersection according to the prevailing flow and signal control conditions. The other component is a signal processor that is essentially a collection of various signal control logics.

In the flow processor, the location and the speed of each simulated vehicle are updated once per second. Each simulated vehicle is probabilistically assigned a set of attributes related to vehicle length, maximum desired speed, directional movement, desired space headway from the vehicle ahead in a stationary queue, desired stopped location with respect to the stop line, driver reaction time, and driver sensitivity in a car-following situation. The model simulates the traffic movements at intersections where the following features may or may not exist: right turns on red, auxiliary turning bay, mixed directional movements from a given lane, and opposed left turns.

The signal processor determines when the signal indications should be changed, according to the control logic being analyzed. For traffic-actuated signal operations, this processor can accept inputs from a variety of detectors. Each traffic lane may have a combination of several motion detectors and presence detectors. Such detectors may have call-delay or call-extension features.

The vehicular movements as simulated by the Clarkson model agree reasonably well with the observed characteristics related to right turns on red, opposed left turns, and queue dissipation. Because the model is developed for the purpose of comparative analysis of alternative signal controls, special care has been taken to realistically duplicate the interactions between vehicles and detectors. Some aspects of the model output are described elsewhere (*7*).

The model has been tested in terms of its ability to provide accurate estimates concerning the operations of traffic-actuated signals. The data used in this test were related to six hourly flow patterns observed at four intersections. The observed and simulated values of average cycle lengths, average green intervals, and the average delays in certain lanes are shown in Table 1. The largest difference between the observed and the simulated average greens is 1.9 sec. The simulated stopped delays deviate from the observed values by no more than 1.4 sec per vehicle.

Because opposed left-turn movements are the focus of this study, it is especially important that the simulation model realistically represents the interactions between the left-turn vehicles and their opposing flow. In this regard, simulated and observed saturation flows of the opposed left turns at two intersections were compared. The results of this comparison are shown in Figure 2. Data related to Case E of Table 1,

TABLE 1 OBSERVED CHARACTERISTICS OF SIX SIGNAL OPERATIONS AND ESTIMATES OBTAINED FROM CLARKSON MODEL

		Average Green, sec				Average Stopped Delay, sec/vh	
		Observed		Simulated			
Case (1)	Phase (2)	Mean (3)	S.D.[a] (4)	Mean (5)	S.D. (6)	Observed (7)	Simulated (8)
A	1	6.5	2.6	5.1	2.1	7.7[b]	6.6
	2	32.4	25.2	30.3	24.6		
B	1	33.8	18.2	31.9	17.6	11.1[c]	10.7
	2	5.4	2.1	5.1	2.3		
	3	24.0	0.0	24.0	0.0		
C	1	27.8	12.8	27.6	12.1	14.7[c]	14.9
	2	12.4	6.3	13.0	6.3		
D	1	18.8	9.2	17.5	8.1	not available	
	2	10.7	5.8	9.5	5.2		
E	1	29.3	7.6	30.2	9.1	42.3[d]	43.7
	2	20.5	0.7	20.2	0.9		
F	1	12.9	3.9	12.6	3.5	14.0[c]	13.5
	2	9.2	4.1	9.4	4.0		
	3	33.6	7.3	32.1	6.7	3.1[e]	2.7

[a]S.D. = Standard Deviation
[b]single-lane flow with right turns and left turns
[c]exclusive left-turn flow
[d]shared-permissive left-turn flow (85% left turns)
[e]exclusive right-turn flow with right-turn-on-red

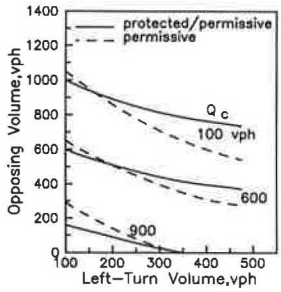


FIGURE 6 Maximum allowable combinations of left-turn volume and opposing volume to avoid unacceptable operations due to capacity constraint (one opposing lane).

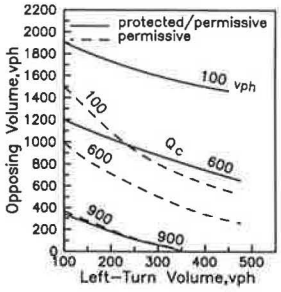


FIGURE 7 Maximum allowable combinations of left-turn volume and opposed volume to avoid unacceptable operations due to capacity constraint (two opposing lanes).

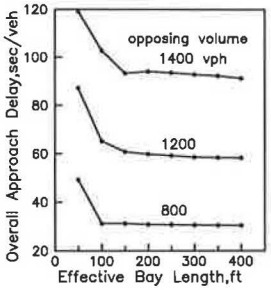


FIGURE 8 Example relationships between delay and the effective length of left-turn bay.

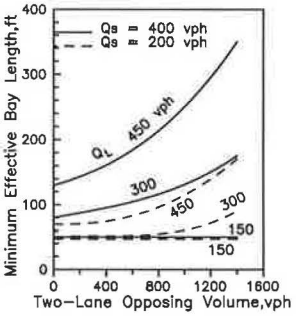


FIGURE 9 Minimum required effective length of left-turn bay (low cross traffic with critical movement volume Q_c = 100 vph).

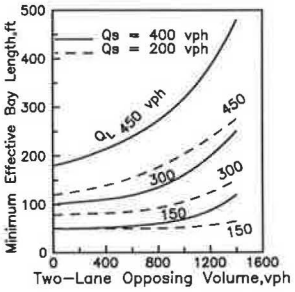


FIGURE 10 Minimum required effective length of left-turn bay (moderate cross traffic with critical movement volume Q_c = 600 vph).

can become rather limited. An example of the detrimental impact of inadequate bay lengths is shown in Figure 8. This figure is developed in two stages. First, for a bay length of 100 ft, the timing settings are adjusted to minimize the overall delay and to avoid, whenever possible, excessive delays and queue lengths. Next, the resulting timing settings are held constant while the bay length is varied. Thus, the delay curves shown in the figure are a function of the bay length alone.

Figure 8 shows that the overall delay can increase dramatically when the bay length is shorter than the critical length. For the flow pattern shown in the figure, the critical length is about 100 ft when the left-turn vehicles are faced with a two-lane opposing volume of 800 vph. This critical length is raised to about 150 ft when the opposing volume is increased to 1,200 vph. Beyond such critical lengths, a very large increase in the bay length is needed in order to produce a noticeable improvement in the control efficiency.

Based on the critical bay lengths for a variety of flow conditions, Figures 9, 10, and 11 assist in the determination of the minimum bay length requirements. These figures can be used directly when left turns encounter two opposing lanes. To apply them to cases involving only one opposing lane, the one-lane opposing volume must first be transformed into an equivalent two-lane opposing volume. Through comparison of simulated bay length requirements, it is found that 1 vehicle in a one-lane opposing flow can be transformed into about 1.3 vehicles in a two-lane opposing flow. This conversion factor approximates the ratio of typical observed left-turn saturation flow facing two opposing lanes to that facing one opposing lane (*8*). The following example illustrates the applications of Figures 9, 10, and 11.

Given a left-turn volume of Q_L = 350 vph, a two-lane opposing volume of Q_o = 700 vph, and adjacent straight-through flow of Q_s = 300 vph, and a critical movement vol-

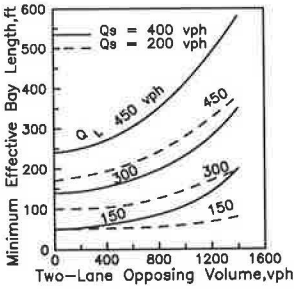


FIGURE 11 Minimum required effective length of left-turn bay (high cross traffic with critical movement volume Q_c = 900 vph).

ume of Q_c = 350 vph on the cross street, it is desired to know how long the minimum length of the left-turn bay should be. Because the critical movement volume on the cross street is 350 vph, Figures 9 and 10 should be used. Through interpolation, Figure 9 gives an estimated required length of 100 ft for a critical movement volume of 100 vph on the cross street. Similarly, Figure 10 gives an estimated required length of 140 ft for a critical movement volume of 600 vph on the cross street. Thus, the required length is approximately the average of 100 ft and 140 ft, that is, 120 ft.

For the conditions given in this example, Figure 5 shows that protected/permissive phasing may be a better choice when the opposing volume exceeds approximately 450 vph. Because the given opposing volume of 700 vph is much larger, protected/permissive phasing warrants serious consideration. If left-turn bays of at least 120 ft in length are provided, protected/permissive phasing can be effectively used. In such a case, Figure 7 shows that the intersection capacity will unlikely become inadequate unless the opposing volume exceeds about 1,100 vph. Because the given opposing volume is only 700 vph, there are flexibilities in using protected/permissive phasing to improve signal operations. In contrast, Figure 7 also shows that the intersection capacity will be inadequate if permissive phasing is implemented instead. With this additional understanding, it appears reasonable to conclude that protected/permissive phasing should be chosen over permissive phasing.

CONCLUSIONS

The selection of left-turn phasing plans for full-actuated signal control is a complicated problem. Ideally, such a problem should be solved with the aid of computer simulation. When an easy access to a simulation model is not available, the information presented here is useful, particularly for planning purposes.

The choice between permissive phasing and protected/permissive phasing is governed primarily by left-turn volume, opposing volume, the number of opposing lanes, and the level of cross traffic. Simple rules of thumb are not adequate in guiding such a choice. Protected/permissive phasing is generally preferred to permissive phasing if the intersection capacity cannot accommodate signal operations with permissive phasing but is still adequate to support operations with protected/permissive phasing.

In implementing protected/permissive phasing, the left-turn bays should be sufficiently long. Otherwise, the ability of such a phasing arrangement to improve signal operations can be seriously compromised.

REFERENCES

1. K. R. Agent and R. C. Deen. Warrants for Left-Turn Signal Phasing. *Transportation Research Record 737*, TRB, National Research Council, Washington, D.C., 1979, pp. 1–10.
2. *Traffic Control Devices Handbook*. FHWA, U.S. Department of Transportation, 1983.
3. N. M. Rouphail. Analytical Warrant for Separate Left-Turn Phasing. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 20–24.
4. J. E. Upchurch. Guidelines for Selecting Type of Left-Turn Phasing. In *Transportation Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 30–38.
5. B. H. Cottrell, Jr. Guidelines for Protected/Permissive Left-Turn Signal Phasing. In *Transportation Research Research Record 1069*, TRB, National Research Council, Washington, D.C., 1986, pp. 54–61.
6. *Traffic Control Systems Handbook*. Report FHWA-IP-85-11. FHWA, U.S. Department of Transportation, 1985.
7. F. B. Lin. Estimating Average Cycle Lengths and Green Intervals of Semiactuated Signal Operation for Level-of-Service Analysis. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
8. P. G. Michalopoulos, J. O'Connor, and S. M. Nova. Estimating of Left-Turn Saturation Flows. In *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1978, pp. 35–41.
9. F. B. Lin and T. T. Nadratowski. Estimation of Left-Turn Traffic Parameters. *Journal of Transportation Engineering*, ASCE, Vol. 109, No. 3, May 1983, pp. 347–362.
10. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Post-Mounted Delineators and Raised Pavement Markers: Their Effect on Vehicle Operations at Horizontal Curves on Two-Lane Rural Highways

RAYMOND A. KRAMMES AND KEVIN D. TYER

Post-mounted delineators (PMDs) and retroreflective raised pavement markers (RPMs), either individually or in combination, have been recommended in previous research as supplemental delineation treatments at horizontal curves on two-lane rural highways. However, these recommendations have been based on limited amounts of operational data or accident models that show little correlation between accident rates and the type of delineation treatment. No attention has been paid to the short-term effects of changing from one delineation treatment to another or to the long-term operational effectiveness of the treatments. To evaluate how vehicle operations changed, existing PMDs were removed and replaced with RPMs supplementing the existing painted centerline at five horizontal curves on two-lane rural highways in Texas. Vehicle operations were monitored on the outside lane of the curves first with the existing PMDs in place and then with the RPMs after 1 day (short-term), 6 to 11 weeks (intermediate-term), and 11 months (long-term). Operational measures of effectiveness that have been suggested by previous research to be correlated to accident experience were evaluated, including the speed at the midpoint of the curve, speed change from the beginning to midpoint of the curve, lateral placement at the midpoint of the curve, and number of vehicle encroachments into the opposing lane at the midpoint of the curve. Vehicle operations with the RPMs compared favorably with the existing PMDs in both short-term and intermediate-term evaluations. The long-term evaluation at one curve indicated that the RPMs, which had lost most of their reflectivity, continued to provide adequate near delineation, but that their far delineation was somewhat degraded.

Post-mounted delineators (PMDs) are commonly used as a supplement to standard pavement markings at horizontal curves on two-lane rural highways. Maintenance problems associated with PMDs have proven to be a nuisance. As a result, the Texas State Department of Highways and Public Transportation (SDHPT) has sought an alternative on-pavement delineation treatment to replace PMDs. Research was performed to evaluate one alternative to PMDs, retroreflective raised pavement markers (RPMs) supplementing the existing painted centerline.

Adequate path delineation is particularly important on horizontal curves. In a study on the accident characteristics of horizontal curves on two-lane rural highways, Glennon et al. (*1*) found that the average accident rate on horizontal curves is three times that of tangent sections, that the average rate of single-vehicle run-off-the-road accidents on horizontal curves is four times the rate on tangent sections, and that single-vehicle run-off-the-road accidents were proportionally greater than other accidents under wet, icy, or nighttime conditions. Glennon (*2*) also reported that more than two-thirds of the single-vehicle run-off-the-road accidents on curves were on the outside of the curves.

A review of previous research on delineation treatments for horizontal curves on two-lane highways suggests that PMDs and RPMs, either individually or in combination, are effective supplements to painted centerlines. However, these findings have been based on limited amounts of operational data or accident models that show little correlation between accident rates and the type of delineation. No attention has been paid to the short-term effects of changing from one delineation treatment to another or to the long-term operational effectiveness of the treatments.

Therefore, a study was undertaken to evaluate the operational effectiveness of removing existing PMDs at horizontal curves on two-lane rural highways and replacing them with RPMs supplementing the existing painted centerline. The RPM treatment consisted of placing RPMs between the centerline markings at 40-ft intervals within the curve and placing four RPMs at 80-ft intervals on the tangents approaching both ends of the curve.

LITERATURE REVIEW

Two approaches have been used to evaluate the safety and operational effectiveness of PMDs and RPMs on two-lane rural highways. Some studies looked directly at the accident experience on roadway sections with various combinations of centerlines, edgelines, PMDs, and RPMs. The difficulty in directly evaluating the safety effectiveness of alternative delineation treatments prompted other studies to evaluate operational measures of effectiveness (MOEs) that are correlated to accident experience and, therefore, could be used as surrogates for accident experience in safety evaluations.

Taylor et al. (*3*) presented the state of the art in roadway delineation systems through 1972, which was the basis for most subsequent research. Operational data were collected with various combinations of centerlines, edgelines, PMDs, and RPMs at two horizontal curve locations. They concluded

Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843.

that RPMs, either alone or in conjunction with PMDs, used as supplements to existing centerlines, improve driver performance through horizontal curves, when compared to weathered painted centerlines.

Stimpson et al. (*4*) performed field studies comparing various combinations of centerlines, edgelines, PMDs, and RPMs at eight sites. They recommended 2- to 4-in. centerline and edgeline striping for continuous delineation on two-lane rural highways. They recommended RPMs to supplement the centerline where severe visibility problems caused by fog or blowing sand are common. For isolated horizontal curves, based on field studies at two locations, they concluded that RPMs are preferred over PMDs as supplements to centerlines. They also stated, however, "When RPMs cannot be used because of economic problems, consideration should be given to the installation of post delineators on the outside of the curve. Although not likely to be as beneficial as RPM supplements, PMDs apparently do provide some degree of near as well as far delineation" (*4*).

Bali et al. (*5*) developed a cost-benefit methodology for evaluating delineation treatments based on safety effectiveness. The continuous delineation treatments studied included various combinations of centerlines, edgelines, PMDs, and RPMs. Treatments for isolated horizontal curves included combinations of centerlines, edgelines, and PMDs. Accident data were obtained for more than 500 sites in 10 states. Regression models were developed to predict accident rates based on roadway, traffic, environmental, and delineation variables. Separate models were developed for tangent and winding alignments and for isolated horizontal curves. For tangent and winding alignments, they found that highways with centerlines had lower accident rates than highways without centerlines, that highways with RPM centerlines had lower accident rates than highways with painted centerlines, and that highways with PMDs had lower accident rates than highways without PMDs (both with and without edgelines). The analysis of the effects of edgelines was inconclusive. In the models for isolated horizontal curves, the type of delineation did not explain accident rate variance.

Capelle (*6*) reviewed roadway delineation research up to 1978 and noted, "Although the literature suggests that RPMs can be a very effective supplemental treatment at curves on two-lane roads, there have been very few studies of the effect of this system." Regarding PMDs, he concluded, "The evidence to date does not permit a positive recommendation of a standard for the use of post delineators as a supplement, but there is sufficient information which indicates that their use can be effective under certain conditions."

In the early 1980s, Niessner (*7,8*) coordinated separate field evaluations of PMDs and RPMs. He concluded, based on field evaluations of PMDs in eight states, "It is not possible to state that the installation of post delineators under all conditions will result in a reduction in the number of run-off-the-road-type accidents. The data that was collected indicates a trend toward reducing this type of accident with the installation of post delineators" (*7*). His finding with respect to RPMs, based on field evaluations in 12 states, was as follows: "The general consensus was that the raised pavement markers do provide improved nighttime pavement delineation when compared to and used in conjunction with conventional paint stripes. However, they should not be construed as a panacea for reducing the potential hazards at all locations" (*8*).

Several other studies have also evaluated the operational effects of RPMs and PMDs at horizontal curves on two-lane highways. Nemeth et al. (*9*) measured the distance from which a curve could be detected with various combinations of centerlines, edgelines, PMDs, and RPMs and found that, compared to no delineation, the addition of RPMs to centerline and edgelines gave the largest increase in detection distance. Zador et al. (*10*) evaluated chevrons, PMDs, and RPMs and found that vehicles moved toward the centerline when PMDs were added but moved away from the centerline when chevrons and RPMs were added. The variability in speeds and lateral placements were slightly reduced when chevrons and RPMs were used.

STUDY APPROACH

This study focused on the short-, intermediate-, and long-term operational effects of replacing the existing PMDs with RPMs supplementing the painted centerline at isolated horizontal curves on two-lane rural highways in Texas. The scope of the study was restricted to two-lane rural highways whose existing delineation consisted of weathered painted centerlines and PMDs on the outside shoulder.

The basic study approach was to focus on operational MOEs that could be observed in the field and that were good surrogates for accident experience. Study sites were selected, and data were collected with the existing PMDs, new PMDs, new RPMs, and weathered RPMs (6 weeks, 10 to 11 weeks, and 11 months old). Data were collected only at night and under clear, dry weather conditions. Statistical analyses were performed to identify the differences in the operational MOEs among the delineation treatments.

Operational Measures of Effectiveness

The study used MOEs that previous research suggests are correlated with accident rates on horizontal curves. The following MOEs were used in the study:

- Speed at the midpoint of the curve,
- Speed change from the beginning to the midpoint of the curve,
- Lateral placement at the midpoint of the curve, and
- Vehicle encroachments into the opposing lane at the midpoint of the curve.

Several researchers have argued that run-off-the-road accidents result from vehicles traveling too fast and, therefore, that it is desirable for delineation treatments to reduce mean speeds (*11,12*). Taylor et al. (*3*), however, found that there was not a statistically significant correlation between accident rates and speed measures including the mean, variance, and skewness of the speed distribution. In spite of Taylor's finding, the mean and standard deviation of speeds at the midpoint of the curves were studied because they are such fundamental measures of traffic operations.

The speed change from the beginning to the midpoint of a curve is a measure of the deceleration within the curve. Taylor et al. (*3*) concluded, "Although strong evidence does not exist in support of the hypothesis that accident rates are correlated

with deceleration rates on horizontal curves, there seems to be some justification in concluding that this correlation may also exist. It would seem that delineation treatments that reduce this statistic are ones that provide advance warning of curves." Thompson and Perkins (*13*) also identified the speed differential between the approach and midpoint of the curve as a good surrogate for the accident rate on the outside lane of isolated horizontal curves.

Stimpson et al. (*4*) suggested that the ideal vehicle path is parallel to the centerline and centered on the lane and reported that accident frequencies on tangent and winding alignments are correlated with the variance of lateral placement and with a centrality index that measures the extent to which the mean lateral placement deviates from the center of the lane. Taylor et al. (*3*) found, "A fairly strong correlation between accident rates and the variance of lateral placement on the horizontal curve seems to exist. Thus, if delineation treatments can be shown to reduce the variance in lateral placement, accident rates probably will also be reduced."

The number of vehicle encroachments is related to the variance of the lateral placements. In this study, an encroachment is said to occur if the left front wheel crosses the center of the roadway. Thompson and Perkins (*13*) reported positive correlations between the total encroachment rate (i.e., "number of edgeline plus centerline touches per 100 vehicles entering curve") and accident experience at horizontal curves. A smaller number of encroachments would be indicative of a more effective delineation treatment.

Study Sites

Five horizontal curves were studied. Site selection criteria were as follows:

- Isolated simple circular curve,
- Existing weathered painted centerlines and PMDs, but no edgelines,
- Speed limit of 45 mph or higher,
- Shoulders, if present, no wider than 4 ft,
- Minimal roadside development and, therefore, low nighttime ambient light level,
- Few, if any, intersecting driveways in the vicinity of the site, and
- Average annual daily traffic (AADT) of at least 2,000 vehicles per day.

Table 1 shows the characteristics of the curves that were studied. The degree of curvature ranged from 3 to 5 degrees, curve lengths from 850 to 1,670 ft, and pavement widths from 19 to 28 ft. All of the sites had weathered painted centerlines and PMDs on the outside of the curve. None of the curves had edgelines.

Delineation Treatments

Table 2 shows the delineation treatments studied at each site. Treatments were monitored in essentially the same sequence at all sites. During the first evening at a site, data were collected with the existing PMDs. During the next day, the PMDs were removed and new RPMs were installed. During the second evening, data were collected with the new RPMs. At two sites (FM 219 and FM 933), new delineators were placed on the posts and data were collected with the new PMDs during the second evening. At these sites, the new PMDs were removed during the third day and new RPMs were installed, and data were collected with the new RPMs during the third evening. Also at these sites, follow-up studies were conducted to monitor vehicle operations after the RPMs had been in place 6 weeks and again after 10 to 11 weeks. At the FM 1753 study site, data were collected after the RPMs had been in place 11 months.

Data Collection Procedures

The speeds and lateral placements of vehicles were measured at the beginning, midpoint, and end of the horizontal curve. Data were collected in both lanes at each location.

An automated data collection system with tapeswitches as axle sensors was used. As each vehicle axle crossed a tapeswitch, an electronic impulse was transmitted to a Golden River environmental computer. The computer recorded the time and a code for which tapeswitch was actuated, from which speeds and lateral placements were computed.

The tapeswitches were covered with a flat gray material that blended with the roadway surface. Their ¼-in.-high profile caused a barely audible rumble within vehicles passing over them. Observations of drivers passing over the tapeswitches in a previous study showed no noticeable effect on driver behavior (*14*).

Statistical Analysis Methodology

The statistical analysis was performed to identify any differences in the operational MOEs at the curves between the

TABLE 1 GEOMETRICS OF STUDY SITES

Site	AADT	Degree of Curvature	Length of Curve (ft)	Pavement Width (ft)
FM 1753	2700	5	1020	20
FM 730	1650	3	1670	19
FM 2280	3700	3	1110	22
FM 219	1350	5	850	28
FM 933	1350	4	890	25

TABLE 2 DELINEATION TREATMENTS TESTED AND SAMPLE SIZE

Site	Existing PMDs	New PMDs	New RPMs	6-Week-Old RPMs	10-11-Week-Old RPMs	11-Month-Old RPMs
FM 1753	52	--	33	--	--	27
FM 730	8	--	24	--	--	--
FM 2280	62	--	34	--	--	--
FM 219	28	28	31	28	35	--
FM 933	47	50	56	46	35	--

existing PMDs and new RPMs, and between new RPMs and RPMs that had lost some of their retroreflectivity.

The data base was screened to include only those vehicles that could be tracked through the entire study section and whose operations were unaffected by other vehicles. Vehicles that could not be tracked included vehicles on the curve when data collection began and ended or vehicles that left the roadway at driveways within the study section. Vehicles were considered to be unaffected by other vehicles if they were neither closely following another vehicle in their lane nor within the study section at the same time as a vehicle in the opposing lane. Drivers closely following other vehicles receive visual cues from the leading vehicle as well as from the roadway delineation; therefore, vehicles were removed from the data base if their headway was 4 sec or less. Illumination from the headlights of oncoming vehicles affects driver behavior; therefore, vehicles were also removed from the data base if they were within the study section during the same time as a vehicle in the opposing lane.

Prior to data collection, it was estimated that a sample size of approximately 50 vehicles would be required with each treatment at each site in order to be reasonably confident of detecting a 2-mph difference in mean speeds and 0.5-ft difference in mean lateral placements. Table 2 shows the actual number of vehicles in the data base for each treatment and site. The total sample size was 624. For all but one site, the data base included at least 25 vehicles for each treatment.

Analyses were performed to evaluate the short-, intermediate-, and long-term operational effectiveness of RPMs. The statistical analyses were performed separately for each MOE at each study site. A 0.05 significance level was used for all tests. The short-term analysis (existing PMDs versus new RPMs) compared the mean speed at the midpoint, speed change from the beginning to the midpoint, and lateral placement at the midpoint using *t*-tests. The standard deviations of these MOEs were compared using *F*-tests. The number of encroachments was analyzed using a chi-squared test. The intermediate-term effectiveness of the 6 to 11-week-old RPMs and long-term effectiveness of the 11-month-old RPMs also involved separate analyses of each MOE at each site. A single-factor analysis of variance (ANOVA) was performed to compare the speed and lateral placement MOEs with the various treatments at each site. If the ANOVA results suggested that differences existed, then a pairwise (least-significant difference) *t*-test was performed to determine which treatments were significantly different.

Separate analyses were performed for the inside and outside lanes of the curves. Previous research indicates that the run-off-the-road accident problem and the delineation requirements are greater on the outside lane than on the inside lane of a horizontal curve (*2*). The painted centerline is better illuminated by vehicles traveling on the inside lane and satisfies most of the guidance requirements. As a result, vehicle operations on the inside lane of a curve would be expected to be less affected by replacing PMDs with RPMs than on the outside lane. Indeed, no differences between the PMDs and RPMs were observed on the inside lane of the curves studied. Therefore, only results for the outside lane are presented here. Results for the inside lane are reported elsewhere (*15*).

ANALYSIS RESULTS

Short-Term Operational Data Analysis

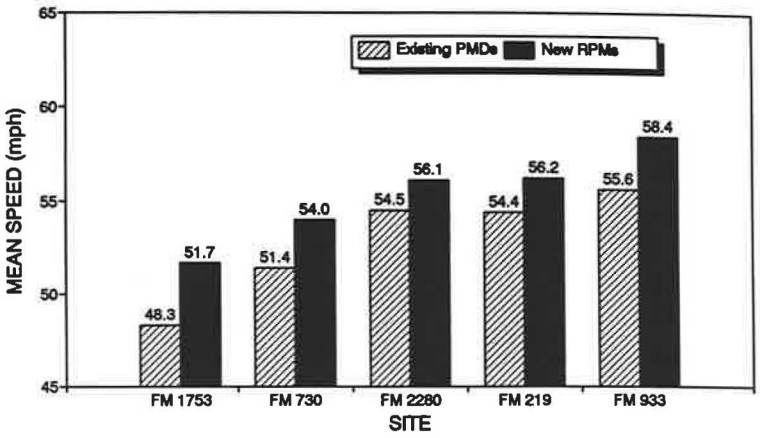
Figure 1 shows that the mean speeds with the new RPMs are consistently 1 to 3 mph higher than with the existing PMDs. The *t*-test results indicate that the differences in mean speeds between the two treatments were statistically significant only at the FM 1753 and FM 933 sites. The *F*-test results indicate that the standard deviation of speeds with the two treatments did not differ significantly at any of the sites. It is not clear whether the higher speeds with the new RPMs are good or bad. Allen et al. (*16*) found that drivers' preferred speed increases as their visual range increases. So, higher speeds may indicate that the new RPMs provided better delineation than the existing PMDs and that drivers had more confidence traveling through the curves. Others argue that speeds are a factor in run-off-the-road accidents and that it is desirable to reduce speeds at curves (*12,13*). However, previous research has not found a correlation between speeds and accident rates with different delineation treatments (*3*).

The speed change from the beginning to the midpoint of the curve was computed for each vehicle as the vehicle's speed at the midpoint minus the speed at the beginning of the curve. Therefore, a negative speed change indicates that the vehicle decelerated from the beginning to the midpoint of the curve, and a positive speed change indicates that the vehicle accelerated. Figure 2 shows the mean and standard deviation of the speed changes at each study site. There were no statistically significant differences in either the means or standard deviations for the two treatments at any of the sites.

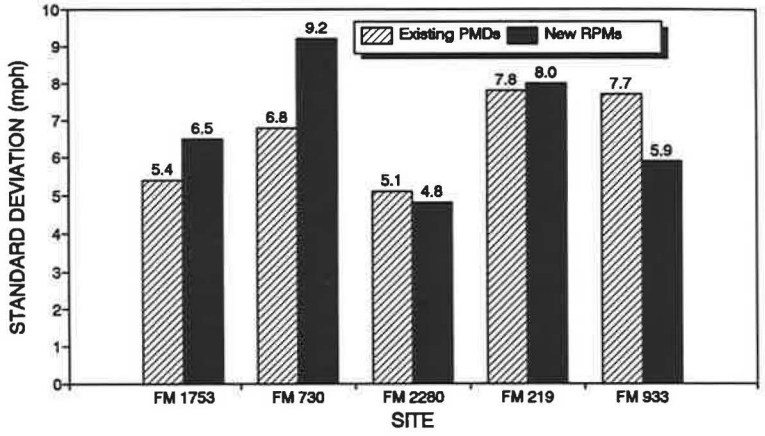
Lateral placement was measured at the midpoint of the curve from the center of the roadway to the outside edge of the left front wheel of the vehicle. As shown in Figure 3, the mean lateral placement with the new RPMs was at least 0.9 ft further from the center of the roadway than with the existing PMDs at all of the study sites. The mean lateral placement was significantly greater with the new RPMs than with the existing PMDs at all but the FM 730 site. These results demonstrate that motorists are less inclined to flatten their path through horizontal curves with RPMs than with PMDs. Previous research has suggested that the ideal vehicle path is centered in the lane. Therefore, these results suggest that the new RPMs compare favorably with the existing PMDs. The standard deviation of lateral placement was smaller with the new RPMs than with the existing PMDs at four of the five sites. The differences were statistically significant at the FM 1753 and FM 730 sites. Previous research suggests that a smaller variance in lateral placement tends to be associated with lower accident rates. Therefore, the new RPMs compare favorably to the existing PMDs with respect to this MOE.

Figure 4 shows the percentage of vehicles in the outside lane that crossed the center of the roadway at the midpoint of the curve. There were fewer encroachments with the new RPMs than with the existing PMDs at all of the sites. The differences between the treatments were statistically significant at all sites except FM 219.

The short-term data analysis suggests that drivers operated differently in the outside lane of horizontal curves when the existing PMDs were replaced with new RPMs supplementing

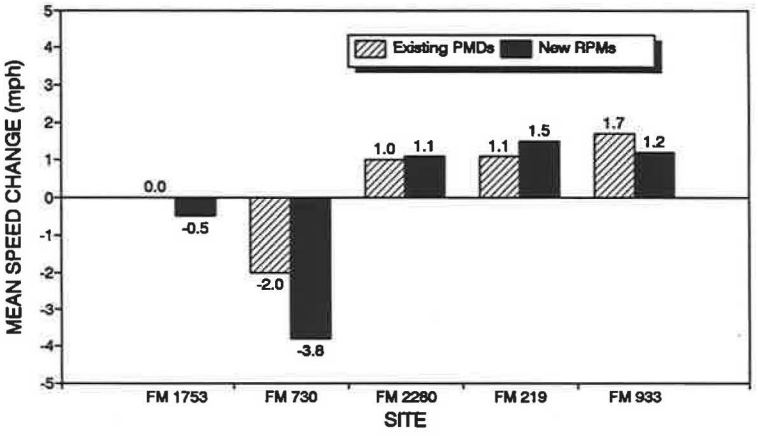


(a) Mean Speeds

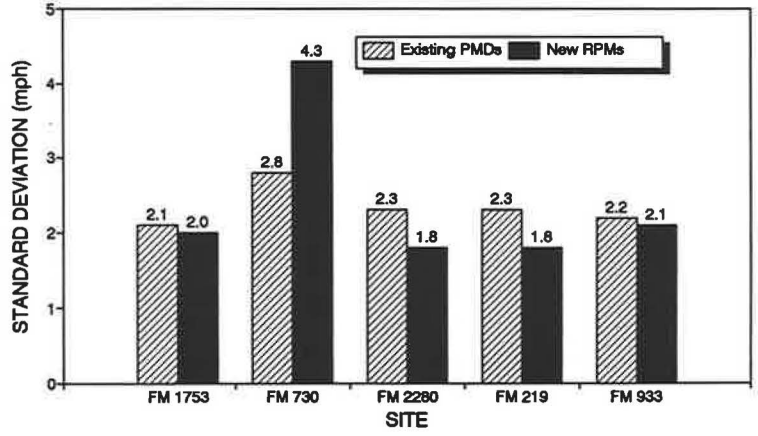


(b) Standard Deviation of Speeds

FIGURE 1 Speeds at midpoint of curve: short-term.

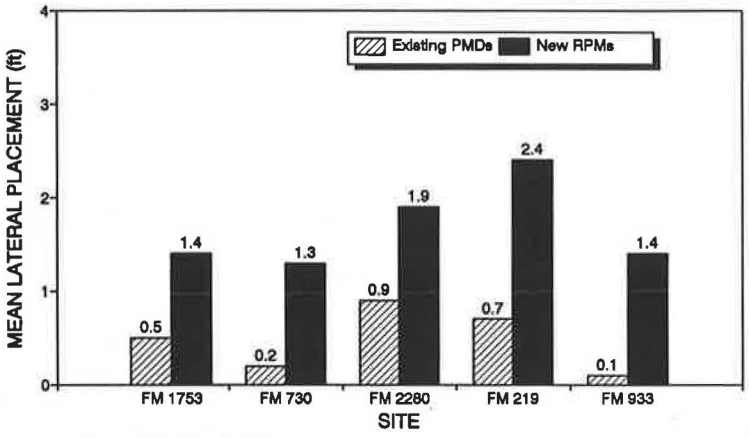


(a) Mean Speed Changes

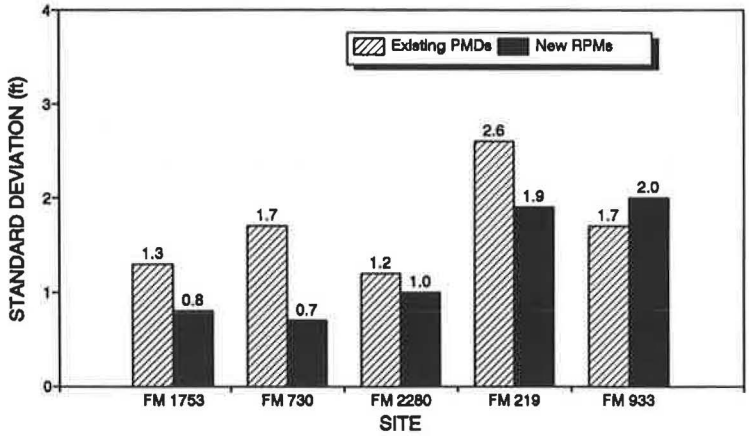


(b) Standard Deviation of Speed Changes

FIGURE 2 Speed changes from beginning to midpoint of curve: short-term.



(a) Mean Lateral Placements



(b) Standard Deviation of Lateral Placements

FIGURE 3 Lateral placements at midpoint of curve: short-term.

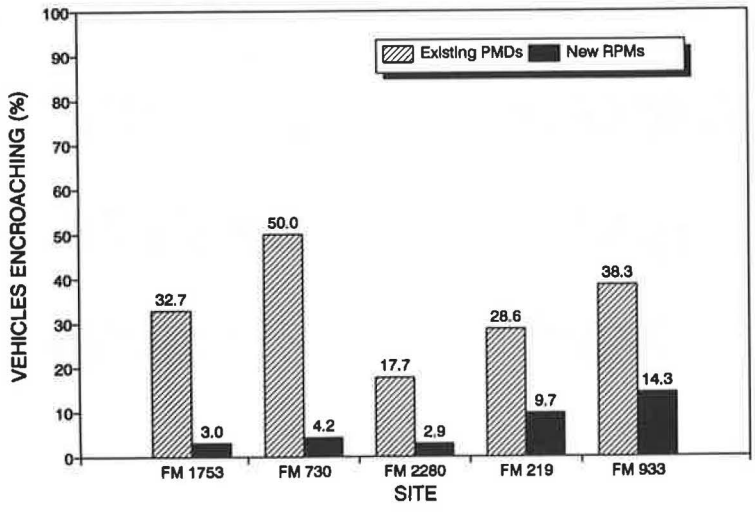


FIGURE 4 Percentage of vehicles encroaching into opposing lane at midpoint of curve: short-term.

the existing painted centerline. The following differences were observed:

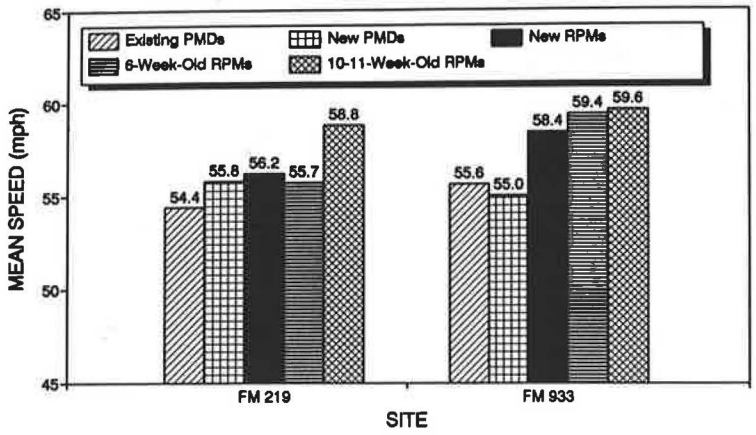
- The mean speeds at the midpoint of the curves were consistently 1 to 3 mph higher with the new RPMs than with the existing PMDs.
- The mean lateral placement was consistently 1 to 2 ft further from the center of the roadway at the midpoint of the curve with the new RPMs than with the existing PMDs.
- There was less variability in lateral placement at the midpoint of the curve with the new RPMs than with the existing PMDs.
- Fewer vehicles crossed the center of the roadway with the new RPMs than with the existing PMDs.

Overall, operations with the new RPMs compare favorably with the existing PMDs. The results suggest that the change in delineation treatments did not cause any operational problems and that the new RPMs provided better path delineation (as evidenced by the findings related to lateral placement and encroachments), which may have given drivers the confidence to operate at higher speeds through the curves.

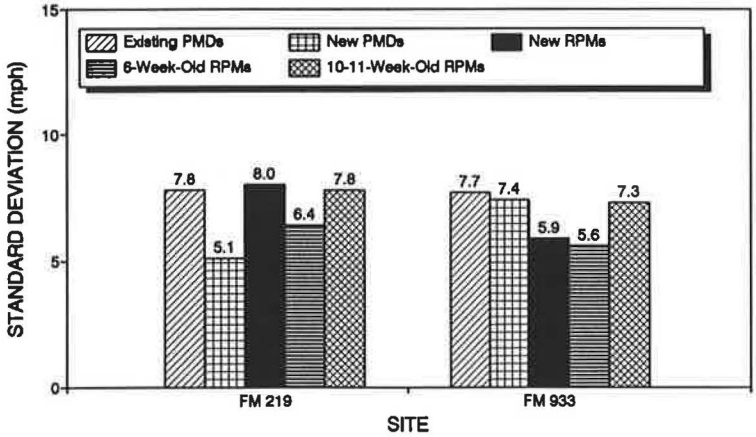
Intermediate-Term Operational Data Analysis

At two sites, follow-up field studies were conducted to monitor changes in vehicle operations through the curves after the RPMs had lost some of their retroreflectivity. Five delineation treatments were monitored at the FM 219 site: existing PMDs, new PMDs, new RPMs, 6-week-old RPMs, and 11-week-old RPMs. Retroreflectivity measurements (at a 20-degree incidence angle) for the 6-week-old and 11-week-old RPMs, 2.4 and 2.1 candlepower per foot candle (cp/ft-c), respectively, exceeded the 2.0 cp/ft-c initial-brightness specification for new RPMs in Texas. The delineation treatments at the FM 933 site were similar: existing PMDs, new PMDs, new RPMs, 6-week-old RPMs (with 2.1 cp/ft-c), and 10-week-old RPMs (with 1.0 cp/ft-c).

The speeds observed in the outside lane at the FM 219 and FM 933 sites are shown in Figure 5. The ANOVA results for the FM 219 site indicate that none of the treatments had significantly different mean speeds. The results of the single-factor ANOVA and pairwise *t*-test for the FM 933 site indicate that (*a*) the mean speeds with the three RPM treatments were not significantly different, (*b*) the mean speeds with the two



(a) Mean Speeds



(b) Standard Deviation of Speeds

FIGURE 5 Speeds at midpoint of curve: intermediate-term.

PMD treatments were not significantly different, but (*c*) the mean speeds were significantly higher with the RPM treatments than with the PMD treatments.

Figure 6 shows that the mean speed increased from the beginning to the midpoint of the curve. At the FM 219 site, the mean speed increase with the 11-week-old RPMs was significantly greater than with either the existing or the new PMDs, but none of the other pairs of treatments were significantly different. At the FM 933 site, none of the treatments differed significantly.

The lateral placement at the midpoint of the curve is shown in Figure 7 for the FM 219 and FM 933 sites. The mean lateral placements with all of the RPM treatments were significantly greater than with any of the PMD treatments at both sites.

Figure 8 shows that the percentage of vehicles encroaching was less for the RPM treatments than for the PMD treatments at both the FM 219 and FM 933 sites. The results of the chi-squared tests indicate that the differences among the treatments were statistically significant.

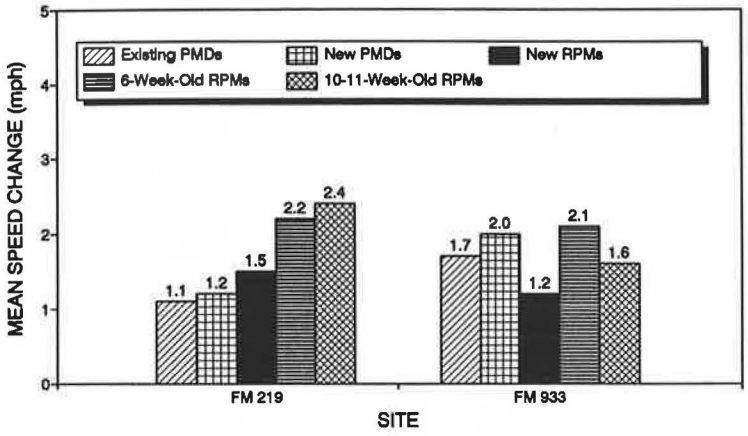
Few changes in the operational effectiveness of RPMs were observed as the RPMs aged and lost some of their retroreflectivity. At the FM 219 and FM 933 sites, little reduction in retroreflectivity was observed, and vehicle operations changed very little up to 11 weeks after the new RPMs were installed. Therefore, the results at the FM 219 and FM 933 sites reinforce the findings of the short-term analysis.

Long-Term Operational Data Analysis

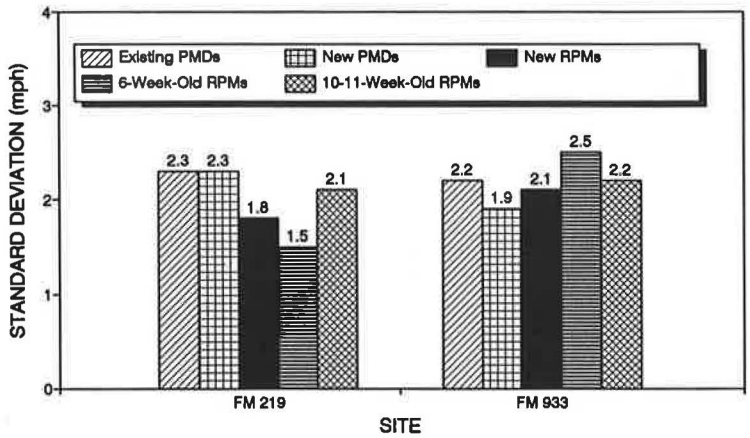
Vehicle operations were monitored at the FM 1753 site after the RPMs had been in place 11 months. The mean specific intensity of the RPMs at that time was 0.1 cp/ft-c.

The delineation treatments at the FM 1753 site were the existing PMDs, new RPMs, and 11-month-old RPMs. Figure 9 shows the mean and standard deviation of the speeds with each treatment. The single-factor ANOVA and pairwise *t*-tests results indicate that the mean speed with the new RPMs was significantly higher than with either the existing PMDs or the 11-month-old RPMs, but that the latter two did not differ significantly.

The means and standard deviations of the speed change with the three treatments at the FM 1753 site are shown in

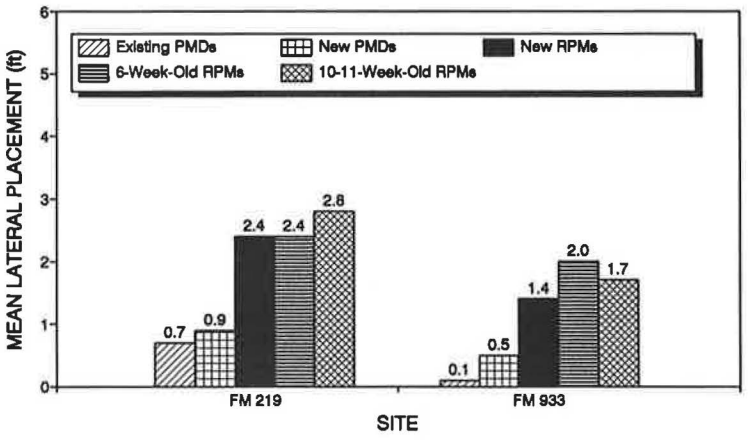


(a) Mean Speed Changes

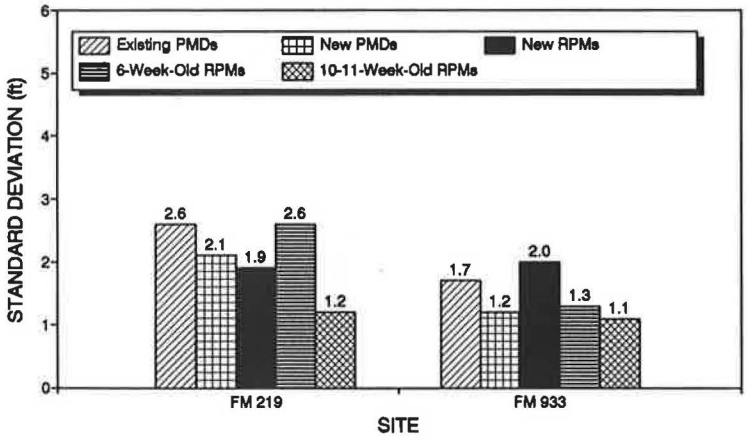


(b) Standard Deviation of Speed Changes

FIGURE 6 Speed changes from beginning to midpoint of curve: intermediate-term.



(a) Mean Lateral Placements



(b) Standard Deviation of Lateral Placements

FIGURE 7 Lateral placements at midpoint of curve: intermediate-term.

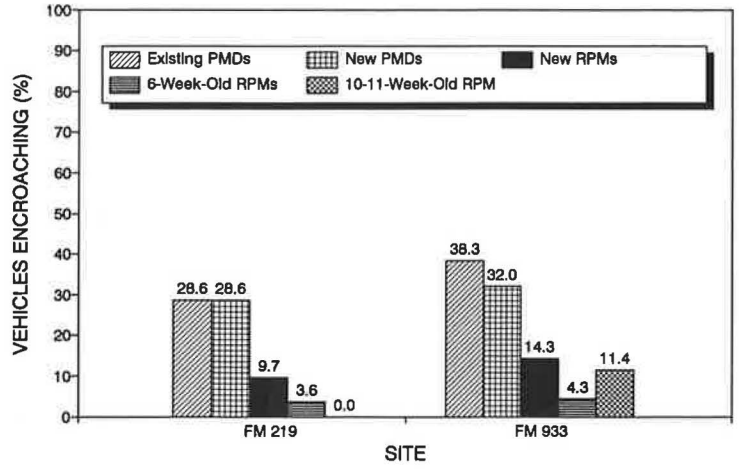
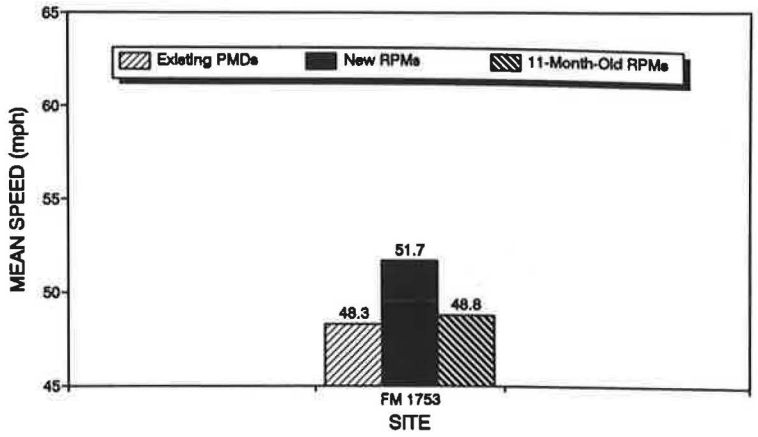
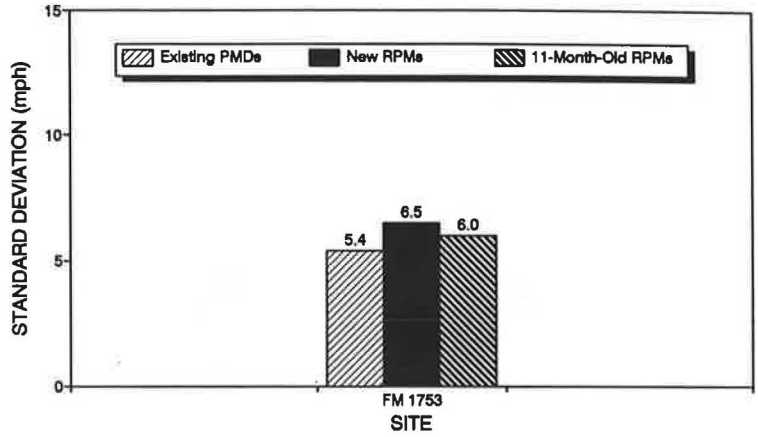


FIGURE 8 Percentage of vehicles encroaching into opposing lane at midpoint of curve: intermediate-term.



(a) Mean Speeds



(b) Standard Deviation of Speeds

FIGURE 9 Speeds at midpoint of curve: long-term.

Figure 10. The mean speed reduction with the 11-month-old RPMs was significantly greater than with either the existing PMDs or new RPMs, but the latter two did not differ significantly.

Summary statistics for the lateral placement at the midpoint of the curve are shown in Figure 11 for the FM 1753 site. The mean lateral placement with both the new and 11-month-old RPMs was greater than with the existing PMDs, but the two RPM treatments were not significantly different.

Figure 12 shows that the proportion of vehicles encroaching was less with the RPM treatments than with the existing PMDs. The results of the chi-squared tests indicate that the differences among the treatments were statistically significant.

The 11-month-old RPMs at the FM 1753 compared favorably with the existing PMDs and were similar to the new RPMs with respect to the mean lateral placement and the number of encroachments. These results suggest that the RPMs, in spite of their loss of retroreflectivity, continued to provide adequate near delineation. The mean speed at the midpoint of the curve with the 11-month-old RPMs was the same as with the existing PMDs but was significantly lower than with the new RPMs. These results are an indication of the relative visual range provided by the treatments. The only MOE that caused concern with the 11-month-old RPMs was a small but statistically significant increase in deceleration from the beginning to the midpoint of the curve. This increase suggests that the RPMs' effectiveness at providing far delineation was somewhat degraded. Unfortunately, there is no objective basis for determining whether the far delineation provided by the 11-month-old RPMs was adequate.

The results from the FM 1753 site suggest that the operational effectiveness of RPMs is due in part to their retroreflectivity and in part to their profile above the pavement surface. Even with low retroreflectivity, it appears that the RPMs continue to serve at least part of their intended function because of their profile.

SUMMARY AND RECOMMENDATIONS

The operational effectiveness of RPMs as an alternative to PMDs at horizontal curves on two-lane rural highways was

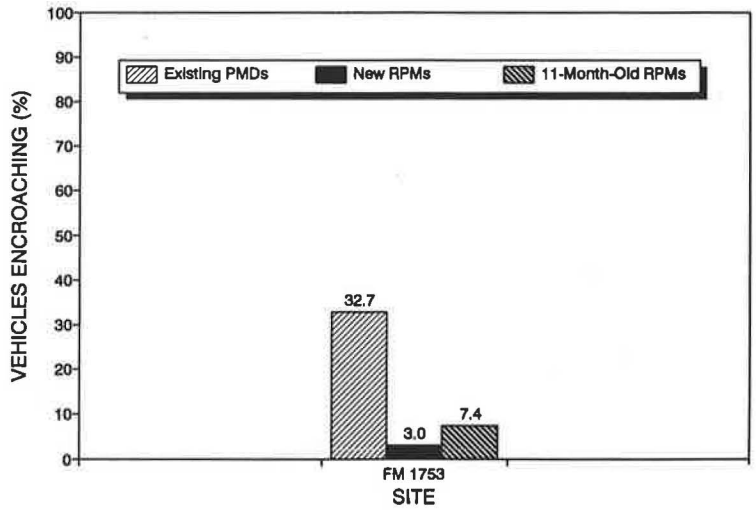


FIGURE 12 Percentage of vehicles encroaching into opposing lane at midpoint of curve: long-term.

evaluated based upon nighttime speed and lateral placement data at five horizontal curves. The analysis focused on those MOEs that previous research suggests are correlated to accident rates at horizontal curves.

The statistical analysis of the short-term operational data suggested that vehicle operations on the inside lane of the curves were not significantly affected by the removal of the PMDs and installation of new RPMs. However, several significant differences were observed on the outside lane of the curves. The mean speeds at the midpoint of the curves were consistently 1 to 3 mph higher with the new RPMs than with the existing PMDs. The mean lateral placement was consistently 1 to 2 ft further from the center of the roadway at the midpoint of the curves with the new RPMs than with the existing PMDs. The variability in lateral placement at the midpoint of the curve was less with the new RPMs than with the existing PMDs. Fewer vehicles crossed the center of the roadway with the new RPMs than with the existing PMDs. The short-term evaluation suggests that the new RPMs provided better path delineation (as evidenced by the findings related to lateral placement and encroachments), which may have given drivers the confidence to operate at higher speeds through the curves.

Intermediate-term operational data were collected at two sites. Data were collected after the RPMs had been in place 6 weeks and again after 10 to 11 weeks. The RPMs retained much of their retroreflectivity at these sites, and the results of the data analysis reinforce the findings of the short-term evaluation.

At one site, data were collected after the RPMs had been in place 11 months and had lost much of their retroreflectivity (mean specific intensity of 0.1 cp/ft-c). The mean lateral placement and number of encroachments with the 11-month-old RPMs were not significantly different than with the new RPMs and compared favorably with the PMDs. The fact that the mean speed at the midpoint of the curve with the 11-month-old RPMs was similar to the existing PMDs but less than with the new RPMs is an indication of the relative visual range provided by the treatments. The only MOE that caused concern with the 11-month-old RPMs was the small, but statistically significant, increase in deceleration from the beginning to the midpoint of the curve, which may indicate that motorists did not receive sufficient advance warning of the curve. These results suggest that after 11 months the RPMs continued to provide near delineation but that their far delineation was at least partially degraded.

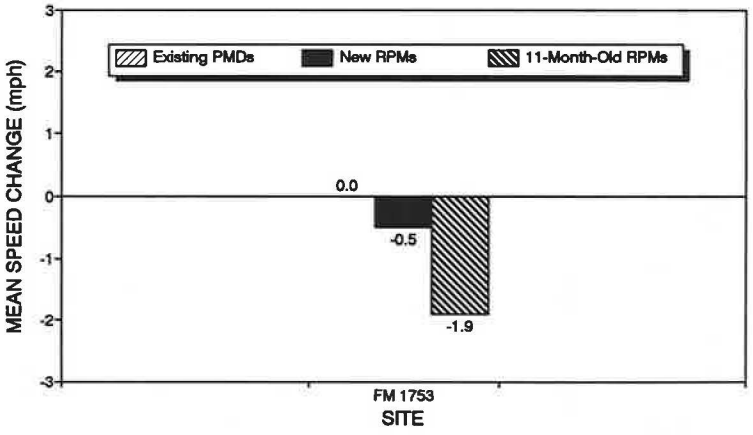
There is no objective basis for defining minimum performance levels for RPMs. Because previous research has not addressed the long-term effectiveness of RPMs and because this study involved a long-term evaluation at only one site, additional research would be necessary to determine at what point RPMs no longer function adequately (i.e., the service life of RPMs). It would be desirable to determine how the operational effectiveness of RPMs changes as they lose retroreflectivity and to define a minimum performance level in terms of the operational MOEs used here.

ACKNOWLEDGMENTS

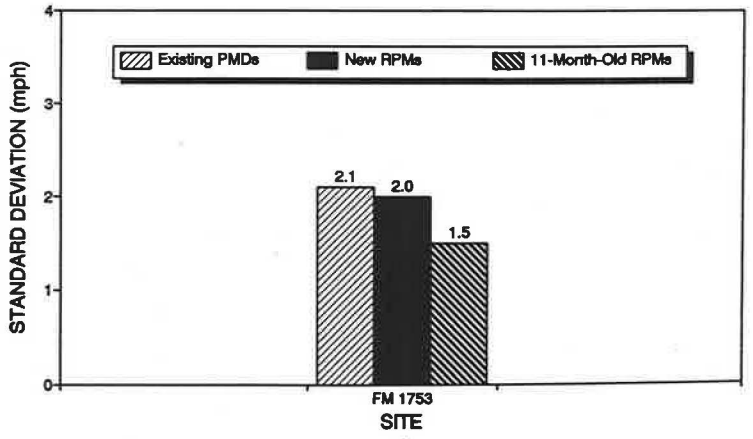
The research reported herein was sponsored by the Texas State Department of Highways and Public Transportation (SDHPT) in cooperation with FHWA, U.S. Department of Transportation. Lewis R. Rhodes, Jr. served as the Texas SDHPT technical coordinator for the study and provided much valuable guidance.

REFERENCES

1. J. C. Glennon, T. R. Neuman, and J. E. Leisch. *Safety and Operational Considerations for Design of Rural Highway Curves.* Report FHWA-RD-86/035. FHWA, U.S. Department of Transportation, 1985.
2. J. C. Glennon. Effect of Alignment on Highway Safety. In *State-of-the-Art Report 6*, TRB, National Research Council, Washington, D.C., 1987.
3. J. I. Taylor, H. W. McGee, E. L. Seguin, and R. S. Hostetter. *NCHRP Report 130: Roadway Delineation Systems.* HRB, National Research Council, Washington, D.C., 1972.

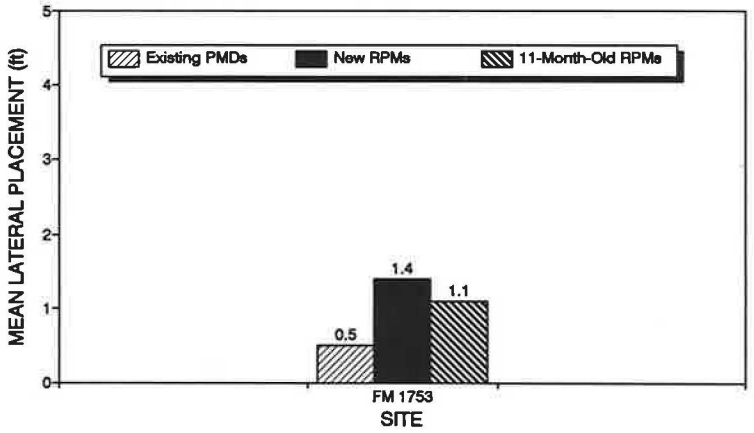


(a) Mean Speed Changes

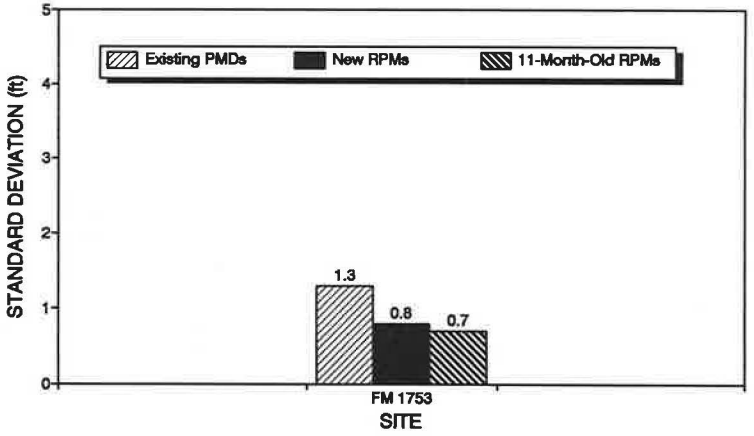


(b) Standard Deviation of Speed Changes

FIGURE 10 Speed changes from beginning to midpoint of curve: long-term.



(a) Mean Lateral Placements



(b) Standard Deviation of Lateral Placements

FIGURE 11 Lateral placements at midpoint of curve: long-term.

4. W. A. Stimpson, H. W. McGee, W. K. Kittelson, and R. H. Ruddy. *Field Evaluation of Selected Delineation Treatments on Two-Lane Rural Highways*. Report FHWA-RD-77-118. FHWA, U.S. Department of Transportation, 1977.
5. S. Bali, R. Potts, J. A. Fee, J. I. Taylor, and J. Glennon. *Cost-Effectiveness and Safety of Alternative Roadway Delineation Treatments for Rural Two-Lane Highways*. Report FHWA-RD-78-51. FHWA, U.S. Department of Transportation, 1978.
6. D. G. Capelle. *Overview of Roadway Delineation Research*. Report FHWA-RD-78-111. FHWA, U.S. Department of Transportation, 1978.
7. C. W. Niessner. *Post Mounted Delineators*. Report FHWA-TS-83-208. FHWA, U.S. Department of Transportation, 1983.
8. C. W. Niessner. *Raised Pavement Markers at Hazardous Locations*. Report FHWA-TS-84-215. FHWA, U.S. Department of Transportation, 1984.
9. Z. A. Nemeth, T. H. Rockwell, and G. L. Smith. *Recommended Delineation Treatments at Selected Situations on Rural State Highways*. Report FHWA/OH-85/002. Ohio Department of Transportation, Columbus, 1985.
10. P. Zador, H. S. Stein, P. Wright, and J. Hall. Effects of Chevrons, Post-Mounted Delineators, and Raised Pavement Markers on Driver Behavior at Roadway Curves. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987, pp. 1–10.
11. T. H. Rockwell and J. C. Hungerford. *Use of Delineation Systems to Modify Driver Performance on Rural Curves*. Report FHWA/OH-79/007. Ohio Department of Transportation, Columbus, 1979.
12. P. H. Wright, J. W. Hall, and P. L. Zador. Low-Cost Countermeasures for Ameliorating Run-off-the-Road Crashes. In *Transportation Research Record 926*, TRB, National Research Council, Washington, D.C., 1983, pp. 1–7.
13. H. T. Thompson and D. D. Perkins. Surrogate Measures for Accident Experience at Rural Isolated Horizontal Curves. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 142–147.
14. C. L. Dudek, R. D. Huchingson, F. T. Creasey, and O. Pendleton. Field Studies of Temporary Pavement Markings at Overlay Project Work Zones on Two-Lane, Two-Way Rural Highways. In *Transportation Research Record 1160*, TRB, National Research Council, Washington, D.C., 1988, pp. 22–34.
15. R. A. Krammes, K. D. Tyer, D. R. Middleton, and S. A. Feldman. *An Alternative to Post-Mounted Delineators at Horizontal Curves on Two-Lane Highways*. Report FHWA/TX-90/1145-1F. Texas State Department of Highways and Public Transportation, Austin, 1990.
16. R. W. Allen, J. F. O'Hanlon, and D. T. McRuer. *Driver's Visibility Requirements for Roadway Delineation, Volume 1, Effects of Contrast and Configuration on Driver Performance and Behavior*. Report FHWA-RD-77-165. FHWA, U.S. Department of Transportation, 1977.

The contents of this paper reflect the views of the authors, who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of FHWA or the Texas State Department of Highways and Public Transportation. This paper does not constitute a standard, specification, or regulation.

Publication of this paper sponsored by Committee on Traffic Control Devices.

Scheme To Optimize Circular Phasing Sequences

NADEEM A. CHAUDHARY, ANULARK PINNOI, AND CARROLL J. MESSER

The provision of traffic progression along an arterial street has long been accepted as a desirable traffic control objective to improve the level of service. Bandwidth maximization is often used to optimize progression. A new scheme optimizes two circular phasing sequences in addition to those available in existing bandwidth maximizing programs. The new circular sequences, having the form main1-cross1-main2-cross2, can be clockwise or counterclockwise. The MAXBAND 89T program uses this enhanced optimization capability to find maximum progression bands on an arterial. This scheme expands the arterial formulation used in MAXBAND 86. It uses the mixed-integer linear programming method for optimizing progression bandwidth. Experimental results for some cases show that the new formulation can produce wider progression bands on an arterial than those of the MAXBAND 86 formulation.

The problem of finding signal timings that produce the maximum sum of progression bands along a two-way arterial was first modeled by Little (*1*) as a mixed-integer linear program (MILP). This formulation found the cycle length and offsets for a two-phase signal system. The basic MILP formulation was later enhanced by Little et al. (*2*) for optimizing National Electrical Manufacturers Association (NEMA) left-turn phasing sequences. Little et al. also developed the MAXBAND computer program (*2*) to optimize signal timing on signalized arterials and triangular networks. In 1986, MAXBAND was upgraded to MAXBAND 86 (*3*). The enhanced program was capable of optimizing signal timing on suburban arterials and urban grid networks. In recent years, Tsay et al. (*4*) and Gartner et al. (*5*) have produced further enhancements to the arterial bandwidth formulation used in MAXBAND; both of these enhancements deal with the shape of the progression bands. In addition, Chaudhary et al. (*6*; see also companion paper in this Record) and Mireault (*7,8*) have recently developed more efficient schemes to optimize arterial and network signal timing problems.

We present another enhancement to Little's basic MILP formulation for the arterial signal timing optimization problem (*2*). This enhancement provides the capability to optimize two new circular phasing sequences in addition to the existing NEMA phasing sequence. The form of these four-phase sequences is main1-cross1-main2-cross2 (i.e., main lead–cross lead–main lag–cross lag), as opposed to the conventional four-phase sequences for which both green phases on one arterial are provided in a single contiguous block as main-cross (i.e., main lead–main lag–cross lead–cross lag). Figure 1 shows the circular phasing, which can be either clockwise or counterclockwise. Evaluation of the new formulation shows that it can produce wider bands than the original formulation and improve arterial performance in some cases.

Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843-3135.

ENHANCED MATHEMATICAL FORMULATION

In this section we develop the enhanced arterial formulation. Readers not interested in the mathematical details may wish to scan this section. First, the original MILP arterial formulation is reproduced. A formulation that has the capability to optimize only circular phasing sequences is then shown. Finally, we show how these two formulations can be combined to produce a comprehensive formulation capable of optimizing both NEMA and circular phasing sequences.

Original MILP Arterial Formulation

The arterial formulation for maximizing progression bandwidth is obtained by using the basic progression bandwidth geometry shown in Figure 2. The variable are defined as follows:

- b ($\bar{b}$) = outbound (inbound) bandwidth, cycles;
- z = signal cycle, inverse of cycle length;
- S_i = signal i, $i = 1, \ldots, n$;
- w_i ($\bar{w}_i$) = time from right (left) side of red at S_i to left (right) edge of outbound (inbound) green band, cycles;
- t_i ($\bar{t}_i$) = outbound (inbound) travel time from S_i to S_{i+1} (S_{i+1} to S_i), cycles;
- Δ_i = time in cycles from the center of $\bar{r}_i$ to the nearest center of r_i: positive if the center of r_i is to the right of the center of $\bar{r}_i$;
- δ_i ($\bar{\delta}_i$) = 0-1 variables for phasing sequence selection; and
- m_i = an integer number.

The constants are defined as follows:

- r_i ($\bar{r}_i$) = outbound (inbound) red time at S_i, cycles;
- g_i ($\bar{g}_i$) = outbound (inbound) green time for through traffic at S_i, cycles;
- ℓ_i ($\bar{\ell}_i$) = time allocated for outbound (inbound) left-turn green at S_i, cycles;
- τ_i ($\bar{\tau}_i$) = queue clearance time, in cycles, an advance of the outbound (inbound) bandwidth upon leaving S_i;
- c ($\bar{c}$) = outbound (inbound) objective function weight;
- k = inbound to outbound target bandwidth ratio;
- T_1 = lower limit on signal cycle length;
- T_2 = upper limit on signal cycle length;

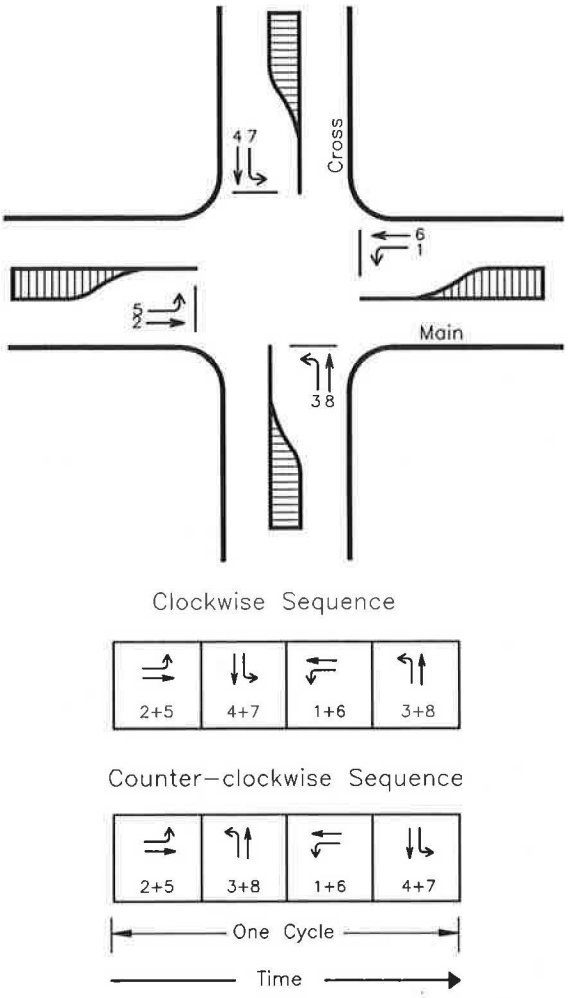


FIGURE 1 Circular phasing sequences.

d_i ($\bar{d}_i$) = outbound (inbound) distance between S_i and S_{i+1};
e_i ($\bar{e}_i$) = lower limit on outbound (inbound) speed on link between S_i and S_{i+1};
f_i ($\bar{f}_i$) = upper limit on outbound (inbound) speed on link between S_i and S_{i+1};
$\frac{1}{h_i}\left(\frac{1}{\bar{h}_i}\right)$ = lower limit on outbound (inbound) reciprocal change between two adjacent links; and
$\frac{1}{g_i}\left(\frac{1}{\bar{g}_i}\right)$ = lower limit on outbound (inbound) reciprocal change between two adjacent links.

The fundamental loop equation for formulating the arterial problem obtained from Figure 2 is as follows:

$$(w_i + \bar{w}_i) - (w_{i+1} + \bar{w}_{i+1}) + (t_i + \bar{t}_i) + \Delta_i - \Delta_{i+1} - m_i$$

$$= -\frac{1}{2}(r_i + \bar{r}_i) + \frac{1}{2}(r_{i+1} + \bar{r}_{i+1}) + (\bar{\tau}_i + \tau_{i+1}) \quad (1)$$

Figure 3 shows the four conventional left-turn green phases. The time from the center of $\bar{r}_i$ to the next center of r_i in terms of ℓ_i and $\bar{\ell}_i$ for each case, (Δ_i), can be expressed as a single equation having two binary variables as

$$\Delta_i = \frac{1}{2}\left[(2\delta_i - 1)\ell_i - (2\bar{\delta}_i - 1)\bar{\ell}_i\right] \quad (2)$$

Each of the four possible left-turn patterns can be determined by the following combinations of binary decision variables:

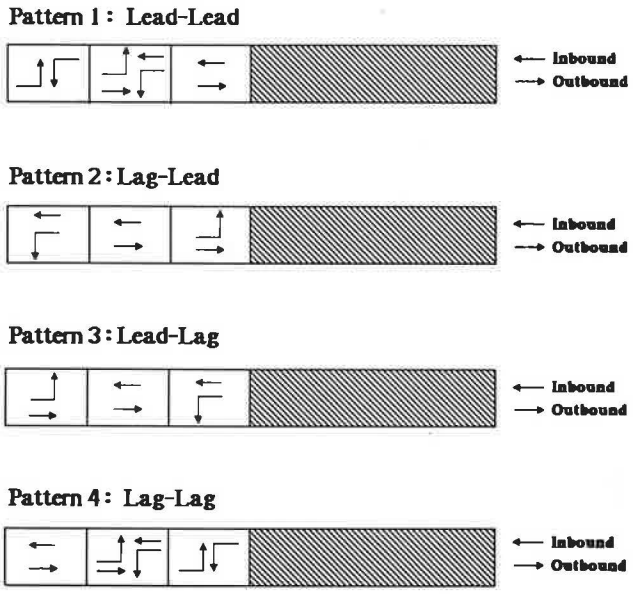


FIGURE 3 Four conventional phasing sequences.

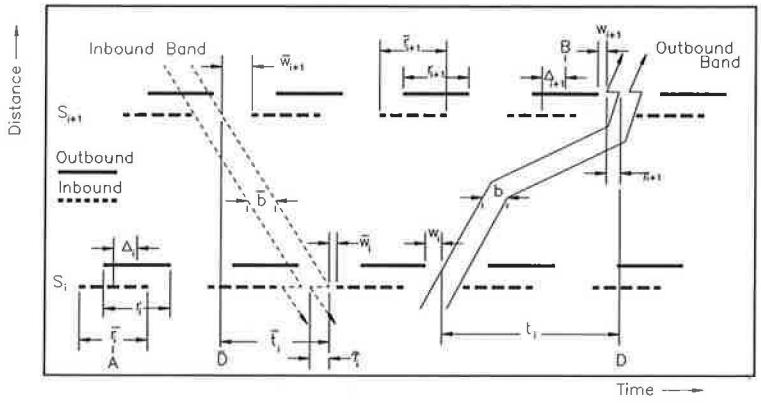


FIGURE 2 Basic progression bandwidth geometry.

Pattern	Signal Phase Sequence	δ_i	$\bar{\delta}_i$
1	Outbound leads–inbound leads	0	1
2	Outbound lags–inbound leads	1	0
3	Outbound leads–inbound lags	0	0
4	Outbound lags–inbound lags	1	1

Substituting Equation 2 in Equation 1 eliminates Δ_i and Δ_{i+1} and results in a more simplified equation. However, for ease in presenting this material, we use the two equations as they are. Following is the complete MILP arterial formulation:

MILP 1 Find $b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i, \delta_i, \bar{\delta}_i, m_i$ and Δ_i to maximize $cb + \bar{c}\bar{b}$ subject to

$$(w_i + \bar{w}_i) - (w_{i+1} + \bar{w}_{i+1}) + (t_i + \bar{t}_i) + \Delta_i - \Delta_{i+1} - m_i = -\frac{1}{2}(r_i + \bar{r}_i) + \frac{1}{2}(r_{i+1} + \bar{r}_{i+1}) + (\bar{\tau}_i + \tau_{i+1})$$

$$i = 1, \ldots, n-1 \quad (1)$$

$$\Delta_i = \frac{1}{2}[(2\delta_i - 1)\ell_i - (2\bar{\delta}_i - 1)\bar{\ell}_i] \qquad i = 1, \ldots, n \quad (2)$$

$$-kb + \bar{b}\begin{cases} = 0 & \text{if } k = 1 \\ \geq 0 & \text{if } k < 1 \\ \leq 0 & \text{if } k > 1 \end{cases} \quad (3)$$

$$\frac{1}{T_2} \leq z \leq \frac{1}{T_1} \quad (4)$$

$$w_i + b \leq 1 - r_i \qquad i = 1, \ldots, n \quad (5a)$$

$$\bar{w}_i + \bar{b} \leq 1 - \bar{r}_i \qquad i = 1, \ldots, n \quad (5b)$$

$$\left(\frac{d_i}{f_i}\right)z \leq t_i \leq \left(\frac{d_i}{e_i}\right)z \qquad i = 1, \ldots, n-1 \quad (6a)$$

$$\left(\frac{\bar{d}_i}{\bar{f}_i}\right)z \leq \bar{t}_i \leq \left(\frac{\bar{d}_i}{\bar{e}_i}\right)z \qquad i = 1, \ldots, n-1 \quad (6b)$$

$$\left(\frac{d_i}{h_i}\right)z \leq \left(\frac{d_i}{d_{i+1}}\right)t_{i+1} - t_i \leq \left(\frac{d_i}{g_i}\right)z \qquad i = 1, \ldots, n-2 \quad (7a)$$

$$\left(\frac{\bar{d}_i}{\bar{h}_i}\right)z \leq \left(\frac{\bar{d}_i}{\bar{d}_{i+1}}\right)\bar{t}_{i+1} - \bar{t}_i \leq \left(\frac{\bar{d}_i}{\bar{g}_i}\right)z \qquad i = 1, \ldots, n-2 \quad (7b)$$

Optional Left-Turn Choice Constraints on δ_i and $\bar{\delta}_i$

$$i = 1, \ldots, n$$

$b, \bar{b}, z, w_i, \bar{w}_i, t_j, \bar{t}_j$, and $m_j \geq 0$

$$i = 1, \ldots, n, j = 1, \ldots, n-1$$

Δ_i unrestricted continuous variables

$$i = 1, \ldots, n$$

m_i general integer variables

$$i = 1, \ldots, n-1$$

δ_i and $\bar{\delta}_i$ binary variables

$$i = 1, \ldots, n \quad (8)$$

Circular Phasing Sequence Optimization Capability

In this section we present details on an arterial formulation that selects the cycle length, offsets, and only circular phasing sequences to maximize the sum of progression bands in both arterial directions. There are two reasons for devoting a section to this development. The first is to document completely this major step in the production of the complete mathematical formulation derived in the next section. The second is to emphasize that circular phasing sequence is not conventionally used in progression-based signal timings, and that reprogramming conventional signal controllers may be needed to implement a circular phasing sequence. However, this is not a difficult procedure with microprocessor-based systems.

The circular phasing sequences shown in Figure 1 have four signal phases, each of which displays a green indication to both through and left-turn movements on a signalized approach. These phasing sequences are different from the conventional NEMA sequences in two ways. First, the green splits for circular phasing sequences, in general, are not the same as those for NEMA phases. Second, the time between centers of red for inbound and outbound directions is different. Therefore, there is a need to calculate new green splits for all phases, and to develop a new equation representing time between the centers of red for inbound and outbound movements on an arterial.

Green Split Calculation

As opposed to Webster's method (*9*) of computing green splits for the NEMA sequences used in the original formulation, the green splits for the circular phasing sequences are calculated as follows:

$$\max(g_{1,i}, g_{6,i}) + \max(g_{2,i}, g_{5,i}) + \max(g_{4,i}, g_{7,i}) + \max(g_{3,i}, g_{8,i}) = 1$$

$G_{1,i} = G_{6,i} = \max(g_{1,i}, g_{6,i})$

= main-street outbound movement

$G_{2,i} = G_{5,i} = \max(g_{2,i}, g_{5,i})$

= main-street inbound movement

$G_{4,i} = G_{7,i} = \max(g_{4,i}, g_{7,i})$

= cross-street outbound movement

$$G_{3,i} = G_{8,i} = \max(g_{3,i}, g_{8,i})$$

$$= \text{cross-street inbound movement}$$

where

$g_{m,i}$ = calculated green based on volume to saturation flow ratio for movement m of signal i, and
$G_{m,i}$ = green split for movement m of signal i for a circular phasing sequence.

The labels "inbound" and "outbound" are assigned to the movements for consistency with the original MILP formulation. The split calculation given above requires the following modifications to Constraints 1, 5a, and 5b given previously:

$$(w_i + \bar{w}_i) - (w_{i+1} + \bar{w}_{i+1}) + (t_i + \bar{t}_i) + \Delta_i - \Delta_{i+1} - m_i$$

$$= -\frac{1}{2}(R_{6,i} + R_{2,i}) + \frac{1}{2}(R_{6,i+1} + R_{2,i+1})$$

$$+ (\bar{\tau}_i + \tau_{i+1}) \qquad i = 1, \ldots, n-1 \tag{9}$$

$$w_i + b \leq 1 - R_{6,i} \qquad i = 1, \ldots, n \tag{10a}$$

$$\bar{w}_i + \bar{b} \leq 1 - R_{2,i} \qquad i = 1, \ldots, n \tag{10b}$$

Equation for Selecting Best Circular Sequence

Following is the derivation of a single equation describing the time between the centers of red for the two arterial phases in a circular phasing sequence.

For the clockwise phasing sequence:

$$\Delta_{6-2,i} = \frac{1}{2} - \frac{1}{2}(G_{8,i} - G_{4,i}) = \frac{1}{2} - \frac{1}{2}(R_{4,i} - R_{8,i})$$

$$\Delta_{4-8,i} = \frac{1}{2} - \frac{1}{2}(G_{6,i} - G_{2,i}) = \frac{1}{2} - \frac{1}{2}(R_{2,i} - R_{6,i})$$

For the counterclockwise phasing sequence:

$$\Delta_{6-2,i} = \frac{1}{2} + \frac{1}{2}(G_{8,i} - G_{4,i}) = \frac{1}{2} + \frac{1}{2}(R_{4,i} - R_{8,i})$$

$$\Delta_{4-8,i} = \frac{1}{2} + \frac{1}{2}(G_{6,i} - G_{2,i}) = \frac{1}{2} + \frac{1}{2}(R_{2,i} - R_{6,i})$$

where

$\Delta_{j-k,i}$ = difference between the centers of red for outbound movement j and inbound movement k on signal i of the artery, and
$R_{m,i} = 1 - G_{m,i}$ is the red split for movement m of signal i.

Subscripts 6 and 4 represent outbound movements on the main and cross arteries, respectively. Subscripts 2 and 8 represent inbound movement on the main and cross arteries, respectively.

As opposed to the original formulation, the Δ values derived here for the main artery contain red splits for the cross artery, and vice versa. Combining the above equations to obtain a single set of equations for both clockwise and counterclockwise phasing sequences, we have

$$\Delta_{6-2,i} = \Delta_i = \frac{1}{2} + \left(\frac{1}{2} - \beta_i\right)(R_{4,i} - R_{8,i}) \tag{11}$$

for the main artery, and

$$\Delta_{4-8,i} = \Delta_{c,i} = \frac{1}{2} + \left(\frac{1}{2} - \beta_i\right)(R_{2,i} - R_{6,i}) \tag{12}$$

for the cross artery.

The binary decision variable (β_i) selects clockwise phasing for the ith signal when its value is 1 and selects counterclockwise phasing for the ith signal when its value is 0. Equation 11 is a replacement for Equation 2 of the original formulation. Equation 12 is needed only when one also desires to simultaneously optimize bands on the cross artery at signal i (i.e., in a multiarterial network problem).

Substituting Equations 1, 2, 5a, and 5b in the original arterial formulation with Equations 9, 11, 10a, and 10b, we obtain a new formulation that is capable of optimizing only circular phasing sequences.

The new formulation was manually tested on several real-world test problems. These problems were optimized using the LINDO optimization package (*10*) on a personal computer. Figure 4 shows the time-space diagram for a five-intersection test problem using only circular phasing sequences. The intersections at 906 ft and 1,965 ft have only three phases. This illustrates the fact that for a three-legged intersection, a circular phasing sequence reduces to a conventional three-phase sequence with either lead-lag or lag-lead phasing for the two-way street.

Combined NEMA and Circular Phasing Optimization Capability

Next, we develop a comprehensive formulation having the capability to select either NEMA or circular phasing sequences that produce the maximum total progression bands on an arterial. We accomplish this by combining the original and new formulations described previously. The process of combining these two formulations is slightly more complicated because Constraints 1, 2, 5a, and 5b in the original formulation and corresponding Constraints 9, 11, 10a, and 10b in the new formulation are mutually exclusive. We combine these constraints by introducing binary variables into the problem formulation. The purpose of these variables is to provide a systematic way of selecting either NEMA or circular phasing sequences. Additional variables are defined as follows:

$R_{6,i}$ = main-street outbound approach (Movements 1 + 6) red split for signal i in a circular phasing;
$R_{2,i}$ = main-street inbound approach (Movements 2 + 5) red split for signal i in a circular phasing;
$R_{8,i}$ = circular phase red split for cross-street Movements 3 + 8 at node i;
$R_{4,i}$ = circular phase red split for cross-street Movements 4 + 7 for node i;

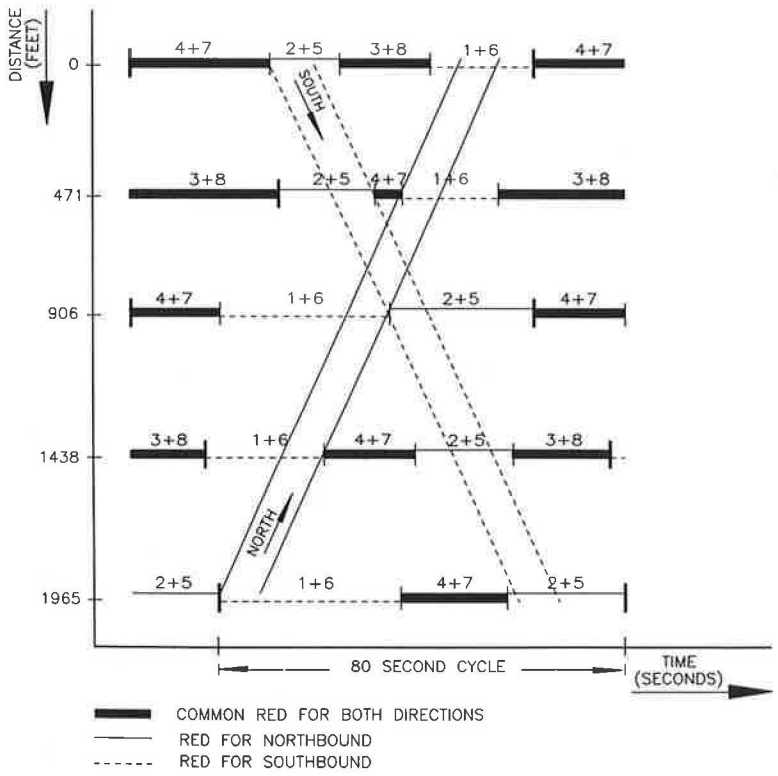


FIGURE 4 Time-space diagram for solution with only circular phases.

β_i = binary variable that selects one of the two circular phasing sequences: a value of 0 selects the counterclockwise sequence, and a value of 1 selects the clockwise sequence; and

α_i = binary variable that selects between the NEMA and the circular phasing sequences: a value of 1 selects the NEMA sequence, and a value of 0 selects the circular sequence.

Constraints on Progression Bands

Constraints 5a and 5b are combined with Constraints 10a and 10b, respectively, to form the following two constraints:

$$w_i + b + (r_i - R_{6,i})\alpha_i \leq 1 - R_{6,i} \tag{13a}$$

$$\bar{w}_i + \bar{b} + (\bar{r}_i - R_{2,i})\alpha_i \leq 1 - R_{2,i} \tag{13b}$$

Loop Constraints

The Loop Constraints 1 and 9 are combined using the binary variable α_i, defined above, as follows:

$$(w_i + \bar{w}_i) - (w_{i+1} + \bar{w}_{i+1}) + (t_i + \bar{t}_i) - m_i + \Delta_i$$

$$+ \frac{1}{2}(r_i + \bar{r}_i - R_{6,i} - R_{2,i})\alpha_i - \Delta_{i+1}$$

$$- \frac{1}{2}(r_{i+1} + \bar{r}_{i+1} - R_{6,i+1} - R_{2,i+1})\alpha_{i+1} =$$

$$-\frac{1}{2}(R_{6,i} + R_{2,i} - R_{6,i+1} - R_{2,i+1}) + (\bar{\tau}_i + \tau_{i+1}) \tag{14}$$

Phasing Sequence Selection Equations

During optimization, the proper values of Δ_i used in Equation 14 depend on whether NEMA or circular phasing is selected by the optimization program. This means that for a signal, either Constraint 2 or Constraint 11 is active. In order to implement these either/or (disjunctive) constraints, each of them is replaced by two inequality constraints. Because the Δ_i variables are unrestricted (i.e., their values can also be negative), we add a value of 2 to these variables to ensure that they are able to achieve a lower bound of -2. However, this transformation does not change the formulation, because the added values are canceled when Δ_i and Δ_{i+1} are substituted in Loop Constraint 4. Finally, using the binary variable α_i defined earlier, we obtain the following set of constraints:

$$\Delta_i + 2\alpha_i - \delta_i\ell_i + \bar{\delta}_i\bar{\ell}_i \leq \frac{1}{2}(-\ell_i + \bar{\ell}_i) + 4 \tag{15a}$$

$$\Delta_i - 2\alpha_i - \delta_i\ell_i + \bar{\delta}_i\bar{\ell}_i \geq \frac{1}{2}(-\ell_i + \bar{\ell}_i) \tag{15b}$$

$$\Delta_i - 2\alpha_i + (R_{4,i} - R_{8,i})\beta_i \leq \frac{1}{2}(R_{4,i} - R_{8,i}) + \frac{5}{2} \tag{16a}$$

$$\Delta_i + 2\alpha_i + (R_{4,i} - R_{8,i})\beta_i \geq \frac{1}{2}(R_{4,i} - R_{8,i}) + \frac{5}{2} \tag{16b}$$

Note that for simplicity, Movements 1 + 6 are assumed to be outbound movements on the main artery and Movements 4 + 7 are assumed to be outbound movements on the cross artery. Further, this notation is used for consistency with that used in the original MILP formulation presented earlier.

The final comprehensive formulation is as follows:

MILP 2 Find $\underline{b}$, $\overline{b}$, z, w_i, $\overline{w}_i$, t_i, $\bar{t}_i$, δ_i, $\bar{\delta}_i$, α_i, β_i, m_i and Δ_i to maximize $cb + \bar{c}\bar{b}$ subject to

$$(w_i + \overline{w}_i) - (w_{i+1} + \overline{w}_{i+1}) + (t_i + \bar{t}_i) - m_i + \Delta_i + \frac{1}{2}(r_i + \bar{r}_i - R_{6,i} - R_{2,i})\alpha_i - \Delta_{i+1} - \frac{1}{2}(r_{i+1} + \bar{r}_{i+1} - R_{6,i+1} - R_{2,i+1})\alpha_{i+1} = -\frac{1}{2}(R_{6,i} + R_{2,i} - R_{6,i+1} - R_{2,i+1}) + (\bar{\tau}_i + \tau_{i+1}) \qquad i = 1, \ldots, n-1 \tag{14}$$

$$\Delta_i + 2\alpha_i - \delta_i \ell_i + \bar{\delta}_i \bar{\ell}_i \leq \frac{1}{2}(-\ell_i + \bar{\ell}_i) + 4 \qquad i = 1, \ldots, n \tag{15a}$$

$$\Delta_i - 2\alpha_i - \delta_i \ell_i + \bar{\delta}_i \bar{\ell}_i \geq \frac{1}{2}(-\ell_i + \bar{\ell}_i) \qquad i = 1, \ldots, n \tag{15b}$$

$$\Delta_i - 2\alpha_i + (R_{4,i} - R_{8,i})\beta_i \leq \frac{1}{2}(R_{4,i} - R_{8,i}) + \frac{5}{2} \qquad i = 1, \ldots, n \tag{16a}$$

$$\Delta_i + 2\alpha_i + (R_{4,i} - R_{8,i})\beta_i \geq \frac{1}{2}(R_{4,i} - R_{8,i}) + \frac{5}{2} \qquad i = 1, \ldots, n \tag{16b}$$

$$-kb + \bar{b} \begin{cases} = 0 & \text{if } k = 1 \\ \geq 0 & \text{if } k < 1 \\ \leq 0 & \text{if } k > 1 \end{cases} \tag{3}$$

$$\frac{1}{T_2} \leq z \leq \frac{1}{T_1} \tag{4}$$

$$w_i + b + (r_i - R_{6,i})\alpha_i \leq 1 - R_{6,i} \qquad i = 1, \ldots, n \tag{13a}$$

$$\overline{w}_i + \bar{b} + (\bar{r}_i - R_{2,i})\alpha_i \leq 1 - R_{2,i} \qquad i = 1, \ldots, n \tag{13b}$$

$$\left(\frac{d_i}{f_i}\right)z \leq t_i \leq \left(\frac{d_i}{e_i}\right)z \qquad i = 1, \ldots, n-1 \tag{6a}$$

$$\left(\frac{\bar{d}_i}{\bar{f}_i}\right)z \leq \bar{t}_i \leq \left(\frac{\bar{d}_i}{\bar{e}_i}\right)z \qquad i = 1, \ldots, n-1 \tag{6b}$$

$$\left(\frac{d_i}{h_i}\right)z \leq \left(\frac{d_i}{d_{i+1}}\right)t_{i+1} - t_i \leq \left(\frac{d_i}{g_i}\right)z \qquad i = 1, \ldots, n-2 \tag{7a}$$

$$\left(\frac{\bar{d}_i}{\bar{h}_i}\right)z \leq \left(\frac{\bar{d}_i}{\bar{d}_{i+1}}\right)\bar{t}_{i+1} - \bar{t}_i \leq \left(\frac{\bar{d}_i}{\bar{g}_i}\right)z \qquad i = 1, \ldots, n-2 \tag{7b}$$

NEMA Left-Turn Choice Constraints on δ_i and $\bar{\delta}_i$

$$i = 1, \ldots, n$$

b, $\bar{b}$, z, w_i, $\overline{w}_i$, t_j, $\bar{t}_j$, and $m_j \geq 0$

$$i = 1, \ldots, n, j = 1, \ldots, n-1$$

m_i general integer variables

$$i = 1, \ldots, n-1$$

δ_i, $\bar{\delta}_i$, α_i and β_i binary variables

$$i = 1, \ldots, n \tag{8}$$

EXPERIMENTS WITH ENHANCED FORMULATION

The comprehensive formulation presented in the previous section was manually tested using four real-world arterial test problems with 3, 5, 6, and 12 intersections. Constraint 3, which enforces a user-desired relationship between the inbound and outbound bandwidths, was relaxed for this set of experiments. All optimizations were performed on a personal computer using the LINDO optimization package. The optimum time-space diagram for the 12-intersection problem, obtained using the original formulation, is shown in Figure 5. The optimum time-space diagram for the same test problem, obtained using the enhanced formulation, is shown in Figure 6. Table 1 summarizes the results of these two optimization runs on the 12-intersection problem. Following is a description of results for this problem.

1. The enhanced formulation resulted in a cycle length of 78 sec, compared to a value of 71 sec with the original formulation.
2. The enhanced formulation produced larger progression bands with a total increase of 15.04 sec in the total bandwidth. This increase can also be verified by a comparison of bands in terms of percent of cycle length.
3. The enhanced formulation resulted in different phasing sequences on intersections at 0, 3,480, 4,520, 11,050, 12,145, and 12,950 ft. A counterclockwise circular phasing sequence was selected at the intersection at 12,145 ft. For the other intersections, different NEMA sequences were selected.

Optimizing combined NEMA and circular phasing sequences (using MAXBAND 89T) for the problems with three, five, and six intersections produced the same total bandwidths as the optimization of NEMA phasing sequences alone

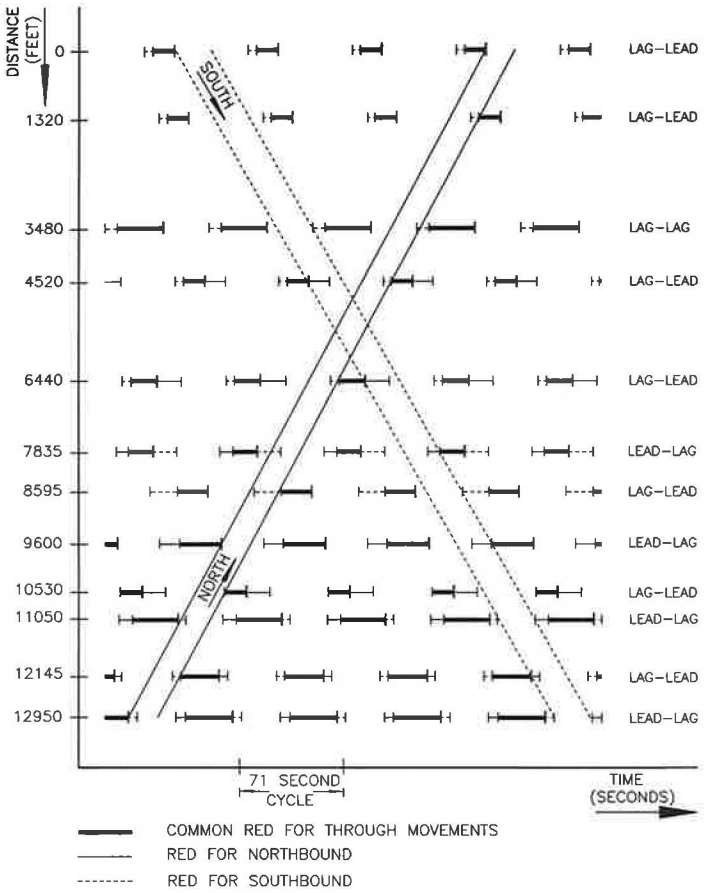


FIGURE 5 NEMA-only phasing sequence optimization for 12-intersection problem.

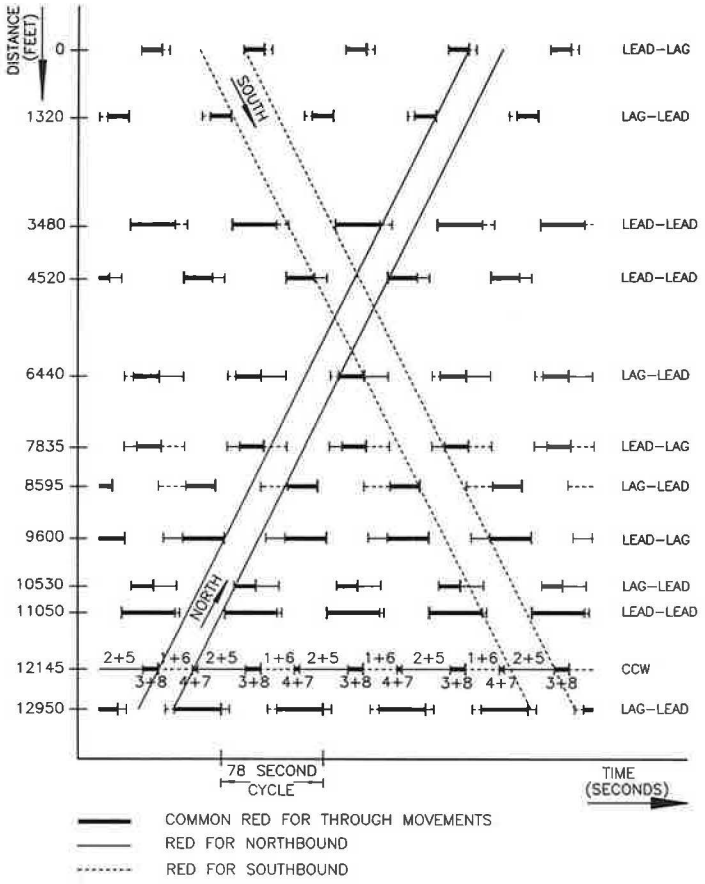


FIGURE 6 Combined NEMA and circular phasing sequence optimization for 12-intersection problem.

TABLE 1 MAXBAND 86 AND MAXBAND 89T RESULTS FOR 12-INTERSECTION PROBLEM

	MAXBAND 86	MAXBAND 89T
Cycle length (Sec)	71	78
Northbound Band		
Seconds (% of Cycle)	19.0 (28.1)	26.0 (33.3)
Southbound Band		
Seconds (% of Cycle)	24.6 (34.5)	33.6 (43.0)
Phasing Sequence		
Signal 1	Lag-Lead	Lead-Lag
Signal 2	Lag-Lead	Lag-Lead
Signal 3	Lag-Lag	Lead-Lead
Signal 4	Lag-Lead	Lead-Lead
Signal 5	Lag-Lead	Lag-Lead
Signal 6	Lead-Lag	Lead-Lag
Signal 7	Lag-Lead	Lag-Lead
Signal 8	Lead-Lag	Lead-Lag
Signal 9	Lag-Lead	Lag-Lead
Signal 10	Lead-Lag	Lead-Lead
Signal 11	Lag-Lead	Counter-Clock Circular
Signal 12	Lead-Lag	Lag-Lead

(MAXBAND 86). These results were not unexpected for the following reasons:

1. The maximum progression bandwidth cannot be greater than the minimum through green time. Thus, the upper limit on total bandwidth is the sum of minimum inbound green and minimum outbound green. For many arterial problems, only NEMA phasing optimization produces this maximum. For these problems, the enhanced formulation will result in the same total bandwidth, even when a circular phasing is selected.
2. Even if the total band produced by NEMA- only optimization is less than the upper limit, the signal spacings, combined with practical travel speeds, may not allow full utilization of the additional green time windows provided by circular phasing sequences.

DEVELOPMENT OF MAXBAND 89T

Because the manual procedure was too tedious and time-consuming, an automated method was developed to perform this task. We accomplished this by modifying the arterial signal-timing optimization capability of MAXBAND 86. The modified program, MAXBAND 89T, is capable of optimizing arterial signal-timing problems only, and allows the selection of best values for cycle length, offset, link travel speeds, and either NEMA or combined NEMA and circular phasing sequences. The following phasing restrictions are programmed in MAXBAND 89T:

1. Circular phasing sequence optimization is not allowed for a three-legged intersection because this is a special case of the conventional NEMA phasing sequence.
2. If circular phasing optimization is desired in addition to the NEMA phasing, at least one of the two signalized approaches on an arterial must have left-turn demand.

The time-space diagram of MAXBAND 89T prints characters 6666, 2222, 4444, and 8888, to indicate Signal Phases 1 + 6, 2 + 5, 4 + 7, and 3 + 8, respectively. The length of a character string indicates the duration of corresponding phase.

EXPERIMENTS WITH ENHANCED FORMULATION USING MAXBAND 89T

Eight real-world arterial problems were used to test MAXBAND 89T on a DecStation 3100 computer. This computer is about two times faster than a Compaq 386/25 personal computer with a math coprocessor (*11*). For these test problems, we used fixed cycle lengths to ensure proper comparison with the original arterial formulation. In addition, we forced the inbound and outbound progression bandwidths to be the same.

Figures 7 and 8 show optimal time-space diagrams for Problem 2 (Ridgewood Avenue) obtained from the two programs. The optimal MAXBAND solution produced bands equal to 28.9 sec in each direction. The travel speeds selected for northbound and southbound directions were equal to 34.5 mph. The phasing sequences selected for Signals 1 through 4 were lag-lead, lead-lag, lead-lag, and lag-lead, respectively. In comparison, MAXBAND 89T produced 37.5-sec bands in each direction, an increase of over 17 sec in the total band. As compared to the solutions from MAXBAND 86, the enhanced program selected higher travel speeds of 37.7 and 37.8 mph for northbound and southbound directions, respectively. The phasing sequences selected by MAXBAND 89T were also different from those selected by MAXBAND 86. It selected counterclockwise circular phasing for the first intersection, and selected lead-lead phasing for the second and third intersections as compared to lag-lead phasing selected by MAXBAND 86.

Table 2 compares the MAXBAND 86 and MAXBAND 89T optimization results for all eight test problems. A description of the overall test results follows:

1. Combined NEMA and circular optimization produced wider progression bands for four out of eight problems. The improvement in bandwidth was from 15.35 to 29.55 percent.
2. Bandwidth improvement in terms of actual time was 3.5, 5.73, 6.51, and 17.09 sec for Problems 2, 5, 7, and 8, respectively.
3. As expected, the computational time required to optimize the enhanced formulation (MAXBAND 89T) increased slightly.

The TRANSYT-7F (*12*) program was used to further analyze the two solutions for Problems 2, 5, 7, and 8. For each optimal solution obtained from MAXBAND 86 and MAXBAND 89T, we performed (a) evaluation of delay, (b) delay optimization without constraining the bands, and (c) bandwidth-constrained delay minimization. For delay op-

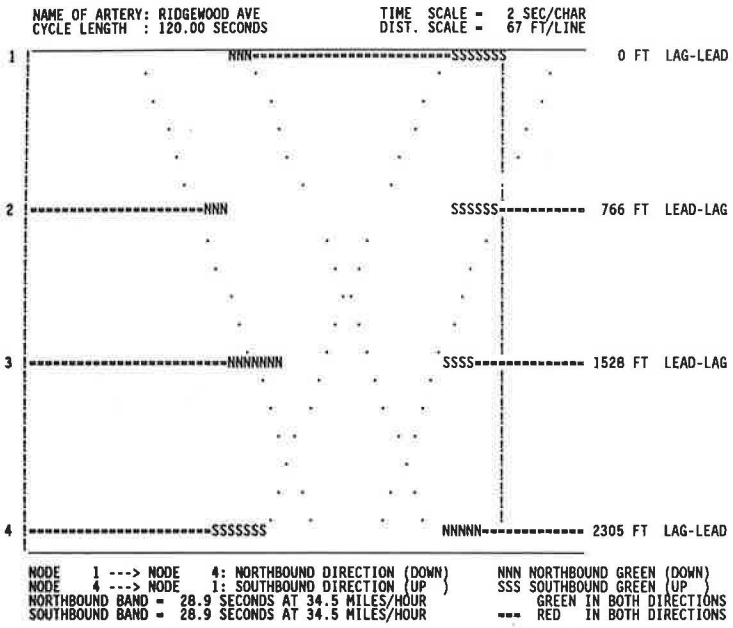


FIGURE 7 MAXBAND 86 time-space diagram for Ridgewood Avenue problem.

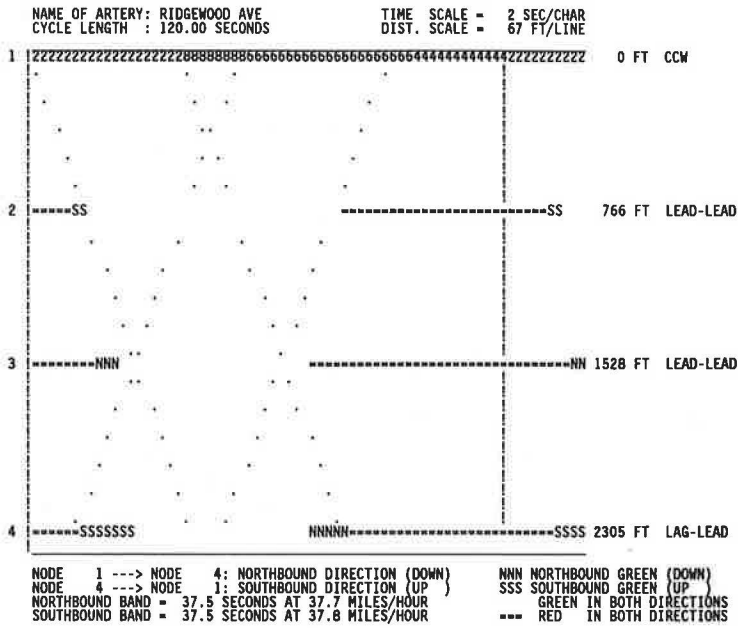


FIGURE 8 MAXBAND 89T time-space diagram for Ridgewood Avenue problem.

timization, we requested a stop penalty that minimized fuel consumption. The results are shown in Table 3. Following is a summary of these additional TRANSYT-7F studies:

1. For Problem 8 (Fannin), combined NEMA and circular phasing optimization (MAXBAND 89T) resulted in smaller values of delay, stops, and fuel consumption. For the other three problems, these measures were worse than those produced by optimizing NEMA sequences alone (MAXBAND 86).

2. For all cases, the use of an optimal bandwidth solution as a starting point for TRANSYT optimization resulted in reduced delay, stops, and fuel consumption. However, minimization without constraining the bands was always better than the constrained bandwidth option.

SUMMARY

We present an enhanced MILP arterial formulation that produces maximal progression bands by finding the best cycle

TABLE 2 COMPARISON OF MAXBAND 86 AND MAXBAND 89T OPTIMIZATION RESULTS

No	Problem Name	No. of Signals	Cycle Length (Sec)	Original $b = \bar{b}$ (Sec)	Original CPU Time (Sec)	Original+Circular $b = \bar{b}$ (Sec)	Original+Circular CPU Time (Sec)	Bandwidth Increase (Sec)	Bandwidth Increase (%)
1.	Washington	4	80	29.71	1.4	29.71	2.5	--	--
2.	Ridgewood	4	120	28.92	1.6	37.47	2.0	17.09	29.55
3.	Fourth Street	5	100	11.19	1.1	11.19	1.6	--	--
4.	M Street	8	80	29.30	1.6	29.30	2.1	--	--
5.	N 33rd	9	75	10.07	25.4	11.82	25.3	3.50	17.38
6.	Nicholasville	12	80	27.06	88.8	27.06	47.8	--	--
7.	N Michigan	13	90	18.68	22.5	21.54	25.8	5.73	15.35
8.	Fannin	15	80	20.87	30.7	24.13	165.9	6.51	15.59

-- indicate no difference in the total bandwidth

TABLE 3 DELAY COMPARISON OF SOLUTIONS FROM MAXBAND 86 AND MAXBAND 89T

Test Problem	TRANSYT Option	Total Delay (Veh-hr/hr)	Ave. Delay (sec/veh)	Stops (veh/hr)	Fuel Cons. (gal/hr)
Ridgewood (N)	1	79.00	25.40	7426.0	144.00
	2	75.00	24.10	7101.0	139.00
	3	75.00	24.10	7101.0	139.00
Ridgewood (N+C)	1	84.00	27.10	6694.0	145.00
	2	80.00	26.00	6466.0	140.00
	3	81.00	26.10	6459.0	141.00
N 33rd (N)	1	243.27	26.01	24119.4	552.92
	2	221.90	23.72	23621.0	532.90
	3	225.66	24.12	23314.4	533.21
N 33rd (N+C)	1	289.44	30.94	22337.8	571.24
	2	259.53	27.74	22613.0	551.46
	3	263.97	28.22	22293.9	552.26
N Michigan (N)	1	182.98	15.15	16687.6	362.28
	2	152.04	12.58	15172.7	331.00
	3	154.43	12.78	15766.4	336.15
N Michigan (N+C)	1	249.60	20.66	16289.5	408.83
	2	186.78	15.46	15015.0	355.00
	3	258.00	21.40	15135.0	409.00
Fannin (N)	1	150.95	12.47	19224.2	458.56
	2	140.78	11.63	16984.3	433.48
	3	140.10	11.57	17429.4	436.34
Fannin (N+C)	1	146.76	12.12	18842.5	452.27
	2	137.86	11.39	17434.6	434.70
	3	137.60	11.36	18018.1	439.03

(N) - NEMA optimization using MAXBAND 86

(N+C) - NEMA+Circular optimization using MAXBAND 89T

TRANSYT Options (1) Evaluation, (2) Unconstrained Optimization, (3) Bandwidth Constrained Optimization

length, offsets, and either NEMA or circular phasing sequences. We also present results of computational experience with MAXBAND 89T, which allows easy use of the enhanced formulation. In addition, the results from MAXBAND 86 and MAXBAND 89T were further analyzed using TRANSYT-7F. Although experiments were performed using a limited number of test problems, the results demonstrate that, in some cases, combined NEMA and circular phasing sequence optimization may produce wider progression bands than NEMA phasing optimization alone. In addition, arterial performance may improve as a result of the improved bandwidth solutions. The results can be summarized as follows:

1. Combined NEMA and circular phasing sequence optimization can produce wider arterial progression bands in some cases.
2. In some cases, selection of circular phasing may produce lower total vehicular delay, depending on factors such as demand on approaches.
3. As expected, the CPU time for combined NEMA and circular phasing optimization is, in general, slightly more than that for NEMA phasing optimization only.

The results demonstrate that for some cases, MAXBAND 89T can produce solutions with improved progression bands and arterial performance.

CONCLUSIONS AND RECOMMENDATIONS

These results are based on a small set of test problems. However, the results show that combined NEMA and circular phasing sequences can produce improved signal timings; therefore, optimization using the circular phasing sequence should not be excluded from the choices examined.

Further research is needed to more fully understand the advantages and disadvantages of the circular phasing optimization. Some of the questions that need to be answered are

1. What is the effect of circular phasing selection on the cross street? On pedestrian traffic?

2. What causes the total delay to be increased or decreased?
3. Under what signal conditions (i.e., demand, signal spacings) can maximum benefits be obtained from circular phasing?
4. How would drivers react when they encounter a signal with the nontraditional circular phasing sequence?
5. What type of results would be obtained from using this capability in the other recent nonuniform arterial bandwidth models (*4*,*5*)?

ACKNOWLEDGMENTS

The material presented here is based, in part, upon research funded by the Texas State Department of Highways and Public Transportation under the Texas Advanced Transportation Technology Project. The data sets used were supplied by Stephen L. Cohen of FHWA. The authors would like to thank him for assisting this work. The authors would also like to thank the anonymous referees for providing useful comments that helped to improve this paper.

REFERENCES

1. J. D. C. Little. The Synchronization of Traffic Signals by Mixed-Integer Linear Programming. *Operations Research*, Vol. 14, 1966, pp. 568–594.
2. J. D. C. Little, M. D. Kelson, and N. H. Gartner. MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks. In *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981, pp. 40–46.
3. E. C. Chang, S. L. Cohen, C. Liu, C. J. Messer, and N. A. Chaudhary. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequence in Multiarterial Closed Networks. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1989, pp. 61–67.
4. H. S. Tsay and L. J. Lin. A New Algorithm for Solving the Maximum Progression Bandwidth. Presented at the 67th Annual Meeting of the Transportation Research Board, Washington D.C., Jan. 1988.
5. N. H. Gartner, S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND—A Variable-Bandwidth Arterial Progression Scheme. In *Transportation Research Record 1287*, TRB, Presented at the 69th Annual Meeting, National Research Council, Washington, D.C., 1990.
6. N. A. Chaudhary, C. J. Messer, and A. Pinnoi. Efficiency of Mixed Integer Linear Programs for Traffic Signal Synchronization Problems. *Proc., 25th Annual SE TIMS Meeting*, Oct. 1989, pp. 155–157.
7. P. Mireault. Solving the Single Artery Traffic Signal Synchronization with Benders Decomposition. Presented at CORS/ORSA/TIMS Joint National Meeting, Vancouver, Canada, May 8–10, 1990.
8. P. Mireault. *An Integer Programming Approach To The Traffic Signal Synchronization Problem*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Feb. 1988.
9. F. V. Webster. *Traffic Signal Settings*. Road Research Technical Paper 39. Her Majesty's Stationery Office, London, England, 1958.
10. L. Schrage. *User's Manual for LINDO*, 3rd ed. The Scientific Press, Redwood City, Calif., 1987.
11. P. Magney. DECstation 3100: A Leader. *Computer Reseller News*, Sept. 4, 1989, pp. 57–58.
12. C. E. Wallace, K. G. Courage, D. P. Reaves, G. W. Schoene, G. W. Euler, and A. Wilbur. *TRANSYT-7F User's Manual*. FHWA, U.S. Department of Transportation, 1988.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

TRANSYT-7F or PASSER II, Which Is Better—A Comparison Through Field Studies

SHUI-YING WONG

Several studies have compared the arterial signal timings optimized by TRANSYT-7F and PASSER II. The comparisons, however, were based on simulated results. In this study, the TRANSYT-7F timing plans were compared with the PASSER II timing plans based on operational characteristics, field results, and simulated results. These comparisons were possible because (a) the signals on two arterials in San Francisco were optimized by TRANSYT-7F and implemented in 1987, (b) the same signals were retimed by PASSER II and implemented in 1988, and (c) before-and-after studies were conducted. From the field results, the overall effectiveness of TRANSYT-7F and PASSER II was about the same in terms of travel time and stops along the arterial (excluding cross streets). On one of the arterials, the offset pattern and operational characteristics of the TRANSYT timing were very different from those of the PASSER timing; on the other arterial, they were very similar. The TRANSYT-7F simulated travel times were reasonably close to the field travel times. However, the simulated measures of effectiveness in general were inclined in favor of the timing plans optimized by TRANSYT-7F. The field data for travel time were reliable and easy to collect. Statistically, one to five samples were required to attain a 95 percent level of confidence for the example arterials, each with 30 or more signalized intersections.

TRANSYT-7F and PASSER II are two popular programs for signal timing. TRANSYT-7F optimizes signals by minimizing vehicle delay and stops to all approaches, and PASSER II optimizes signals by maximizing the bandwidth along the arterial. Several studies have compared TRANSYT-7F, PASSER II, and other bandwidth programs. Skabardonis and May (*1*) compared the arterial signal timings optimized by TRANSYT-7F, PASSER II, and MAXBAND and found that TRANSYT-7F produced the best result in terms of a performance index (a combination of delay and stops expressed as a number). Cohen (*2*) compared the arterial signal timings optimized by TRANSYT-7F and MAXBAND and found that TRANSYT-7F produced a better result in terms of delay and stops. Liu (*3*) compared the arterial signal timings optimized by TRANSYT-7F without bandwidth constraint, TRANSYT-7F with bandwidth constraint, MAXBAND, and PASSER II. He found that TRANSYT-7F without bandwidth constraint produced the best result in terms of delay and stops.

Although these findings suggest that TRANSYT-7F produces better results, many traffic engineers prefer PASSER II. A recent survey of traffic engineers indicated that 63 percent used PASSER II and 26 percent used TRANSYT-7F to analyze coordinated signalized intersections; however, the same survey indicated that 93 percent of them obtained information about alternative traffic computer programs through literature (*4*). It appeared that the findings did not convince many traffic engineers. One reason may be that PASSER II is easier to use and provides visible and verifiable progression along the arterial. Because signal progression is readily perceived along the arterial, complaints from the public are minimized. Another reason may be that traffic engineers have reservations about findings based on simulation results. Skabardonis and May's findings were based on the TRANSYT-7F simulated results. Cohen's and Liu's findings were based on the TRANSYT-7F and NETSIM simulated results. However, simulation has its merits. Different scenarios can be easily analyzed at minimal cost. Furthermore, it is difficult to implement the signal timing plans from different models on the same arterial for the sake of comparison. However, simulation may be different from what is actually happening in the field. Comparison through field studies may provide better insight.

The signals on two arterials, Geary Boulevard and 19th Avenue, in San Francisco were optimized by TRANSYT-7F and implemented in January 1987 as part of the California Department of Transportation's Fuel Efficient Traffic Signal Management (FETSIM) program. The original offsets on both arterials were set manually with a double alternate pattern. Although the TRANSYT-7F timings resulted in annual savings of $3.5 million (based on simulation results) (*5*), about 10 complaints on 19th Avenue were received during the first 3 months after implementation.. About 10 to 15 complaints on the same arterial were received during the next 15 months. The majority of complaints were that the northbound (a.m. inbound) progression was bad. In response to the complaints, the a.m. timing on 19th Avenue and the p.m. timing on Geary Boulevard were retimed using PASSER II and implemented in June 1988. After the signals were changed to PASSER timings, one response was received during the first 3 months. The response complimented the good progression in northbound 19th Avenue and urged the same in southbound. There was no response on Geary Boulevard during the TRANSYT-7F nor the PASSER timings.

Although these user responses were not a scientific sampling, they did represent some users' perceptions. Because timings from both TRANSYT-7F and PASSER II were being implemented on the same arterials, we had the opportunity to find out whether TRANSYT-7F or PASSER II is really better. We conducted before-and-after field studies by com-

Office of Traffic Operations and IVHS, FHWA, HTV-32, 400 7th St., Washington, D.C. 20590.

paring the operational characteristics and the field results under both timings. Because collecting field data is usually time-consuming, the comparison described how and what data were collected, the required sample size to achieve a 95 percent level of confidence, and the statistical test to compare the data in order to get some idea of how much effort is involved. Because many findings were based on the simulated results of TRANSYT-7F, as mentioned earlier, field results were also compared to TRANSYT-7F simulated results to see if there were discrepancies.

The signal timing plans that were implemented to the arterials were developed from TRANSYT-7F, Release 4, and PASSER II-84. The model-simulated results were from TRANSYT-7F, Release 6.

STUDY ARTERIALS

The study arterials were (a) Geary Boulevard with 30 signals and (b) 19th Avenue and Park Presidio Boulevard (referred to as 19th Avenue) with 33 signals (see Figure 1). The signals are fixed-time. Geary Boulevard is a two-way street with curb parking, left-turn pockets, and three lanes per direction. There are retail stores and parking activity is heavy. 19th Avenue

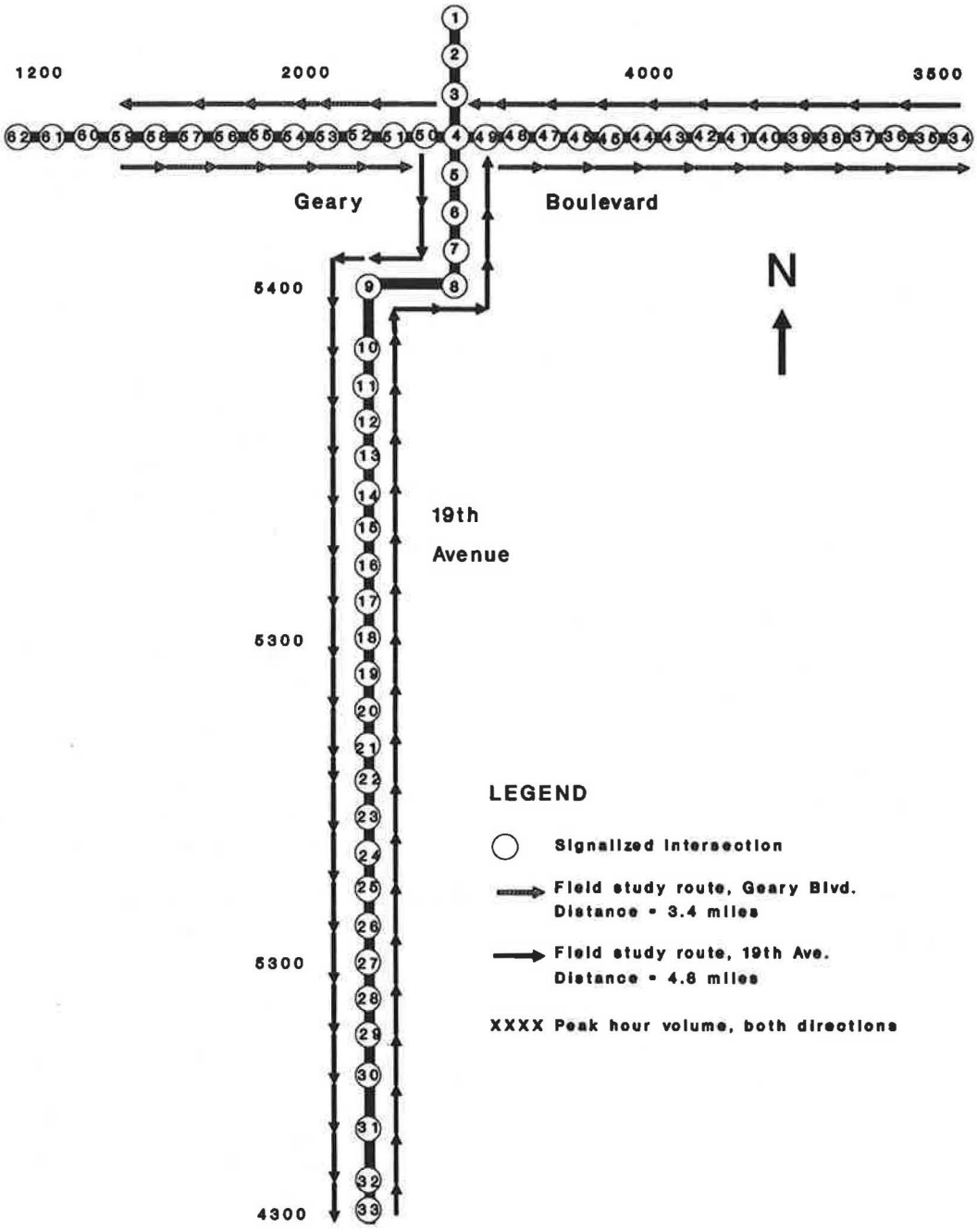


FIGURE 1 Study arterials.

is also a two-way street with three lanes per direction. Left turns are allowed in only a few intersections. Portions of the street have curb parking; however, there are no retail stores and parking activity is insignificant. Figure 1 shows the peak-hour traffic volumes.

Each arterial had three timing plans per day. We chose the p.m. timing plan on Geary Boulevard and the a.m. timing plan on 19th Avenue for comparison because they are the most critical.

TRANSYT-7F TIMINGS

In developing the timing plans, turning-movement counts were collected at each signalized intersection. We went through the processes of model calibration, cycle length selection, optimization, and fine-tuning (*5*).

During model calibration, we selected five key intersections from each arterial. For each selected intersection, we (a) recorded the major platoon arrivals and compared them with TRANSYT's flow profiles, and (b) observed the queue lengths and compared them with TRANSYT's maximum back of queues. From simulation runs, links with at least 95 percent degree of saturation were checked in the field to see if they were congested.

During cycle length selection, we made runs with cycle lengths between 65 and 95 sec. We included runs with double cycles on selected intersections. We selected 85 sec as the cycle length on both arterials based on minimum fuel consumption.

During optimization, we were concerned that the offsets on 19th Avenue were too close to a simultaneous pattern. We therefore explored the following options: (a) performing normal optimization; (b) applying delay and stop weights to links along the arterial; (c) first using PASSER II to optimize the offsets, then inputting the resulting offsets to TRANSYT-7F for optimization of both offsets and splits; and (d) modifying TRANSYT's hill-climb steps to emphasize offset optimization. The resulting offsets from these options were similar. Therefore the normal optimization option was used for both arterials.

During fine-tuning, we drove through the arterials to check any abnormal stops or delay. Several offsets and splits were modified based on field checks.

We continued to make minor adjustments until May 1988. Minor adjustments were necessary because the input coding, from which the signal timing was obtained, might not represent 100 percent of the field conditions. TRANSYT-7F was a versatile tool for fine-tuning. If the signal timing of a particular intersection is not working properly, one can (a) change the split or offset of any affected intersection, then resimulate the changed part along with the rest of the network; (b) update the input data and reoptimize the offsets and splits of the affected intersections while the rest of the network remains fixed; (c) update the input data of affected intersections and reoptimize the whole network; or (d) use any combination of these options. This flexibility allows improving localized drawbacks while preserving system-wide efficiency.

Figures 2 and 3 show the time-space diagrams for TRANSYT-7F timings. These timings were completed after the final adjustments and were the ones under which the field studies were conducted.

PASSER II TIMINGS

Although PASSER II can optimize cycles, splits, offsets, and phase sequences, we optimized only the offsets because (a) we were retiming the same signals on the same arterial; (b) the roadway widths on both arterials were wide, and the minimum green time on cross streets (for pedestrians walking across the arterial) were long enough for traffic volumes on cross streets; and (c) the signals were predominately two-phased. However, PASSER II cannot optimize offsets only. To prevent splits from varying, we coded the minimum green equal to the existing green plus yellow and all-red times for each phase. We used the total directional volume as the bandwidth split. On 19th Avenue, however, because of motorists' complaints, we used a 65 percent bandwidth split to favor the northbound traffic, even though the total flow in this direction was 52 percent.

PASSER II can optimize up to 20 intersections per run. However, the arterials used had 30 and 33 intersections. Instead of arbitrarily dividing the arterial into two runs, intersections with similar volumes were grouped into segments. Each segment was optimized with a separate directional bandwidth split. After optimization, we manually aligned the through bands from each segment so that there was a continuous through band in the major flow direction while as much smooth flow as possible was maintained in the reverse direction. The resulting timing has the following characteristics:

1. The through bands on both directions of each segment are wider than those of arterials that are not segmented. The fewer the number of intersections, the wider the through bands, because fewer intersections means less constraints for PASSER II to maximize.
2. Having wider through bands within each segment means traffic has better progression and fewer stops within the segment, and the segment boundaries become the scheduled stopping points. That is, if a vehicle can pass the boundary intersection, it will not have to stop until the next boundary intersection.

During the first 2 months after implementation, we made minor adjustments to the offsets of intersections both within the same segment and between different segments. Adjusting a few seconds of offsets, at the northbound approach to Lincoln Avenue on 19th Avenue (Intersection 11 in Figure 1), for instance, remedied the spillback. However, PASSER II was not a good tool for minor adjustments. We could not freeze the timings on certain intersections while optimizing the others (although one can freeze the phase lengths and splits by coding the minimum greens, one cannot freeze the offsets). When we reran PASSER II by changing the queue clearance time on one or two intersections to try to avoid the spillback that was observed in the field, for instance, we got different offsets on almost all intersections. Because all of the timings had been set in the field, it was impractical to change all of them to correspond to PASSER II's optimal timing. Whenever we made adjustments, we still reran PASSER II to get some ideas and made the adjustments manually.

Figures 4 and 5 show the time-space diagrams for the PASSER II timings. These timings were completed after the minor adjustments and were the ones under which the field studies were conducted.

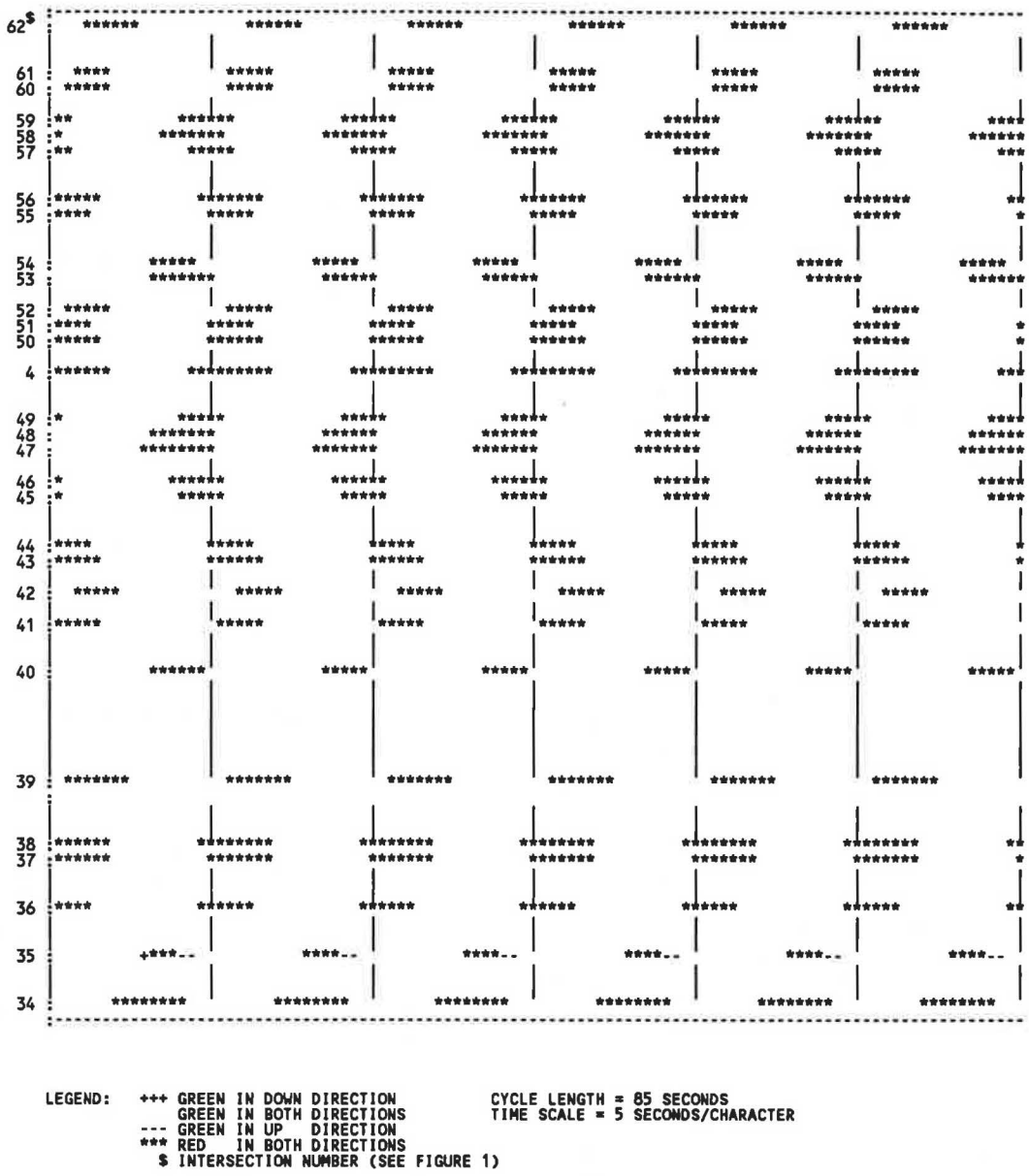


FIGURE 2 Geary Boulevard time-space diagram, from TRANSYT-7F.

FIELD STUDIES

The field studies were on four routes (Geary eastbound, Geary westbound, 19th Avenue northbound, and 19th Avenue southbound), as shown in Figure 1. Field studies under the TRANSYT timings were conducted in June 1988, and those under the PASSER timings were conducted in October 1988. Although the before and after studies were several months apart, the flow pattern would probably remain the same because (a) there were no major changes in land use along the study routes; (b) the study periods (a.m. and p.m. peak hours) were the commuting periods, and commuting traffic is usually not sensitive to monthly or seasonal changes except during major holiday periods; and (c) June and October are not major holiday periods for commuters. To minimize the variation of the before and after field data, the field studies were conducted along the same routers, during the same peak hours, and by the same driver and recorder during the TRANSYT and PASSER timings. Furthermore, we defined the data collection method precisely, especially in defining the number of stops. The number of stops was defined as follows:

- A stop occurred whenever the test vehicle was motionless for 3 sec or more. This avoided the ambiguity of minor stop-and-go situations.
- Only one stop was counted within the same phase on the same approach, even if there were two or more legitimate motionless periods. This avoided counting a stop more than once due to temporary lane obstruction, lane changing, or turning right on red by the preceding vehicles.

We wanted to collect travel time, stopped delay, and number of stops data. From previous studies on the same arterials (*5*), however, stopped delay data were not reliable because

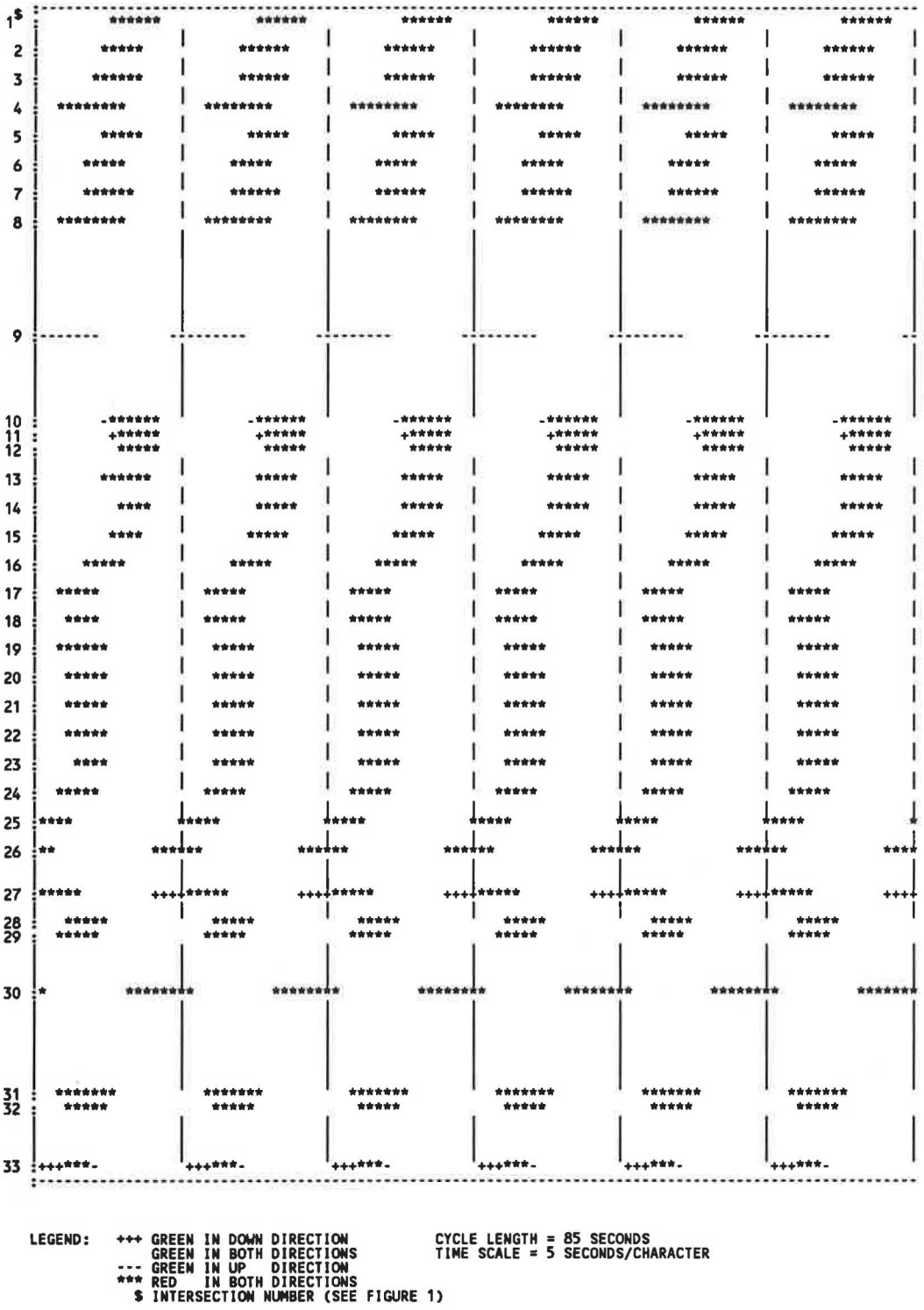


FIGURE 3 19th Avenue time-space diagram, from TRANSYT-7F.

of the ambiguities of slow-moving or stop-and-go situations. Stopped delay times varied so much that the sample size would have to be over 20 to attain a 95 percent level of confidence. We therefore ignored delay and concentrated on getting reliable travel time and stop data.

To determine the sample size, we applied the following equation (*6,7*):

$$N = (KS/E)^2 \quad (1)$$

where

N = number of samples,
K = 1.96 for a 95 percent level of confidence,
S = standard deviation, and
E = tolerable error, equals 1 min per route distance for

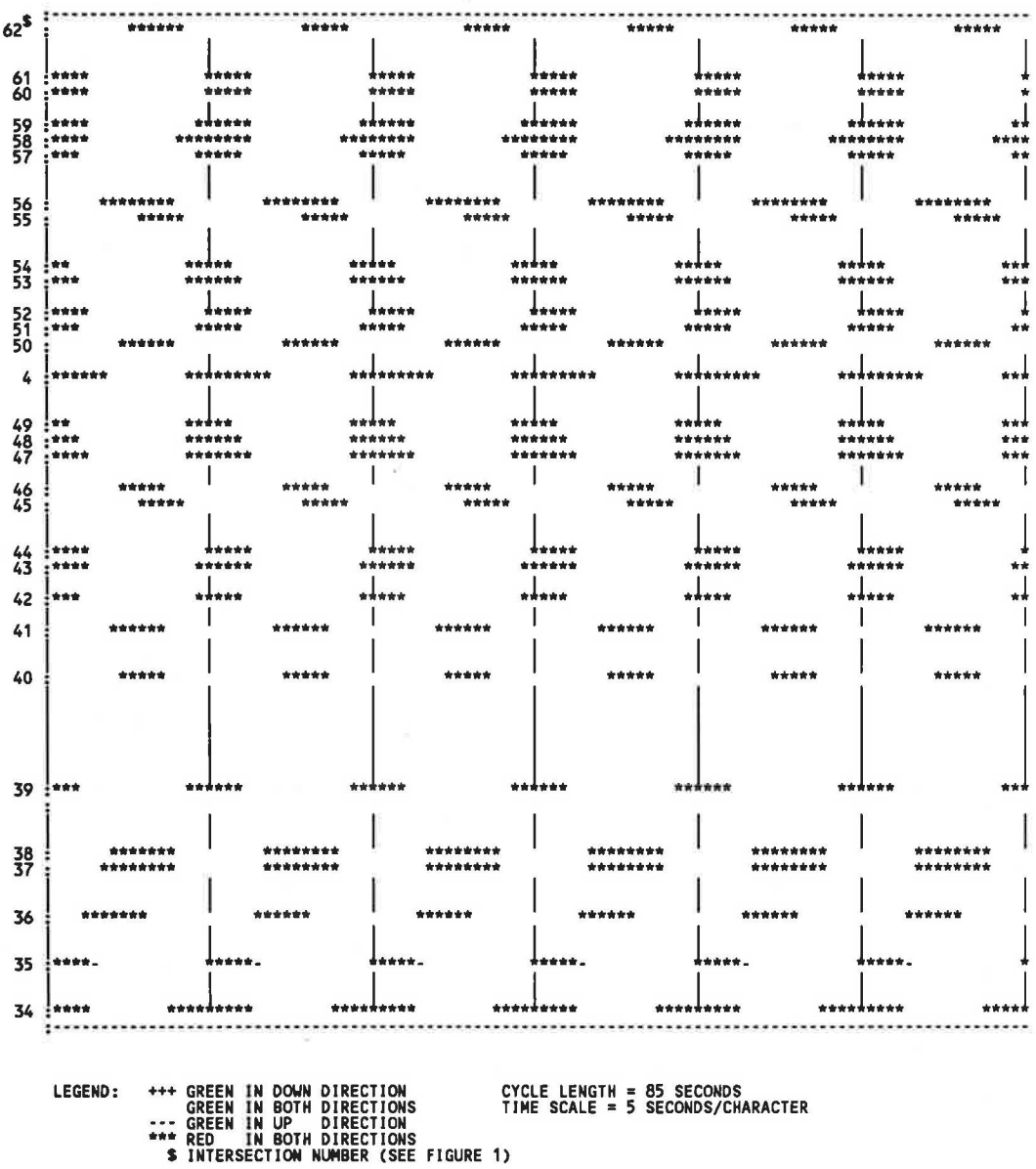


FIGURE 4 Geary Boulevard time-space diagram, from PASSER II.

travel time and 1 stop per route distance for number of stops (route distances on Geary Boulevard and 19th Avenue were 3.4 and 4.8 mi, respectively).

After collecting three samples, we computed the standard deviation and applied Equation 1 to estimate the sample sizes for travel time and stops for each route. We repeated the process after each additional run until the number of field samples was equal to or greater than the computed sample size for travel time. We conducted additional samples to satisfy the computed sample size for stops, if possible.

The last lines on Tables 1 and 2 show that the sample size required to attain a 95 percent level of confidence for travel time ranged from 1 to 5 and that for stops ranged from 2 to 25. Hence, travel time requires less effort. The results also show that a street with less traffic friction requires fewer samples. For example, on 19th Avenue, where there were few left turns and parking activities, the required sample size for travel time ranged from one to five and that for stops ranged from two to seven. On Geary Boulevard, where there were heavy left turns and parking activities, the required sample size for travel time ranged from 3 to 5 and that for stops ranged from 4 to 25.

COMPARISON OF TIMING PLANS AND OPERATIONAL CHARACTERISTICS

The TRANSYT and PASSER timing plans on 19th Avenue were different (see Figures 3 and 5). The offsets by TRANSYT were mostly simultaneous and those by PASSER were double and triple alternates. Through our field observation, the PASSER timing plan had the following characteristics:

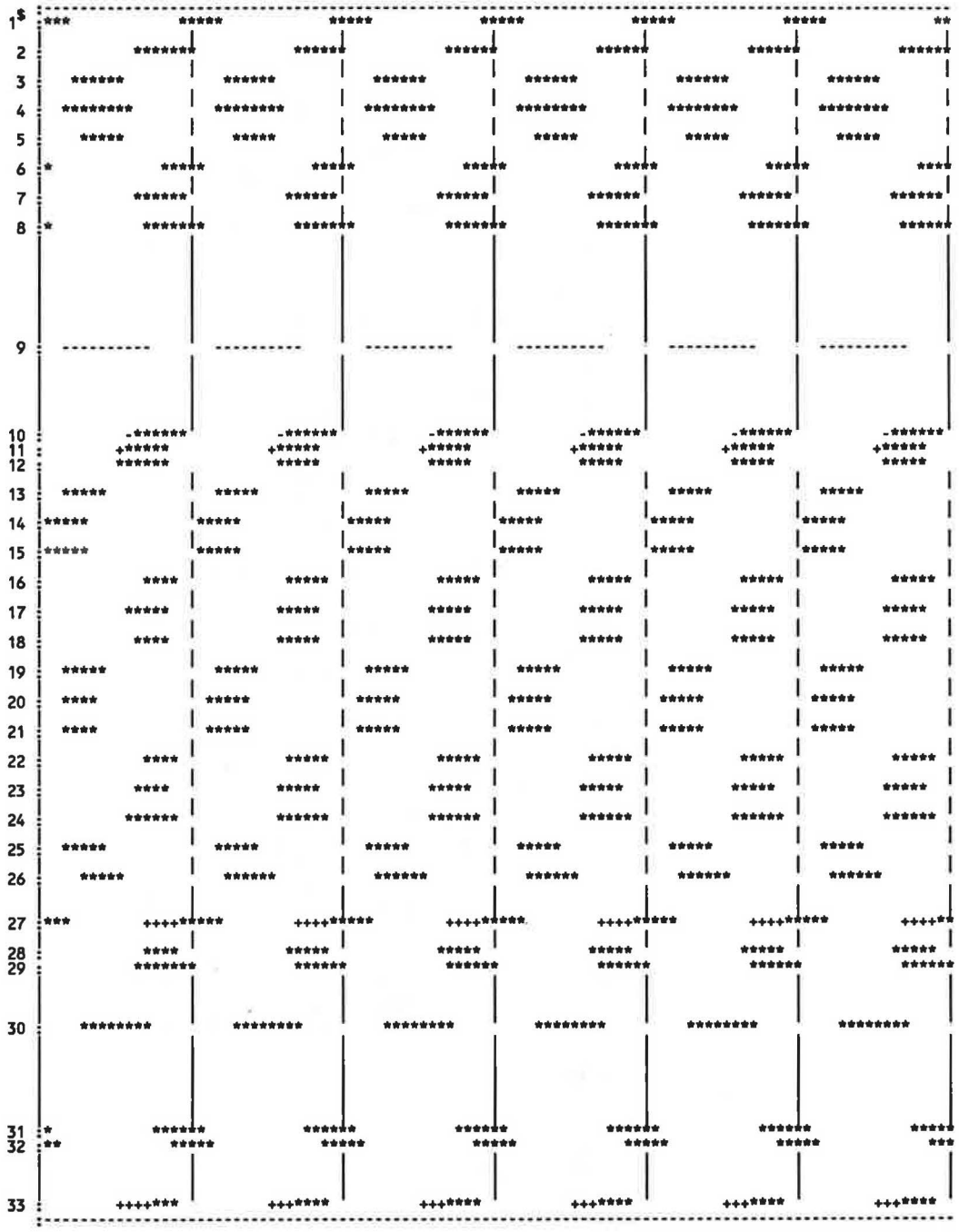


FIGURE 5 19th Avenue time-space diagram, from PASSER II.

1. At the start of green, if our test vehicle was not within the through band, it would hit the red signal at the next one or two intersections (because of double or triple alternating of offsets). After that, we would join the platoon of vehicles within the through band and would go through many intersections without stopping. The platoon of vehicles within the through band became larger as more vehicles joined. The more intersections the platoon of vehicles could go through, the more vehicles would accumulate behind it. Soon the length of the moving platoon became so long that it would oversaturate the green signal. That is, vehicles at the front of the platoon would arrive at the beginning of the green signal and

TABLE 1 FIELD RESULTS UNDER TRANSYT-7F TIMING

	Geary boulevard				19th Avenue/Park Presideo Boulevard			
	Eastbound		Westbound		Northbound		Southbound	
Test Run No.	Travel Time (Min/Mi)	No. of Stops (Per Mi)	Travel Time (Min/Mi)	No. of Stops (Per Mi)	Travel Time (Min/Mi)	No. of Stops (Per Mi)	Travel Time (Min/Mi)	No. of Stops (Per Mi
1	2.89	1.46	3.94	2.92	2.84	1.67	2.59	1.46
2	3.48	2.92	3.08	2.04	2.61	1.25	2.72	1.67
3	2.88	1.17	3.45	2.33	2.77	1.67	2.55	1.46
4	2.54	0.58	3.13	1.75	2.65	1.67	2.61	1.67
5	2.97	1.17	3.39	2.04	2.69	1.67	2.48	1.25
6	2.92	1.17	3.47	2.33	2.62	1.46	2.53	1.46
7	2.90	1.46	3.48	2.33	2.85	1.67	2.65	1.46
8	3.14	2.62	3.41	2.04	2.65	1.67	2.65	1.46
9	3.19	1.46	3.07	2.04	2.57	1.67	-	-
10	-	-	3.48	2.04	2.63	1.46	-	-
Average	2.99	1.55	3.39	2.19	2.69	1.59	2.60	1.49
Req'd Runs[1]	3	25	3	4	1	2	1	2

[1]Required sample size to attain 95% confidence level, computed from equation 1.

TABLE 2 FIELD RESULTS UNDER PASSER II TIMING

	Geary boulevard				19th Avenue/Park Presideo Boulevard			
	Eastbound		Westbound		Northbound		Southbound	
Test Run No.	Travel Time (Min/Mi)	No. of Stops (Per Mi)	Travel Time (Min/Mi)	No. of Stops (Per Mi)	Travel Time (Min/Mi)	No. of Stops (Per Mi)	Travel Time (Min/Mi)	No. of Stops (Per Mi)
1	2.89	1.75	2.60	0.87	2.22	0.42	2.70	1.67
2	3.02	2.33	2.55	0.87	2.64	0.84	3.07	2.30
3	3.29	2.04	3.74	2.33	2.39	1.04	3.35	2.30
4	3.03	1.46	3.33	1.75	2.37	1.04	3.15	2.09
5	3.04	1.75	2.97	1.17	2.30	0.84	3.17	2.51
6	2.85	1.75	2.92	1.17	2.75	1.25	3.27	2.30
7	3.15	2.04	3.42	2.33	3.23*	1.88*	3.81*	2.30*
8	2.47	1.17	3.27	2.04	4.04*	1.88*	3.50*	2.51*
9	3.18	2.33	3.00	1.17	-	-	-	-
10	2.96	1.46	2.59	0.58	-	-	-	-
11	3.22	2.62	3.37	2.04	-	-	-	-
12	3.22	2.33	2.90	2.04	-	-	-	-
12	3.39	2.04	2.93	2.04	-	-	-	-
14	3.34	2.04	2.87	2.04	-	-	-	-
15	3.36	2.04	3.27	1.75	-	-	-	-
Average	3.09	1.94	3.05	1.61	2.44	0.90	3.12	2.19
Req'd Runs[1]	3	7	5	15	4	7	5	7

[1]Required sample size to attain 95% confidence level, computed from equation 1.
*During foggy weather, not included in the averages and other statistical calculations.

would go through without stopping, but vehicles at the back of the platoon would arrive at the same approach beyond the green signal and would have to stop for the red signal.

2. At the start of green, if our test vehicle was within the through band and was the leading vehicle, it could theoretically go through all of the intersections without stopping. However, in a heavy traffic situation such as 19th Avenue, we could not do so because even if we maintained a speed matching the design speed of the through band, we would join the back of another moving platoon after going through about 10 intersections. This "other" moving platoon was from the through band of the previous cycle. Once we joined the back of this other moving platoon, we would no longer be the leading vehicle and would have difficulty maintaining a constant speed (because of frictions from preceding vehicles). We would stop sooner or later because the vehicle at the front

of this other moving platoon would oversaturate the green signal, as mentioned in Point 1.

3. We encountered midblock stops or stop-and-go situations quite often, probably because of the long platoon. However, most of these stops did not fit our definition of stops and were not counted in our field data.

4. Such timing appeared to encourage motorists to travel at the design speed because they would get the best progression at this speed. However, it appeared that it would also encourage motorists to go through yellow signals, because if they could pass the yellow signal, they could remain within the through band and pass many intersections without stopping.

5. During foggy weather, the progression became very bad (see Run Numbers 7 and 8, Table 2). Motorists became cautious and drove slower than the design speed, so, a vehicle originally within the through band would be out of the through band, or "out of sync," after passing a few intersections. Once out of the through band, the vehicle would have to stop once or twice before returning to the through band.

The TRANSYT-7F timing on 19th Avenue, by comparison, did not provide a through band. Our test vehicle would stop after going through several intersections, no matter when we started during the cycle. This was probably because of the simultaneous offsets. Because each vehicle would stop after going through several intersections, the platoons were in small bundles rather than in long queues. There were fewer midblock stops and stop-and-go situations. This is probably because the platoons were in small bundles and the queues were shorter. This timing plan appeared to encourage speeding because the higher the speed, the more intersections the vehicle could go through. However, it appeared that one would not be encouraged to go through yellow signals, because passing one intersection during yellow would not necessarily have the advantage of passing the next intersection. Although we did not experience foggy weather during field studies, we expect the progression would not be as dramatically changed as that of the PASSER timing because the platoons of vehicles were in smaller bundles.

TRANSYT-7F, Release 6, has a link-to-link flow weighting feature (*8*, p. 4-52) which can be used to encourage progression along the arterial. Although this feature was not available during the 1987 project, we subsequently applied it to the 19th Avenue data set to see how the offsets would have been different. We used link-to-link weights along 19th Avenue for (a) both directions and (b) northbound only. Figures 6 and 7 show the time-space diagrams, which were similar to the one in Figure 3. Hence, even if we had applied this new feature to our TRANSYT-7F timings in 1987, the resulting progressions would have been similar. The phenomena described above would still have been true.

The TRANSYT and PASSER offset patterns on Geary Boulevard were about the same. Their operational characteristics were also similar.

COMPARISON OF FIELD RESULTS

To compare whether the changes were statistically significant, we computed the t-statistic as follows (*9*, p. 294; *10*):

$$T_{(\alpha/2, Nt+Np-2)} = \frac{X_t - X_p}{\sqrt{\left[\frac{S_t^2(N_t - 1) + S_p^2(N_p - 1)}{N_t + N_p - 2}\right]\left(\frac{1}{N_t} + \frac{1}{N_p}\right)}} \quad (2)$$

where

T = computed t-statistic, with $(1 - \alpha)$ percent level of confidence and $(N_t + N_p - 2)$ degrees of freedom;
α = 0.05;
X = mean value;
S_t = standard deviation, TRANSYT timings;
S_p = standard deviation, PASSER timings;
N_t = number of samples, TRANSYT timings; and
N_p = number of samples, PASSER timings.

If the absolute value of the computed t-statistic is less than the corresponding critical value of the t-distribution, the change is not significant.

Equation 2 assumes that the data are normally distributed and that the variance of the data from TRANSYT timing is equal to the variance of the data from PASSER timing. To test whether the data are normally distributed, we applied the Kolmogorov-Smirnov test, as follows (*9*, p. 533):

$$D = \max|F_i - S_i| \quad (3)$$

where

D = Kolmogorov-Smirnov test statistic,
F_i = cumulative frequency of the ith category from normal distribution, and
S_i = cumulative frequency of the ith category from field data.

The Kolmogorov-Smirnov test (at a 95 percent level of confidence) indicated that each of the 16 data sets listed in Tables 1 and 2 can be regarded as normally distributed.

To ensure that the variance of the data from TRANSYT timing is equal to the variance of the data from PASSER timing, we collected the data along the same routes, during the same peak hours, and by the same driver and recorder during the TRANSYT and PASSER timings, respectively. Furthermore, the denominator of Equation 2 is from the weighted average of the sample variances of the TRANSYT and PASSER timings, respectively. It is the best estimate of the population variance, which also ensures the equal variance assumption (*9*, p. 293).

Table 3 summarizes the comparison. On Geary Boulevard, there was improvement in westbound (p.m. outbound). The PASSER timing reduced travel time and stops by 10 percent and 26 percent, respectively, along the arterial when compared to the TRANSYT timing. In eastbound, however, it increased travel time and stops by 3 percent and 25 percent, respectively. The changes in westbound were significant at a 95 percent level of confidence, but those in eastbound were not.

On 19th Avenue, there was improvement in northbound (a.m. inbound). The PASSER timing reduced travel time and stops by 9 percent and 43 percent, respectively. In southbound, however, it increased travel time and stops 20 percent

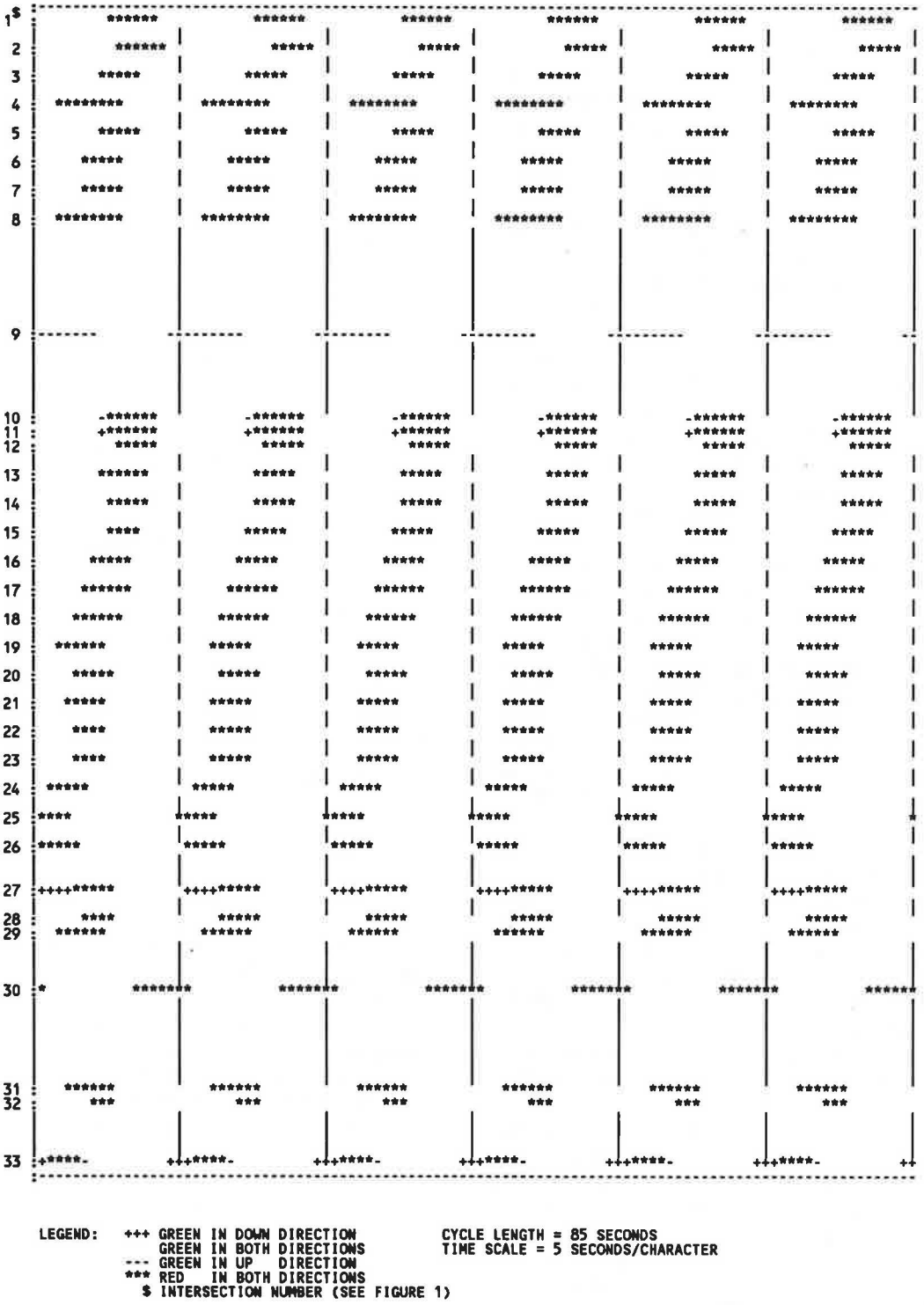


FIGURE 6 19th Avenue time-space diagram, from TRANSYT-7F link-to-link weight in both directions.

and 47 percent, respectively. All of these changes were significant at 95 percent level of confidence.

From the above results, it appeared that the PASSER timings improved one direction but worsened the other. To get an overall picture, we weighted the changes by traffic volumes, as shown in Table 4. The PASSER timings increased the weighted travel time by 0.7 percent and the weighted number of stops by 2.6 percent. The TRANSYT-7F timings appeared to be slightly better; however, if we exclude the data on eastbound Geary Boulevard (because both the travel time and stops were not statistically significant), the PASSER timings would increase the weighted travel time by 0.2 percent

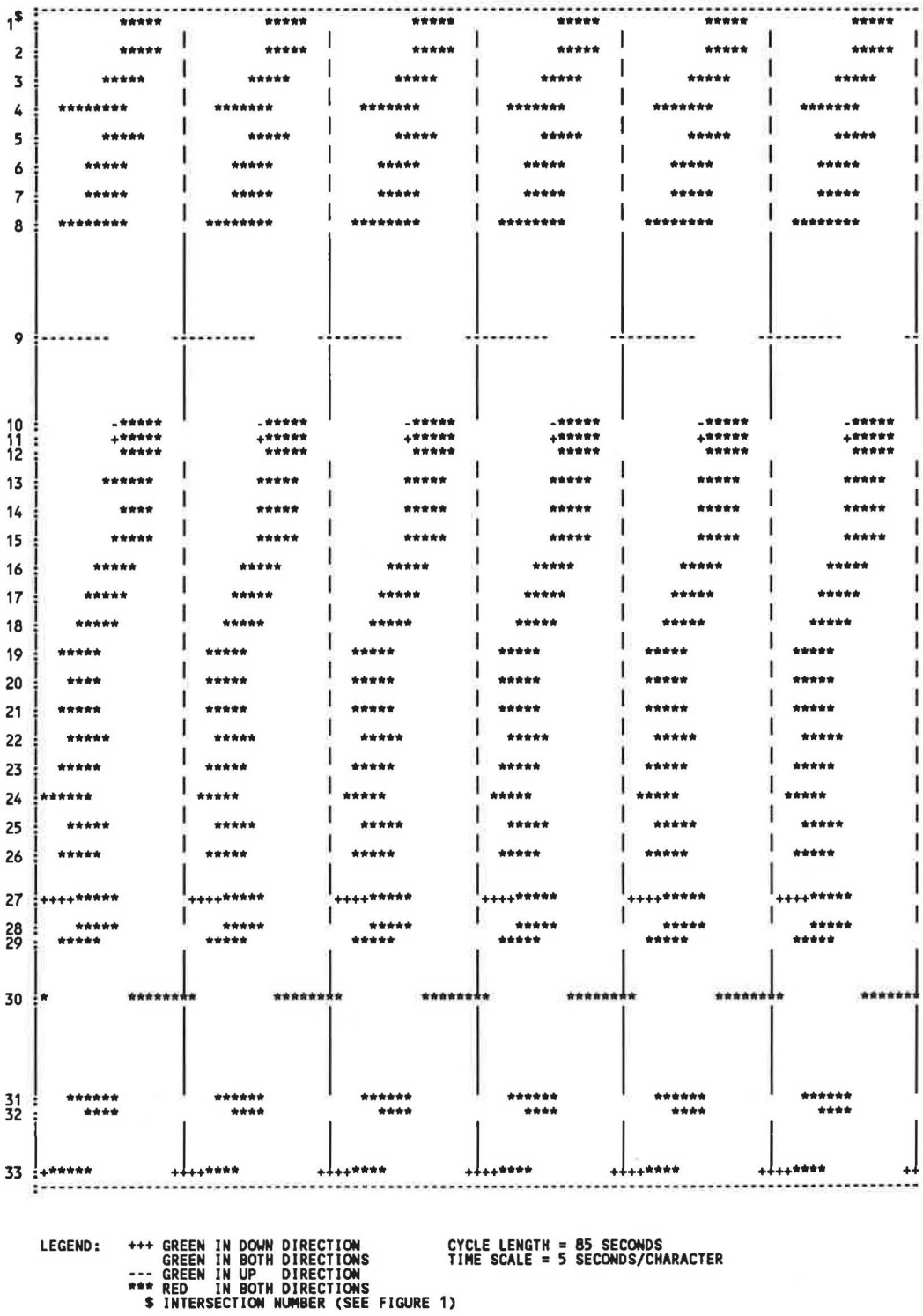


FIGURE 7 19th Avenue time-space diagram, from TRANSYT-7F link-to-link weight in northbound only.

and would reduce the number of stops by 9 percent. Therefore, we conclude that the effectiveness of the TRANSYT and PASSER timings was about the same.

In Table 3, the comparison of stops in eastbound Geary Boulevard was interesting. Although there was a 25 percent change in the number of stops, it was not significant at a 95 percent level of confidence (i.e., the result was due to chance). This happened even though we had defined stop precisely (see "Field Studies"), we had collected 9 to 15 samples (see Tables 1 and 2), and the magnitude of change was 25 percent. Also, as Tables 1 and 2 show, it may require 25 samples to attain a 95 percent level of confidence. Hence, collecting stops

TABLE 3 COMPARISON OF FIELD RESULTS

	Geary Boulevard		19th Avenue	
	Eastbound	Westbound	Northbound	Southbound
TRANSYT Timing				
Travel Time (Min/Mi)				
Mean, Xt	2.99	3.39	2.69	2.60
Standard Deviation, St	0.261	0.258	0.098	0.077
Number of Stops (Per Mi)				
Mean, Xt	1.55	2.18	1.59	1.49
Standard Deviation, St	0.743	0.315	0.146	0.134
PASSER Timing				
Travel Time (MIN/Mi)				
Mean, Xp	3.09	3.05	2.44	3.12
Standard Deviation, Sp	0.242	0.343	0.206	0.226
Number of Stops (Per Mi)				
Mean, Xp	1.94	1.61	0.90	2.19
Standard Deviation, Sp	0.392	0.582	0.285	0.288
Comparison of Travel Time				
Difference, Xt-Xp	-0.10	0.34	0.25	-0.52
% Change, 100(Xt-Xp)/Xt	-3%	10%	9%	-20%
t-Statistics[1]	-0.99	2.69	3.22	-6.14
Significant at 95% Confidence Level?	No	Yes	Yes	Yes
Comparison of Number of Stops				
Difference, Xt-Xp	-0.39	0.57	0.68	-0.70
% Change, 100(Xt-Xp)/Xt	-25%	26%	43%	-47%
t-Statistics[1]	-1.67	2.84	6.39	-6.15
Significant at 95% Confidence Level?	No	Yes	Yes	Yes

[1]Computed from equation 2.

TABLE 4 FIELD RESULTS WEIGHTED BY TRAFFIC VOLUME

	(1) Total Flow (Veh/Hr)	(2) Travel Time (Min/Mi)	(3) No. of Stops (No/Mi)	(1)x(2)	(1)x(3)
Under TRANSYT Timing					
Geary Blvd., E.B.	33255	2.99*	1.55*	99432	51545
Geary Blvd., W.B.	47755	3.39	2.19	161889	104583
19th Ave., N.B.	76806	2.69	1.59	206608	122122
19th Ave., S.B.	70606	2.60	1.49	183576	105203
Weighted Total, All Routes:				651505	383453
Weighted Total, Geary E.B. Excluded:				552073	331908
Under PASSER Timing					
Geary Blvd., E.B.	33255	3.09*	1.94*	102758	92645
Geary Blvd., W.B.	47755	3.05	1.61	145653	76886
19th Ave., N.B.	76806	2.44	0.90	187407	69125
19th Ave., S.B.	70606	3.12	2.19	220290	154627
Weighted Total, All Routes:				656108	393283
Weighted Total, Geary E.B. Excluded:				553350	300638

Change in Weighted Travel Time = (651505-656108)/651505 = -0.7%
Change in Weighted No. of Stops = (383453-393283)/383453 = -2.6%

Change in Weighted Travel Time, Geary E.B. Excluded = (552073-53350)/552073 = -0.2%
Change in Weighted No. of Stops, Geary E.B. Excluded = (331908-300638)/331908 = 9%

*Not significant at 95% level of confidence.

data may require great effort. On the other hand, travel time required five samples or less to attain a 95 percent level of confidence (see Tables 1, 2, and 3). Travel time data are also easy to obtain. Only a stopwatch is needed to record the starting and ending times. Because it is reliable and easy to obtain, one should use travel time data whenever possible.

COMPARISON OF MODEL-SIMULATED AND FIELD RESULTS

Table 5 shows the TRANSYT-7F simulated results under the TRANSYT and PASSER timings. The values were for links along the arterial (excluding cross streets) and were stratified to correspond to the test routes. The results show that the simulated travel time (total time), delays, stops, fuel consumption, and performance index under the TRANSYT timings were 4 to 52 percent better than those under the PASSER timings in all cases. The TRANSYT timings appeared to be better; however, the field results showed that the PASSER timings were better in some cases.

One reason for the disparity may be our choice of platoon dispersion factor (PDF) during simulation (we used PDF = 0.35 in Table 5). PDF affects the predicted flow rates from the upstream stop line to the downstream stop line and hence affects the simulated measures of effectiveness (MOEs). The TRANSYT-7F manual suggests 0.25, 0.35 and 0.5 for low-, moderate-, and heavy-friction roadway characteristics (*8*, p. 4-32). We applied all of these PDFs to simulate travel times. Table 6 (first three rows) shows the results. It shows that no matter which PDFs were used, the simulated travel times under the TRANSYT timings were 3 percent to 9 percent better than those under the PASSER timings. However, the travel times from field data (Table 6, fourth row) showed differently. It shows that the travel times on westbound Geary Boulevard and northbound 19th Avenue under the TRANSYT timings were 10 percent and 9 percent, respectively, worse than those under the PASSER timings. Hence the TRANSYT-7F simulation results were inclined in favor of the timing plans optimized by TRANSYT-7F.

Comparison of the magnitudes of the simulated and field travel times, in Table 6 (fifth row), shows that the differences ranged from 3 to 12 percent, which were reasonably close.

CONCLUSIONS

When timing arterial signals, TRANSYT-7F is expected to be better if the whole system, including cross streets, is considered; PASSER II is expected to be better if only the arterial street is considered. Field results, however, indicated that the effectiveness of TRANSYT-7F and PASSER II was about the same in terms of travel time and stops along the arterial (excluding cross streets). The offset patterns and operational characteristics of the TRANSYT timing might be different than those of the PASSER timing, as shown in 19th Avenue. On the other hand, they might be similar, as shown in Geary Boulevard. Travel time field data were reliable and easy to collect. Statistically, one to five samples were required to attain a 95 percent level of confidence for our example arterials, each with 30 or more signalized intersections. The TRANSYT-7F simulated travel times were reasonably close to the field travel times in terms of magnitude. On the other hand, the simulated travel times were inclined in favor of the timing plans optimized by TRANSYT-7F. Hence, when we make comparisons in relative terms, care must be exercised to avoid drawing the wrong conclusion. When making comparisons, travel time field data should be included whenever possible because they are reliable and easy to obtain. Though TRANSYT-7F may require more work to obtain an optimal timing plan, it is easier to use for fine-tuning or for later modification of any signals. Though PASSER II may be easier

TABLE 5 TRANSYT-7F SIMULATED RESULTS[1]

	Geary Boulevard						19th Avenue					
	Eastbound			Westbound			Northbound			Southbound		
	(1) TRANSYT Timing	(2) PASSER Timing	(1-2)/1 Percent Change	(3) TRANSYT Timing	(4) PASSER Timing	(3-4)/3 Percent Change	(5) TRANSYT Timing	(6) PASSER Timing	(5-6)/5 Percent Change	(7) TRANSYT Timing	(8) PASSER Timing	(7-8)/7 Percent Change
Total Time (Veh-Hr/Hr)	233	243	-4%	318	336	-6%	536	564	-5%	499	549	-10%
Total Delay (Veh-Hr/Hr)	75	85	-13%	98	116	-16%	178	206	-16%	167	218	-31%
Average Delay (Sec/Veh)	8.1	9.2	-14%	7.4	8.8	-19%	8.3	9.7	-17%	8.5	11.1	-31%
Uniform Stops (%)	36	48	-33%	31	42	-35%	31	47	-52%	31	46	-48%
Fuel Consumption (Gal/Hr)	307	332	-8%	403	446	-11%	797	914	-15%	734	854	-16%
Performance Index	117	143	-22%	140	184	-31%	315	433	-37%	290	410	-41%

[1]From TRANSYT-7F, Release 6's Route Summary Report, with arterial links (excluding cross street links) corresponding to the field study routes.

TABLE 6 TRAVEL TIMES FROM TRANSYT-7F SIMULATION AND FROM FIELD DATA

	Geary Boulevard						19th Avenue					
	Eastbound			Westbound			Northbound			Southbound		
	(1) TRANSYT Timing	(2) PASSER Timing	(1-2)/1 Percent Change	(3) TRANSYT Timing	(4) PASSER Timing	(3-4)/3 Percent Change	(5) TRANSYT Timing	(6) PASSER Timing	(5-6)/5 Percent Change	(7) TRANSYT Timing	(8) PASSER Timing	(7-8)/7 Percent Change
From Simulation (Min/Mi)[1]												
With PDF[2]= 25	3.10	3.22	-4%	3.05	3.23	-5%	2.50	2.61	-4%	2.53	2.76	-9%
With PDF = 35	3.12	3.23	-4%	3.09	3.24	-5%	2.53	2.63	-4%	2.56	2.79	-9%
With PDF = 50	3.15	3.23	-3%	3.14	3.26	-4%	2.57	2.67	-4%	2.60	2.84	-9%
From Field Data (Min/Mi)	2.99	3.09	-3%	3.39	3.05	10%	2.69	2.44	9%	2.60	3.12	-20%
Maximum Difference Between Simulation and Field Data[3]	-5%	-5%	NA	10%	-7%	NA	7%	-9%	NA	3%	12%	NA

[1]From TRANSYT-7F, Release 6's Route Summary Report, with Total Flow divided by Flow and the units converted to minutes per mile. The values were for through links on the arterial that correspond to the field study routes.

[2]PDF = Platoon dispersion factor.

[3]100[(Travel time from field data)-(highest or lowest travel time from simulation)]/(Travel time from field data)

NA = Does not apply.

to use for obtaining an optimal timing plan, it is difficult to use for fine-tuning or modifying selected intersections while trying to maintain an optimal setting with the rest of the system.

SUGGESTIONS FOR FURTHER RESEARCH AND DEVELOPMENT

1. These findings represent only two timing plans on two arterials. More studies are needed before we can generalize the results.
2. TRANSYT-7F assumes uniform arrival for each of the first approach entering a network. When timing an arterial, cross-street approaches are usually coded as the first approaches entering the network, hence the arrival patterns on cross streets are uniform. Although TRANSYT considers flows from cross streets, the flows may not be realistic. Future development should consider the ability of specifying arrival patterns.
3. When we try to get a different offset pattern from TRANSYT-7F optimization by applying weights to delay and stops, using link-to-link flow weighting feature, or inputting the PASSER II offsets, the result remains about the same. More research on the hill-climb process may produce a relationship to get different offset patterns.
4. In using PASSER II, one cannot freeze the offsets of certain intersections while optimizing the others. In practice, there is usually a need to change the offsets on certain intersections. When this happens, one would like to reoptimize only the affected intersections rather than the whole system. Future development should consider this possibility.
5. The green band from the PASSER II timing should enable a vehicle to travel without stopping throughout the system, if the vehicle is able to maintain a speed matching the design speed. However, this may not be possible because the vehicle may join the back of another moving platoon or queue. This moving platoon is from the previous cycle. Although PASSER II considers queue clearance time, it does not address the catching up of the moving platoon from the previous cycle. Research should be conducted to include this phenomenon into the optimization process.

ACKNOWLEDGMENTS

The author wishes to thank Gary Euler, Sheldon Strickland, Lyle Saxton, Antoinette Wilbur, and Beverly Russell of FHWA for their comments and Burton Stephens of FHWA for his support. The author thanks Norman Bray, Harvey Quan and Bond Yee of the City of San Francisco for providing the opportunity to work on projects leading to this paper.

REFERENCES

1. A. Skabardonis and A. D. May. Comparative Analysis of Computer Models for Arterial Signal Timing. In *Transportation Research Record 1021*, TRB, National Research Council, Washington, D.C., 1985, pp. 45–52.
2. S. L. Cohen. Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 81–84.
3. C. C. Liu. Bandwidth-Constrained Delay Optimization for Signal Systems. *ITE Journal*, Dec. 1988, pp. 21–26.

4. The UTM Surveys Traffic Engineer Software Users. *McTrans Newsletter*, University of Florida, Gainesville, June 1990.
5. S. Y. Wong. *City of San Francisco Fuel Efficient Traffic Signal Management Program—Final Report to California Department of Transportation*. Bureau of Traffic Engineering and Operations, City of San Francisco, Calif., Jan. 1987.
6. W. S. Homburger and J. H. Kell. *Fundamentals of Traffic Engineering*, 11th ed. Institute of Transportation Studies, University of California, Berkeley, 1984, pp. 7-2 and 7-3.
7. L. J. Pignataro. *Traffic Engineering Theory and Practice*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1973.
8. C. E. Wallace et al. *TRANSYT-7F User's Manual*. FHWA, U.S. Department of Transportation, Oct. 1988.
9. D. L. Harnett. *Introduction to Statistical Methods*. Addison-Wesley Publishing Co., Reading, Mass., 1975.
10. D. F. Votaw and H. S. Levinson. *Elementary Sampling for Traffic Engineers*. Eno Foundation, Westport, Conn., 1962.

The opinions expressed here are entirely those of the author.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Proposed Enhancements to MAXBAND 86 Program

NADEEM A. CHAUDHARY, ANULARK PINNOI, AND CARROLL J. MESSER

MAXBAND 86 is the only operational traffic signal program that allows progression bandwidth optimization in multiarterial, closed-loop traffic signal networks. The program formulates the problem as a mixed integer linear program and is capable of optimizing network-wide cycle length, signal offsets, and signal phasing sequences. However, hours of computer time may be required to optimize a medium-sized network problem, even on a mainframe computer. This computational inefficiency of MAXBAND 86 makes it impractical for use by the traffic engineering community. However, two heuristic methods efficiently optimize network signal timing problems modeled by MAXBAND 86. The experimental results demonstrate that these heuristic methods produce tremendous savings in the computer time required to solve optimization problems in traffic network signal timing. In addition, computational benefits are achieved by explicitly modeling one-way arterials in a network rather than as two-way arterials, as used in MAXBAND 86.

Traffic signal synchronization for maximum progression bandwidth is widely used because progression bands can be easily visualized and understood by traffic engineers as well as drivers. The capabilities of two existing computer programs, PASSER II (*1*) and MAXBAND 86 (*2,3*), allow traffic engineers to obtain progression bandwidth solutions to signal synchronization problems. The advantage that bandwidth optimization programs have over delay-based programs, such as TRANSYT-7F (*4*) and SIGOP (*5*), is their capability to select the best signal phasing sequences from the available set of possibilities.

PASSER II uses an efficient heuristic optimization technique based on the concept of minimizing interference to progression bands (*6*). However, the drawback of this technique is its inability to handle multiarterial networks with closed loops.

MAXBAND 86, on the other hand, is based on mathematical programming and therefore is capable of optimizing signal timing in networks having several arterials and closed loops. In spite of its mathematical elegance, MAXBAND 86 has two problems. First, its traffic flow model is extremely simplistic, and second, the program is computationally inefficient. For these reasons, MAXBAND 86 has not gained acceptance in the traffic engineering community. Research by Cohen (*7*) and Liu (*8*) has demonstrated that the concurrent use of MAXBAND and TRANSYT produces better signal timings than those produced by either program alone. Thus, an efficient MAXBAND program can provide practical technology for solving existing complex urban traffic congestion problems.

Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843-3135.

Almost all recent research related to the bandwidth maximizing approach has dealt with either some sort of enhancement to the arterial model used in MAXBAND 86 (*9,10*; see companion paper in this Record by Chaudhary et al.) or to the computational efficiency of the arterial problems (*11,12*). Little attention has been paid to the computational efficiency of MAXBAND 86 for optimizing multiarterial problems. We present two heuristic algorithms that efficiently solve signal synchronization problems in multiarterial networks. In addition, we demonstrate that explicitly formulating one-way arterials in a network problem produces better, quicker results than the method of treating one-way arterials as two-way arterials used in MAXBAND 86.

BACKGROUND

The mixed integer linear programming (MILP) formulation for maximizing the sum of progression bandwidths on a two-way arterial was originally formulated by Little (*13*). Little et al. (*14*) later expanded the arterial formulation for a triangular network with three arterials and a single closed loop, and developed MAXBAND, a computer program based on this formulation. MAXBAND had the capability to automatically set up and solve an MILP for a given set of traffic data. MPCODE, the optimization package used in MAXBAND, is a set of computer routines developed by Land and Powell (*15*). Messer et al. enhanced MAXBAND for multiarterial, multiple closed-loop signal network problems (*2*). The enhanced program was named MAXBAND 86. Recent experience with the application of MAXBAND 86 to networks with multiple closed loops indicates that the central processing time (CPU) on a mainframe computer may be in hours, even for medium-sized network problems (*16,17*). These researchers thought that the MPCODE optimization package was inefficient and that replacing it with a more efficient package would solve the MAXBAND inefficiency problem (*2*).

Most research dealing with the computational efficiency of the progression bandwidth maximizing approach has dealt with single arterial problems (*6,11,12*). Mireault successfully applied Benders's decomposition technique to arterial and network problems (*11,18*). This research indicates that Benders's decomposition approach for optimizing network problems with fixed cycle length and fixed travel speeds is up to 10 times faster than the branch and bound method (*18*, pp. 161, 274). However, problems become difficult to solve in a practical amount of time (3 hr) when cycle length and travel speeds are made variable (*17*, p. 275). Recently, Chaudhary et al. (*12*) developed a two-step heuristic method for solving

arterial problems efficiently. Chaudhary et al. also compared the performance of MPCODE with that of LINDO (*19*), an efficient commercial optimization package. The results indicate that, at least for bandwidth optimization problems, MPCODE is as efficient as LINDO, and that it is the nature of the problem, rather than any weakness of the MPCODE optimization package, that makes it difficult to solve. Thus, the need remains to develop heuristic methods to enhance the efficiency of MAXBAND 86 for optimizing network signal synchronization problems using MPCODE.

MATHEMATICAL FORMULATION

The mathematical formulation used in MAXBAND 86 for optimizing progression bandwidth in arterial and network problems is given in several publications (*2,3,14,17*). This formulation determines cycle length, offsets, and signal phasing sequences that maximize the sum of progression bandwidths for all arterials. The green splits remain constant during the optimization process. The formulation for a multiarterial, closed-loop network problem consists of three types of integer variables, shown in Figure 1. The following statements describe these integer variables:

1. General integer variable m_i ensures that the sum of offsets around a loop formed by two one-way links connecting a pair of adjacent traffic signals is an integer multiple of the signal cycle length.

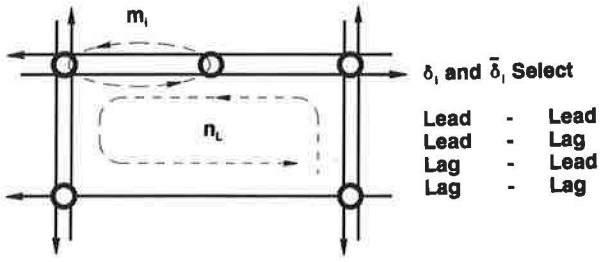


FIGURE 1 Description of integer variables.

2. General integer variable n_L ensures that the sum of offsets around a closed loop formed by three or more arterials is an integer multiple of the cycle length.

3. Binary variables δ_i and $\bar{\delta}_i$ select signal phasing sequences that produce maximum progression bandwidth.

TEST PROBLEMS

Thirteen real-world network problems were used to compare the efficiency of the basic simultaneous optimization procedure and the heuristic optimization methods. Detailed information about most of these networks is given by Cohen (*16*). Table 1 describes the test problems. A constant cycle length was used for each test problem.

The optimization runs were performed on a DecStation 3100 computer. This computer is considered about twice as fast (15,284 Khornerstones/sec) as a Compaq 386/25 80386 computer (7,417 Khornerstones/sec) (*20*). For practical reasons, an upper bound of 2 million branch and bound (BB) iterations was placed on all optimization runs.

OPTIMIZATION USING EXISTING MAXBAND 86 PROGRAM

The simultaneous optimization of test problems was run using MAXBAND 86. The purpose of this set of optimization runs was twofold: to gain insight into the nature of network optimization problems and to develop a basis for determining the effectiveness of the heuristic methods developed later.

Table 2 summarizes the results of simultaneous optimization using the existing MAXBAND 86 program. The following statements describe some of the results:

1. For five problems, the upper limit of 2 million BB iterations was not enough to complete the search. This means that the solutions obtained for these problems are not guaranteed to be the absolute best.

2. These results support previous research findings and show that large amounts of CPU time may be required to optimize

TABLE 1 DESCRIPTION OF NETWORK PROBLEMS

NO.	DATA NAME	NETWORK NAME	NETWORK GEOMETRY				MILP SIZE					
			ARTERIALS	SIGNALS	LINKS	LOOPS	TOTAL CONSTRAINTS	VARIABLES CONTINUOUS	INTEGER	BINARY	TOTAL	NON ZERO ELEMENTS
1.	W311	University/Canyon/12th Street	3	11	11	1	121	57	12	21	90	372
2.	W315	Wisconsin/Massachusetts/Garfield	3	15	15	1	168	73	16	3	92	472
3.	W317	Pennsylvania/Connecticut/K Street	3	17	17	1	190	81	18	7	106	494
4.	W509	Hawthorne Blvd. mini network, California	5	9	10	2	109	61	12	16	89	346
5.	W613	Walnut Creek Network, California	6	13	15	3	164	85	18	33	136	544
6.	W712	Daytona Beach Network, Florida	7	12	17	6	188	97	23	36	156	658
7.	W813B	Post Oak Network, Houston, Texas	8	13	18	6	198	105	24	29	158	654
8.	W813C	Ogden Network, Utah	8	13	18	6	198	105	24	18	147	632
9.	W814	Ann Arbor Michigan	8	14	20	7	221	113	27	31	171	751
10.	W815	Los Angeles, California	8	15	21	7	232	117	28	33	178	777
11.	W816A	Owosso, Michigan	8	16	18	3	195	105	21	27	153	602
12.	W816B	Bay City, Michigan	8	16	20	5	219	113	25	30	168	719
13.	W817	Downtown Memphis Network, Tennessee	8	17	22	6	242	121	28	15	164	771

TABLE 2 RESULTS FROM MAXBAND 86 SIMULTANEOUS OPTIMIZATION

		BEST INTEGER SOLUTIONS					SOLUTIONS WITHIN 95% OF BEST		
NO.	DATA	SOLNS. FOUND	TOTAL BAND(SEC)	FOUND AT ITERATION	BB SEARCH STOPPED AT	CPU TIME HRS:MIN:SEC	SOLUTIONS FOUND	FIRST SOLUTION	FOUND AT ITERATION
1.	W311	17	104.32	6129	12548	1:38	7	100.96	3421
2.	W315	41	95.97	23741	28306	5:18	4	95.97	21246
3.	W317	24	84.08	42809	46212	9:45	4	84.08	42112
4.	W509	16	359.64	4365	4571	0:36	14	342.54	550
5.	W613	72	277.00	857591	955959	3:34:23	10	265.20	458325
6.	W712	47	298.32[a]	1306120	1999818	9:51:44	16	285.48	217712
7.	W813B	42	281.60	1254322	1359574	7:10:07	8	273.60	802663
8.	W813C	29	246.88	1579052	1892872	9:21:51	4	242.48	369606
9.	W814	36	337.20[a]	1287992	1999828	11:59:56	4	328.10	1248071
10.	W815	41	288.50[a]	1767302	1999803	13:49:08	17	275.80	243536
11.	W816A	96	335.60	1159536	1257239	5:47:47	13	327.60	1070847
12.	W816B	56	357.60[a]	1798311	1999809	11:17:23	11	342.60	1473867
13.	W817	62	335.70[a]	1982661	1999759	12:31:45	4	323.50	1888652

[a] Suboptimal solution due to imposed upper limit of 2,000,000 on BB iterations.

a medium-sized problem. For our test problems, CPU time varied from 36 sec to 13 hr and 50 min.

3. All test problems have multiple-integer solutions. Multiple solutions within 95 percent of the best solution were also found; some of these were close to or the same as the best solution.

4. For this set of test problems, an upper limit of 1,888,652 BB iterations would have been required to guarantee that the solutions obtained were within 95 percent of the best solution.

5. Optimal values of the loop variables (m_i) were always in the interval [0,2].

A close examination of the intermediate results from the MPCODE optimization package revealed that several sets of values for loop variables (m_i) may satisfy the loop-closure constraints. Also, for a given set of values for the loop variables (m_i), several different phasing sequence combinations produced the same value of progression bands. These properties of the network problems are proposed as the basic cause of MAXBAND 86's computational inefficiency. Two heuristic algorithms were developed to exploit the properties described. The following sections describe the heuristic algorithms and present the results of optimization using these algorithms. These experiments used lower and upper bounds of 0 and 2 on the arterial loop variables.

TWO-STEP HEURISTIC OPTIMIZATION METHOD

Let 2SMP1 be a problem obtained by relaxing the integrality constraints on the phasing sequence selection variables (δ_i and $\bar{\delta}_i$) in the network optimization problem, and let 2SMP2 be another problem obtained by setting the arterial loop variables (m_i) and network loop variables (n_L) in the network problem to a specific set of values. Then, the two-step heuristic method is as follows:

- Step 1. Optimize Problem 2SMP1 and save the set of values of variables m_i and n_L corresponding to the last six best solutions.
- Step 2. For each set of values of loop variables saved in Step 1, solve Problem 2SMP2. Select the best of these as the optimal solution.

The two-step method was used to optimize the 13 test problems described previously. MPCODE was used to optimize all subproblems. Table 3 summarizes the results of the two-step method on the test problems and compares the results to those obtained from the prior simultaneous optimization (MAXBAND 86). The following statements describe the results:

1. Progression bands obtained are almost the same (within 99.5 percent) or better than those obtained by MAXBAND 86. On the average, the bands are 4.4 percent better than those obtained by MAXBAND 86. This result occurred because MAXBAND 86 did not finish several problems because of the upper limit of 2 million BB iterations.

2. The number of BB iterations required varied from 4.11 to 42.86 percent of that required by MAXBAND 86. In other words, a savings of up to 95.89 percent in BB iterations was achieved. The savings in BB iterations for the nine largest problems was at least 80.12 percent. The average savings in BB iterations was 81.15 percent.

3. The maximum CPU time was about 2.5 hr, compared to the maximum CPU time of 13.82 hr for MAXBAND 86 optimization.

4. Step 2 (phasing sequence selection process) of the heuristic algorithm required the least amount of computational effort.

In summary, the two-step heuristic algorithm is far superior computationally to the simultaneous optimization method. The results are better than those obtained by the simultaneous

TABLE 3 RESULTS FROM TWO-STEP HEURISTIC OPTIMIZATION

NO.	DATA NAME	BEST SOLUTION BAND(SEC)	% OF MAX 86	BRANCH AND BOUND ITERATIONS (STEP 1 + STEP 2 = TOTAL)			% OF MAX 86	CPU TIME HRS:MIN:SEC
1.	W311	104.32	100.00	4301	1077	5378	42.86	0:00:42
2.	W315	95.97	100.00	10984	749	11733	41.45	0:02:07
3.	W317	84.08	100.00	15753	780	16533	35.78	0:03:11
4.	W509	359.64	100.00	1009	342	1351	29.56	0:00:10
5.	W613	275.30	99.39	44331	3722	48053	5.03	0:11:00
6.	W712	349.20	117.01	209043	1200	210243	10.51	1:02:26
7.	W813B	281.60	100.00	124366	5238	129604	9.53	0:39:41
8.	W813C	246.88	100.00	374836	1464	376300	19.88	1:51:50
9.	W814	370.60	109.91	141119	2112	143231	7.16	0:49:55
10.	W815	360.90	125.10	382868	2508	385376	19.27	2:29:09
11.	W816A	335.60	100.00	50552	1064	51616	4.11	0:14:24
12.	W816B	373.20	104.36	153819	2766	156585	7.83	0:53:15
13.	W817	340.80	101.52	239965	1612	241577	12.08	1:31:32
MEAN			104.41				18.85	

optimization method completed in a feasible amount of time. However, the CPU time required may still be more than that desired for practical applications. The following section presents a three-step heuristic method designed to be more efficient than the two-step algorithm.

THREE-STEP HEURISTIC OPTIMIZATION METHOD

Let 3SMP1 be the original network problem obtained by relaxing the integrality constraints on the network loop variables (n_L) and phasing sequence selection variables (δ_i and $\bar{\delta}_i$). Let 3SMP2 be another network problem obtained by fixing arterial loop variables (m_i) to a specific set of values and relaxing the integrality constraints on the phasing sequence selection variables. Finally, let 3SMP3 be the original network problem with loop variables (m_i, and n_L) fixed at specific values. Then, the three-step heuristic method is as follows:

- Step 1. Optimize Problem 3SMP1 and save the set of values of the variables m_i corresponding to the last six best solutions.
- Step 2. For each set of values of arterial loop variables saved in Step 1, solve Problem 2SMP2. Save the values of network loop variables corresponding to the six best solutions.
- Step 3. For each of the six sets of values of loop variables (arterial and network) saved at the second step, optimize 3SMP3. Select the best of these as the optimal solution.

MPCODE was used to test the effectiveness of the three-step method on the same set of 13 problems. Table 4 summarizes the results of these optimization runs and compares the results with those from the simultaneous optimization. The following statements describe the results:

1. The sums of progression bands obtained were, on the average, as good as those obtained by MAXBAND 86. Except for one problem, the solutions were within 90 percent of those obtained by MAXBAND 86. For one problem, the total band was 25 percent more; for this problem, MAXBAND 86 did not finish because of the upper limit on BB iterations.
2. The average number of BB iterations required to optimize the test problems was less than 9 percent of that required by the simultaneous optimization. The savings in BB iterations was from 65.6 percent to 99.41 percent. For the nine largest problems, the savings in BB iterations was at least 98.13 percent.
3. The CPU time required to solve these problems varied from 10.8 sec to 13.1 min.
4. The third step (phasing sequence selection process) of the heuristic method required the least amount of computational effort.
5. For seven problems, total bands obtained were the same as those produced by the two-step method. For five problems, the bands were within 90.8 percent of those from the two-step method.
6. Except for one small problem (118 more BB iterations), the three-step method optimized the problems much more efficiently than the two-step method. Compared to the two-step method, the reduction in BB iterations ranged from 37 to 91 percent. Reduction in BB iterations for the nine largest problems was at least 72 percent.

In summary, the three-step method is much more efficient than the two-step heuristic method. The three-step method solved in less than 13.1 min problems for which the two-step method required up to 2.5 hr of CPU time. The total progression bandwidths produced by the three-step method were generally less than those produced by the two-step method. The reduction in bandwidth, as compared to the two-step

TABLE 4 RESULTS FROM THREE-STEP HEURISTIC OPTIMIZATION

NO.	DATA NAME	BEST SOLUTION BAND(SEC)	% OF MAX 86	BRANCH AND BOUND ITERATIONS (STEP 1 + STEP 2 + STEP 3 = TOTAL)				% OF MAX 86	CPU TIME MIN:SEC
1.	W311	104.32	100.00	1454	808	1127	3389	27.01	0:25
2.	W315	95.97	100.00	4155	1024	751	5930	20.95	0:56
3.	W317	84.08	100.00	6856	832	574	8262	17.88	1:34
4.	W509	359.64	100.00	337	575	557	1469	32.14	0:11
5.	W613	272.90	98.52	5987	2798	4498	13283	1.39	2:36
6.	W712	259.20	86.89	21629	11194	1548	34371	1.72	8:37
7.	W813B	274.00	97.30	6940	13662	3808	24410	1.80	6:24
8.	W813C	229.76	93.07	24084	8142	1279	33505	1.77	8:22
9.	W814	336.60	99.82	2357	26932	1725	31014	1.55	9:34
10.	W815	360.90	125.10	13847	21142	2375	37364	1.87	13:06
11.	W816A	335.60	100.00	3816	2357	1241	7414	0.59	1:39
12.	W816B	357.60	100.00	11217	7926	1700	20843	1.04	5:44
13.	W817	340.80	101.52	17141	13010	1492	31643	1.58	9:34
MEAN			100.17					8.56	

method, may not be significant in view of the computational savings achieved by the three-step method. The total bands produced by the three-step method were, on the average, close to those obtained from MAXBAND 86.

EXPLICIT MODELING OF ONE-WAY ARTERIALS IN NETWORK PROBLEMS

In multiarterial network problems, MAXBAND 86 deals with one-way arterials as two-way arterials. This flaw results in the addition of unnecessary variables and constraints and affects CPU time as well as the quality of the solution. This flaw can be removed by using the linear programming formulation of a one-way arterial instead of the MILP formulation used in MAXBAND 86. For an arterial with n intersections and k left-turn movements, this formulation has $(3n + k - 1)$ to $(2n + 1)$ fewer variables, and a $(6n - 6)$ to $(5n - 4)$ reduction in the number of constraints as compared to the original formulation. This reduction, especially the elimination of $(n - 1)$ general integer variables, is quite significant.

This section shows the results of experimentation using the correct network formulation. The effectiveness of this formulation was tested for Problems W813C and W814. Problem W813C has two one-way arterials, and Problem W814 has one. Table 5 compares the original and corrected problem formulations in terms of the number of variables and constraints that each problem contains. We optimized the corrected formulations for these two problems using simultaneous and heuristic optimization methods. Table 6 compares these results with those given previously. The following statements summarize the results:

1. The size of the new problem formulation is reduced considerably.
2. Simultaneous optimization of the new formulation for Problem W813C resulted in a reduction of almost 72 percent in BB iterations. The quality of the best solution also improved slightly. Simultaneous optimization of Problem W814 produced a better solution; however, like the original problem, the BB search failed to terminate within the specified limit of 2 million BB iterations.

TABLE 5 COMPARISON OF ORIGINAL AND CORRECTED FORMULATIONS

DATA NAME AND DESCRIPTION	VARIABLES CONTINUOUS	INTEGER	BINARY	TOTAL	TOTAL CONSTRAINTS	NON ZERO ELEMENTS
W813C (MAXBAND 86)	105	24	18	147	198	632
W813C (Corrected)	91	19	18	128	168	543
Reduction	14	5	0	19	30	89
W814 (MAXBAND 86)	113	27	31	171	221	751
W814 (Corrected)	103	23	31	157	197	678
Reduction	10	4	0	14	24	73

TABLE 6 COMPARISON OF OPTIMIZATION RESULTS

DATA NAME AND DESCRIPTION	MAX 86 OPTIMIZATION		TWO STEP HEURISTIC		THREE STEP HEURISTIC	
	BAND(SEC)	ITERATIONS	BAND(SEC)	ITERATIONS	BAND(SEC)	ITERATIONS
W813C (MAXBAND 86)	246.88	1,892,872	246.88	376,300	229.76	33,505
W813C (Corrected)	253.28	408,408	253.28	114,610	242.24	15,945
Difference	6.40	-1,484,464	6.40	-261,690	13.48	-17,560
	2.6%	-78.4%	2.6%	-69.5%	5.9%	-52.4%
W814 (MAXBAND 86)	337.20	1,999,828	370.60	143,231	336.60	31,014
W814 (Corrected)	357.20	1,999,859	392.30	71,726	382.90	11,475
Difference	20.00*	*	21.70	-71,505	46.30	-19,539
	5.9%*	*	5.9%	-49.9%	13.8%	-63.0%

* Upper limit of 2 million BB iterations reached. These are sub-optimal solutions.

3. Compared to the two-step optimization of the original formulations for the two problems, the new formulation required 69.54 and 49.92 percent less effort, respectively. In addition, the quality of solutions (total bandwidth) was better.
4. The three-step method also produced quicker results on the new problems. The savings in BB iterations for the test problems were 52.41 and 63 percent, respectively, as compared to optimization of the original formulation.

In summary, the correct formulation is easier to solve using all three optimization methods. In addition, a larger total progression bandwidth is obtained, compared to the original formulation used in MAXBAND 86. This means that the new formulation can optimize signal timings in larger networks than those that could be optimized with the existing MAXBAND 86, especially downtown grid networks in which most of the arterials are one-way. Further, arterial networks can now include one-way frontage roads. This may make it easy to combine freeway ramp metering optimization with optimization of signal timing on the surface street system.

SUMMARY AND RECOMMENDATIONS

We present two ways to increase the computational efficiency of the MAXBAND 86 program. First, we demonstrate that traffic signal network problems can be efficiently optimized using the proposed heuristic methods, without sacrificing the quality of solutions. Second, we demonstrate that the explicit formulation of one-way arterials reduces the network problem size, and produces better progression bands as compared to the network formulation used in MAXBAND 86.

We recommend that the heuristic optimization methods be incorporated in the MAXBAND 86 program for three reasons. First, implementation of these methods is straightforward; second, these methods are robust, that is, any enhancement or modification in the problem formulation will not affect these procedures; and third, they provide more increase in the problem-solving computational efficiency than any other available method. It is recommended that the two-step and three-step optimization methods be added to the simultaneous optimization capability of MAXBAND 86. The choice of the optimization method to be used can then be based on the problem size. It is also recommended that the MAXBAND 86 formulation for one-way arterials be corrected. This will allow signal timing optimization in larger networks than is currently achievable. In addition, wider progression bands can be obtained.

ACKNOWLEDGMENTS

This material is based in part on work supported by the Governor's Energy Management Center, State of Texas Energy Research in Applications Program. The network data used were supplied by Stephen L. Cohen of FHWA. The authors would like to thank him for assisting this work. The authors would also like to thank the anonymous referees for providing useful comments that helped to improve this paper.

REFERENCES

1. E. C. Chang, B. G. Marsden, and B. R. Derr. The PASSER II-84 Microcomputer Environment System—A Practical Signal Timing Tool. *Journal of Transportation Engineering*, Vol. 113, Nov. 1987, pp. 625–641.
2. C. J. Messer, G. L. Hogg, N. A. Chaudhary, and E. C. P. Chang. *Optimization of Left Turn Phase Sequence in Signalized Networks Using MAXBAND 86, Volume 1: Summary Report.* Report FHWA/RD-84/082. FHWA, U.S. Department of Transportation, Jan. 1986.
3. E. C. Chang, S. L. Cohen, C. Liu, C. J. Messer, and N. A. Chaudhary. MAXBAND-86: Program for Optimizing Left-Turn Phase Sequence in Multiarterial Closed Networks. In *Transportation Research Record 1181*, TRB, National Research Council, Washington, D.C., 1989, pp. 61–67.
4. C. E. Wallace, K. G. Courage, D. P. Reaves, G. W. Schoene, G. W. Euler, and A. Wilbur. *TRANSYT-7F User's Manual.* FHWA, U.S. Department of Transportation, 1988.
5. E. B. Lieberman and J. L. Woo. SIGOP II: A New Computer Program for Calculating Optimal Signal Patterns. In *Transportation Research Record 596*, TRB, National Research Council, Washington, D.C., 1976, pp. 16–21.

6. C. J. Messer, R. H. Whitson, C. L. Dudek, and E. J. Romano. A Variable-Sequence Multiphase Progression Optimization Program. In *Highway Research Record 445*, HRB, National Research Council, Washington, D.C., 1973, pp. 24–33.
7. S. L. Cohen. Concurrent Use of MAXBAND and TRANSYT Signal Timing Programs for Arterial Signal Optimization. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 81–84.
8. C. C. Liu. Bandwidth-Constrained Delay Optimization for Signal Systems. *ITE Journal*, Vol. 58, No. 12, Dec. 1988, pp. 21–26.
9. N. H. Gartner, S. F. Assmann, F. Lasaga, and D. L. Hou. MULTIBAND—A Variable-Bandwidth Arterial Progression Scheme. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990.
10. H. S. Tsay and L. J. Lin. A New Algorithm for Solving the Maximum Progression Bandwidth. Presented at the 67th Annual Meeting of the Transportation Research Board, Washington D.C., Jan. 1988.
11. P. Mireault. Solving the Single Artery Traffic Signal Synchronization with Benders Decomposition. Presented at CORS/ORSA/TIMS Joint National Meeting, Vancouver, Canada, May 8–10, 1990.
12. N. A. Chaudhary, C. J. Messer, and A. Pinnoi. Efficiency of Mixed Integer Linear Programs for Traffic Signal Synchronization Problems. *Proc., 25th Annual SE TIMS Meeting*, Oct. 1989, pp. 155–157.
13. J. D. C. Little. The Synchronization of Traffic Signals by Mixed-Integer Linear Programming. *Operations Research*, Vol. 14, 1966, pp. 568–594.
14. J. D. C. Little, M. D. Kelson, and N. H. Gartner. MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks. In *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981, pp. 40–46.
15. A. Land and S. Powell. *Fortran Codes for Mathematical Programming*. John Wiley & Sons, Ltd., London, England, 1973.
16. S. L. Cohen. *Optimization of Left Turn Phase Sequence in Signalized Closed Networks*. Final Report. FHWA, U.S. Department of Transportation, 1988.
17. N. A. Chaudhary. *A Mixed Integer Linear Programming Approach For Obtaining an Optimal Signal Timing Plan in General Traffic Networks*. Ph.D. dissertation. Texas A&M University, College Station, Aug. 1987.
18. P. Mireault. *An Integer Programming Approach to The Traffic Signal Synchronization Problem*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Feb. 1988.
19. L. Schrage. *User's Manual for LINDO*, 3rd ed. The Scientific Press, Redwood City, Calif., 1987.
20. P. Magney. DECstation 3100: A Leader. *Computer Reseller News*, Sept. 4, 1989, pp. 57–58.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Evaluation of Optimized Policies for Adaptive Control Strategy

NATHAN H. GARTNER, PHILIP J. TARNOFF, AND CHRISTINA M. ANDREWS

Optimization Policies for Adaptive Control (OPAC) is an on-line control algorithm designed to optimize the performance of individual traffic signals. It is a building block for demand-responsive control of a distributed signal system. OPAC-RT is a traffic signal control system that implements the OPAC strategy in real time. The system uses traffic data collected from detectors located well upstream (400 to 600 ft) of the stop bar on all approaches to an intersection. Signal timings are dynamically optimized in a demand-responsive manner using a rolling horizon scheme. Results of the first implementation and field testing of the on-line OPAC strategy at individual intersections indicate that OPAC performs better than well-timed actuated signals, particularly at greater demand levels. The payback period of the incremental costs of installing the current version of OPAC is estimated to be less than 1 year. Further enhancements of the OPAC system operation are likely to significantly increase its effectiveness.

Many different traffic signal control strategies are available to the traffic engineer. These strategies can be grouped into the following basic categories: isolated intersection control, arterial control, or network control (*1*). These categories are further distinguished between off-line strategies, which process manually collected data using batch computer programs to produce signal timing plans, and on-line strategies, which use detector inputs to calculate signal timings for immediate implementation.

Ever since the inception of modern traffic signal controls, traffic engineers and signal system designers have attempted to make them as responsive as possible to the prevailing traffic conditions on the premise that increased responsiveness would lead to improved traffic performance. This premise was broadly applied to individual intersection signals as well as to area-wide systems. However, the extent to which traffic responsiveness can be achieved depends on a variety of factors, including the type of control hardware, software capabilities, surveillance and communication equipment, and operator qualifications.

The Optimized Policies for Adaptive Control (OPAC) strategy is an on-line traffic signal timing optimization algorithm that represents the most recent development in traffic control research. The development of this strategy was based on the following principles (*2–4*):

1. The strategy must provide better performance than off-line methods. Although it may seem self-evident, this principle was not always explicitly incorporated in the development of previous responsive strategies. Other less relevant criteria, such as "main-street platoon progression" or "variable cycle in time and in space," were often used.
2. The strategy must be truly demand-responsive, that is, it must adapt to actual traffic conditions and not be responsive to historical or predicted values, which are unreliable and may be far from actuality.
3. The strategy must not be restricted to arbitrary control periods but should be capable of providing continuously optimized controls. Effective responsiveness cannot be achieved by implementing off-line methods at shorter and shorter intervals.
4. Development of new control concepts that are better suited to the variability in traffic flows is needed, not merely the extrapolation of existing concepts. Thus, the conventional notions of cycle time, splits, and offsets, which are inherent in all existing signal optimization methods, are unsuited for demand-responsive control. On the other hand, direct minimization of the performance measures provides much improved performance.
5. The strategy should not be encumbered by a rigid network structure; rather, it should be based on decentralized decision making.

As a result, OPAC was developed as a distributed strategy featuring a dynamic optimization algorithm for traffic signal control without requiring a fixed cycle time. Signal timings are calculated to directly minimize performance measures, such as vehicle delays and stops, and are only constrained by minimum and maximum phase lengths. The strategy was originally developed at the University of Lowell under sponsorship of the U.S. Department of Transportation and is well documented in the literature (*5–8*). The following section briefly describes the OPAC methodology.

Following comprehensive simulation studies (*9*), a field implementation and evaluation of the method was sponsored by FHWA (*10*). Because OPAC is a distributed strategy, it was decided to initially test the operation of the single-intersection module. The same module serves as a building block of a multi-intersection system. The following specific objectives of the study were formulated:

- To develop a system that would enable the real-time OPAC program to interface with a full-actuated, modern, solid-state controller, so that the signal phase times would be determined by the OPAC optimization algorithm;
- To determine, in terms of traffic performance, how effectively the OPAC algorithm controls traffic at isolated intersections as compared with well-timed full-actuated control;
- To make recommendations, based on the observed performance of the existing OPAC control algorithm, for

N. H. Gartner, Department of Civil Engineering, University of Lowell, Lowell, Mass. 01854. P. J. Tarnoff and C. M. Andrews, Farradyne Systems, Inc., 3206 Tower Oaks Blvd., Rockville, Md. 20852.

modifications and enhancements that would improve its effectiveness;

- To estimate the cost-effectiveness of using OPAC instead of full-actuated control at isolated intersections;
- To produce a strategy that can be offered to the signal industry for commercial implementation; and
- To develop features that minimize the amount of fine-tuning required by local traffic engineering personnel.

The results of this study are described in a report prepared by the contracting agency (*10*). This paper discusses primarily the technical issues involved in the first field implementation of the OPAC strategy and the analysis of the traffic performance results.

DEVELOPMENT OF OPAC STRATEGY

Dynamic Programming: OPAC-1

The OPAC strategy evaluated in this study is the culmination of a research effort that included the development of three optimization algorithms (*5*). The first, designated OPAC-1, was designed as a basis for future OPAC strategy development. OPAC-1 uses dynamic programming techniques for the solution of the traffic control problem. Dynamic programming is a global optimization strategy for multistage decision processes (*11*). As such, it provides an absolute standard against which all other strategies can be compared.

When applying dynamic programming to the signal control problem, each interval of time is designated as a stage (with a typical length of 5 sec). For each stage, the initial state is defined by the initial queues on each approach and the signal status. The initial signal status for each approach is either green (0) or red (1). The input decision variable for each approach is either 0 (no signal change) or 1 (change). The output of the algorithm at each stage consists of the new queue values and signal indications that will result on each approach from the implementation of the specific decision. The performance measure (delays and/or stops) is calculated to be the sum of the minimized performance associated with the corresponding intersection state at the succeeding interval (which has already been calculated because the procedure moves backwards), the initial queues, and the arrivals minus the departures during the stage.

The dynamic programming procedure as applied to the traffic signal control problem can be summarized as follows:

1. Select an intersection state at Interval i; that is, select a specific queue length and signal status within the valid ranges for each approach.
2. Calculate the total performance (e.g., delay) for intervals i to n (the last interval in the stage) for each input decision (i.e., calculate the delay assuming the signal changes in Interval i and recalculate the delay assuming there is no change).
3. Choose the policy for Interval i to be "change" or "no change" based on which produced the best total performance.
4. Repeat Steps 2 and 3 for all valid input states at Interval i.

The procedure is terminated when Interval 1 has been reached and yields values for the optimum policy and minimized total performance for each initial intersection state.

Although this procedure ensures globally optimal solutions, it requires complete knowledge of arrivals over the entire control period. It cannot be used for real-time implementation because of the amount of processing involved. Much of the output from the program is never implemented because optimized policies are generated for all possible combinations of initial conditions at each stage of the control period. If an "entire optimum policy" is defined as the complete sequence of optimized policies throughout the control period that corresponds to a particular initial state at Interval 1, then the algorithm could produce hundreds of such policies. However, in practice, only one initial state at Interval 1 exists; hence, only one "entire optimum policy" would be implemented. By being able to produce the theoretically optimal control strategy for each input state, OPAC-1 is a standard for the evaluation of the relative effectiveness of other, more practical strategies.

Sequential Optimization: OPAC-2

The second optimization algorithm, OPAC-2, is a simplification of the OPAC-1 algorithm. It was designed as a building block in the development of a distributed on-line strategy. OPAC-2 has the following features:

- The control period is divided into stages T sec long. In this case, T is approximately equal to a typical cycle length, although it could be longer. (It should be remembered that there is no fixed cycle length in OPAC.)
- Each stage is divided into an integral number of intervals s sec long. For development of the algorithm, $s = 5$ sec was chosen.
- Each stage must include at least one signal phase change and may include as many as three. The phase change (switching) times are measured from the start of the stage in time units of s.
- For any given switching sequence at stage n, the performance function for each approach is defined to be the sum over all intervals in the stage of the initial queue length plus the arrivals minus the departures during each interval.

The optimization problem in OPAC-2 is stated as follows: For each stage, given the initial queues on each approach and the arrivals for each interval of the stage, determine the switching times, in terms of intervals, which yield the least delay to vehicles over the whole stage.

The procedure used for solving the problem is an optimal sequential constrained search method. It is an exhaustive search of all possible combinations of valid switching times within the stage to determine the optimum set. Valid switching times are constrained by minimum and maximum phase durations.

Rolling Horizon Approach: ROPAC

Although the OPAC-2 optimization procedure lends itself more readily to operation in real time than OPAC-1, it still requires knowledge of arrivals over the whole stage. Depending on minimum phase durations, the stage might be 1 or 2 min long. Obtaining actual arrivals over this length of time might be difficult. However, OPAC-2 could be imple-

mented with a traffic prediction model that predicts the traffic pattern over the entire stage. Although using a prediction model might simplify the optimization, research and experimentation with predictors has shown that they are less effective than historical data and are unreliable as estimators of traffic arrival patterns. In essence, it is difficult to predict what will happen during the next cycle based on what happened during the previous cycle (*3*).

In order to use only available flow data without degrading the performance of the optimization procedures, a "rolling horizon" concept was applied to the OPAC-2 algorithm; it was renamed ROPAC (*5*). In this version, the stage length consists of k intervals. The stage is called the projection horizon (or simply horizon) because it is the period over which traffic patterns are projected and optimum phase change information is required. The horizon is typically equal to the average cycle length. With intervals of 4 sec and an average cycle length of 60 sec, the horizon would be approximately 15 intervals.

From detectors placed upstream of each approach, actual arrival data for r intervals can be obtained for the beginning, or head, portion of the horizon. For the remaining $k - r$ intervals, the tail of the horizon, flow data may be obtained from a model. A simple model consists of a moving average of all previous arrivals on the approach. An optimal switching policy is calculated for the whole horizon, but only those changes occurring within the head portion are actually being implemented. Thus, there is a chance for dynamically revising the decisions when more recent real-time data are being obtained.

It is important that the detectors be placed well upstream of the intersection (10 to 15 sec travel time) in order to obtain actual arrival information over the head period. As indicated earlier, traffic prediction models have proven unreliable in determining optimum signal timing. Knowing actual arrivals, however, allows for the exact calculation of delay based on particular phase change decisions. Hence, it is important to have actual arrival data over the period for which phase changes will be implemented.

At the conclusion of the current head period, a new projection horizon containing new head and tail periods is defined, with the new horizon beginning at (rolled to) the termination of the old head period. The calculations are then repeated for the new projection horizon. Figure 1 illustrates the procedure. The roll period can be any multiple number of steps, including one. It does not necessarily have to equal the head period. A smaller roll period implies more frequent calculations and, generally, closer to optimum (i.e., ideal) results.

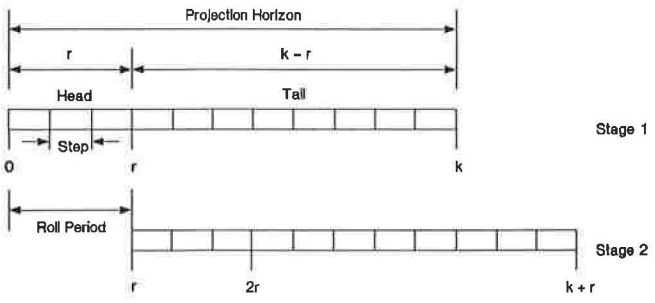


FIGURE 1 OPAC algorithm rolling horizon concept.

Real-Time OPAC: OPAC-RT

The real-time traffic control system developed for implementing the OPAC strategy uses the OPAC software as the signal timing optimization algorithm. The system developed for this study is designated as the Real-Time OPAC Traffic Signal Control System (OPAC-RT).

OPAC-RT Version 1.0: Two-Phase Operation

Version 1.0 of the OPAC-RT Traffic Signal Control System uses the ROPAC software with no modifications or enhancements that would affect its phase change decisions. The primary objective of Version 1.0 of the OPAC-RT system is to effectively control the signal timing at a two-phase, fully actuated, isolated intersection.

OPAC-RT Version 2.0: Dual-Ring, Eight-Phase Operation

Based on the observed performance of Version 1.0, various enhancements were identified. These enhancements were expected to increase its effectiveness and make it more generalized in terms of the locations for which the system could be used. These enhancements were incorporated into Version 2.0. The primary objective of Version 2.0 of the OPAC-RT system is to effectively control the signal timing at an isolated intersection controlled by a dual-ring, eight-phase controller. Only the major phases, typically the through phases, are actually controlled by the system. The minor phases, typically the left-turn phases, are treated by OPAC as part of the intergreen period. The minor phases are controlled by the usual "gap out/max out" strategy.

The system collects vehicle arrival information from upstream detectors as well as signal indications, which are supplied as inputs to a modified version of the ROPAC strategy. The system then implements the switching decisions produced by the optimization algorithm. The system also stores system conditions throughout the control period. This information includes phase returns, HOLD release times, walk requests, the time of the occurrence of any errors, detector occupancies, and arrival patterns. A detailed description of the OPAC-RT software structure is given elsewhere (*10*).

EVALUATION OF REAL-TIME OPAC STRATEGY

Three field tests were conducted to evaluate OPAC-RT: two for evaluating Version 1.0 and one for Version 2.0. Each field test consisted of two phases. In the first phase, the values of three parameters required by the algorithm were fine-tuned to yield the best possible performance. The parameters are

- Saturation flow per approach (Field Tests 1 and 2) or per phase (Field Test 3),

- Travel time in seconds from the OPAC-RT detectors to the stop line, and
- Horizon length.

The horizon is the time period over which the OPAC-RT optimization algorithm calculates its switching decisions. It consists of a head and a tail portion. In the head portion of the horizon, the algorithm has available real-time vehicle arrival information. In the tail portion, the flows are estimated from previous measurements. This estimation requires a "smoothing factor" which was calibrated prior to the field tests.

The second phase of the test plan was performed as a before-and-after study. Two measures of effectiveness were selected for comparing the operation of the intersection under full-actuated and OPAC-RT control. These were delay, defined as vehicle-seconds of delay per vehicle, and percentage of stopped vehicles. Average cycle lengths under both modes of operation were also recorded and compared.

In the before study, a well-timed actuated controller was in control of the intersection. The test plan called for the collection of volume counts, stops, and delay data on each of 3 days for approximately 3 hr each day. Data were to be collected in 25-min segments, yielding 18 sets of observations. An observation consisted of volume counts, stops, and delay by approach per cycle. Delay was calculated by recording the number of stopped vehicles on each intersection approach at a fixed data collection time interval and multiplying this number by the time interval.

The after study was conducted in the same manner, with the OPAC-RT algorithm in control of the intersection. The plan specified that the before and after data collection would be conducted on the same days of the week. Allowances were made for adverse conditions such as bad weather so that a particular field test could be completed in 2 weeks, even if the before and after data were not collected on the same days of the week.

In evaluating the OPAC-RT system, a simple comparison of the observations of delay and percent stops under full-actuated and OPAC-RT control was determined to be inadequate to develop a full understanding of the algorithm's operation. It was postulated that the performance of the OPAC-RT algorithm would be better understood if measured as a function of volume conditions on the major and minor streets. This approach to the evaluation was based on some inherent characteristics of both Version 1.0 of the OPAC algorithm and the actuated controller. Version 1.0 requires at least one phase change per horizon length. Thus, even if there are no calls from the side street, OPAC-RT services the side street once per horizon and causes delay to the major street. The actuated controller does not change phases unless there is a call from the side street, causing no delay to the major street when there is no volume on the side street. This handicap to the OPAC operation was only later removed for Field Test 3.

On the other hand, an actuated controller cannot distinguish one call on the minor street from many calls on the major street. In other words, if flows are high on the major street and low on the minor street, the actuated controller may cause excessive delay to the major street because it assigns the same value to the single vehicle on the minor street as it assigns to the many vehicles on the major street. Such myopia is an inherent handicap of vehicle-actuated control. In this respect, the OPAC-RT algorithm is "smarter" because it counts the many vehicles and considers them more important than a single vehicle on the minor street.

Moreover, because there is no preference to "major" or "minor" directions in the OPAC terminology, what actually counts is the total volume approaching the intersection. It became clear that the performance measure (average delay) should be evaluated as a function of the total volume. To seek a meaningful functional relationship between the two variables, a further decision had to be made about what basic data unit to use in the statistical analysis. Data were collected on a per cycle basis; however, because the phase lengths vary with the changing arrival rates, the cycle length is also variable. Consequently, we have a chain reaction in which the volume per cycle and the average delay are not independent. The statistical data unit must consist of a fixed-time interval that is independent of the resulting control parameters. Therefore, it was decided to aggregate all data into a constant interval approximately 10 min long (600 sec.). These basic data units were then used to compare the performance of OPAC-RT versus actuated control and to derive regression models of average delay versus total volume.

Field Test 1

The location chosen for the first field test, the initial testing of the on-line OPAC strategy, was the intersection of North George Mason Drive and North 16th Street in Arlington, Virginia. This intersection offers low to moderate volume levels and is controlled by a two-phase, full-actuated traffic signal. The north and south approaches are multilane. Detectors supplying information to the OPAC-RT system were located approximately 600 ft (180 m) from the stop lines. Call-only detectors at the stop line in each lane were 6 ft by 50 ft (1.8 m by 15 m). The intersection is normally operated under loop/occupancy control using stop-bar detectors.

OPAC-RT input parameters were calibrated as specified by the test plan. Saturation flow rates were calculated by measuring the discharge rate of standing queues on each approach with adjustments for start-up lost times. Because some of the approaches had very low volumes, the saturation flow rates are thought to be overestimated. This was a further handicap for the OPAC algorithm, because it alone explicitly requires these rates in the optimization procedure. This effect was partly mitigated by the fact that the lowest-volume approaches have only a minor influence on the overall intersection performance optimization. Travel times from the OPAC-RT detectors to the stop lines were all 12 sec. The horizon length was 12 steps or 48 sec. The step size was 4 sec. The head period of the projection horizon was 3 steps.

Delay

Figure 2 shows the scatter plots of the aggregated data under actuated and OPAC-RT control. A hyperbolic model was chosen for the regression analysis. The regression results, summarized in Table 1, indicate a weak correlation, probably

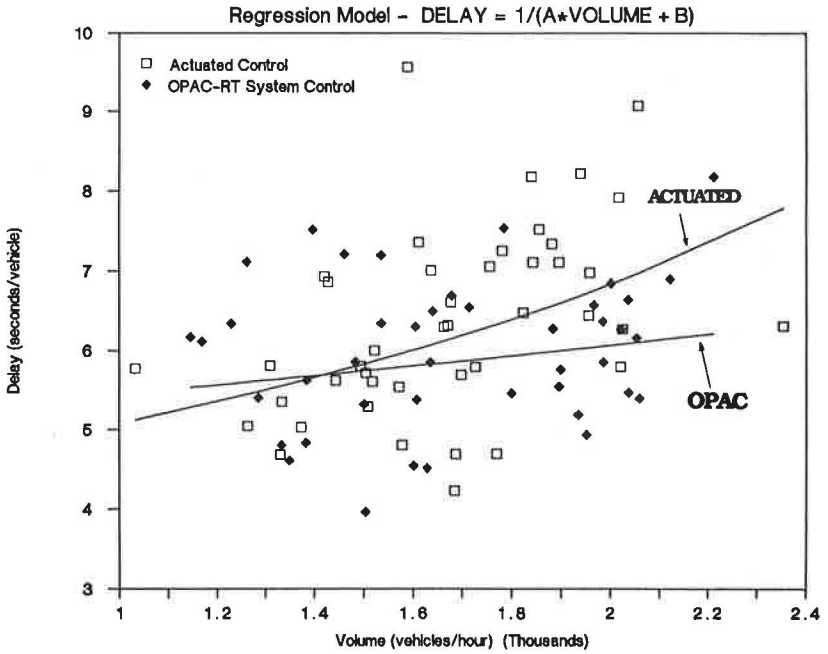


FIGURE 2 Field Test 1 aggregated delay data—regression results.

TABLE 1 SUMMARY OF REGRESSION RESULTS FOR DELAY, FIELD TEST 1, NONLINEAR (HYPERBOLIC) REGRESSION—DELAY = $(A_0 * \text{VOLUME} + A_1)^{-1}$

REGRESSION COEFFICIENTS

VARIABLE	ACTUATED CONTROL	OPAC-RT CONTROL
Volume (A_0)	-0.00005	-0.00001
Constant (A_1)	0.2427	0.2018

OTHER STATISTICS

	ACTUATED CONTROL	OPAC-RT CONTROL
R SQUARED	0.1916	0.0386
SAMPLE SIZE	43	42

because of the small range in values for the independent variable, the flows. These data were later combined with those from Field Test 2, and the resulting regression equation proved much more satisfactory. The average delay for the aggregated data under actuated control was 6.29 sec. Under OPAC-RT system control, the average delay was 6.04 sec. On the average then, OPAC-RT yielded a 3.9 percent reduction in delay to vehicles.

The results of the first field test indicate that OPAC-RT Version 1.0 has the potential to improve the operation of isolated intersections. At the lower volumes at this intersection, the average delays under actuated and OPAC-RT system control were essentially identical. At the higher volumes, however, the difference became larger. It must be recognized in this case, however, that the OPAC-RT Version 1.0 worked under two handicaps: the requirement that the minor side street be serviced even though there are no vehicle calls, and the upward bias of the saturation flow values. To make more definitive statements regarding the operation of the OPAC-RT system, an analysis of an intersection with a wider range in flows—particularly higher flows—was required. This requirement was taken into consideration in selecting the second field test site.

Percent Stops

In a manner similar to the delay data, the stops data (total vehicles and stopping vehicles) were aggregated into time periods of approximately 600 sec. Despite the fact that OPAC-RT Version 1.0 does not optimize for stops, the percentage of stopping vehicles was decreased under OPAC-RT control on the average of 1.6 percent. This decrease is almost insignificant but may be due, in part, to the slightly higher average cycle length observed under OPAC-RT system control. In general, research has shown that shorter cycle lengths increase stopping percentages.

Average Cycle Length

On the average, the cycle lengths under actuated and OPAC-RT control were similar. Under actuated control, the average cycle length was 40 sec. Under OPAC-RT system control, the average cycle length was 44 sec. Despite the slightly higher average cycle length, both delay and stops were decreased on the average under OPAC-RT system control. Although the improvements in performance were modest, it must again be recognized that several requirements of the Version 1.0 system, including required servicing of the minor side street and equal minimum and maximum greens for each phase, degrade the performance of the OPAC algorithm, particularly at low volumes. Another source for degradation is the estimated saturation flow rate.

Field Test 2

The location chosen for the second field test was the intersection of Flowing Wells Road and Prince Road in Tucson, Arizona. At the time of the second field test, it was a two-phase semiactuated intersection with both phases serving multilane approaches. This intersection offered moderate to high volume levels. As part of a computer-controlled network, the intersection was placed off-line during the data collection period.

The remote OPAC detectors were located between 600 and 650 ft from the stop lines. The OPAC-RT input parameters were calibrated according to the test plan. Travel times from the OPAC-RT detectors to the stop lines were all 12 sec. The horizon length was 15 steps, or 60 sec. The step size was 4 sec. The head period of the projection horizon was 3 steps.

Delay

Figure 3 shows the scatter plots of the aggregated data under actuated and OPAC-RT control. Of several linear and nonlinear models, a hyperbolic model yielded the best results for the regression analysis. The regression results, summarized in Table 2, indicate weak correlation. But, when combined with the data from Field Test 1, the resulting regression equation provides a satisfactory explanation. The average delay for the aggregated data under actuated control was 15.81 sec. Under OPAC-RT system control, the average delay was 13.29 sec. In addition, the volumes under OPAC-RT control were found to be higher by an average of 4.12 percent. On the average then, OPAC-RT yielded a 15.94 percent reduction in delay to vehicles despite an increase in volume.

As noted earlier, the intersection in Tucson is part of a computer-controlled network and was placed off-line during the field study. Traffic patterns at the intersection were platooned because surrounding intersections were still on-line.

TABLE 2 SUMMARY OF REGRESSION RESULTS FOR DELAY, FIELD TEST 2, NONLINEAR (HYPERBOLIC) REGRESSION—DELAY = $(A_0 * \text{VOLUME} + A_1)^{-1}$

REGRESSION COEFFICIENTS

VARIABLE	ACTUATED CONTROL	OPAC-RT CONTROL
Volume (A_0)	-8.39×10^{-6}	-2.09×10^{-5}
Constant (A_1)	0.09494	0.1568

ANALYSIS OF VARIANCE - ACTUATED REGRESSION EQUATION

	SUM OF SQUARES	MEAN SQUARES	F	SIGNIFICANCE OF F
Regression	0.00033	0.00033	3.13238	0.0876
Residual	0.00299	0.00011		

ANALYSIS OF VARIANCE - OPAC-RT REGRESSION EQUATION

	SUM OF SQUARES	MEAN SQUARES	F	SIGNIFICANCE OF F
Regression	0.00188	0.00188	9.48554	0.0053
Residual	0.00456	0.00020		

OTHER STATISTICS

	ACTUATED CONTROL	OPAC-RT CONTROL
R SQUARED	0.1006	0.29199
SAMPLE SIZE	30	25

The reduced delay under OPAC-RT control indicates the degree to which the OPAC strategy responds to platooned traffic.

The results of the second field test support the findings from the first field test and indicate that OPAC-RT Version 1.0 has considerable potential for improving the operation of isolated intersections. It must be recognized, however, that the OPAC-RT Version 1.0 requirement that the minor side street be serviced, even though there were no vehicle calls, degraded OPAC's performance in this case as well. Despite the considerable reduction in delay, it was postulated that the

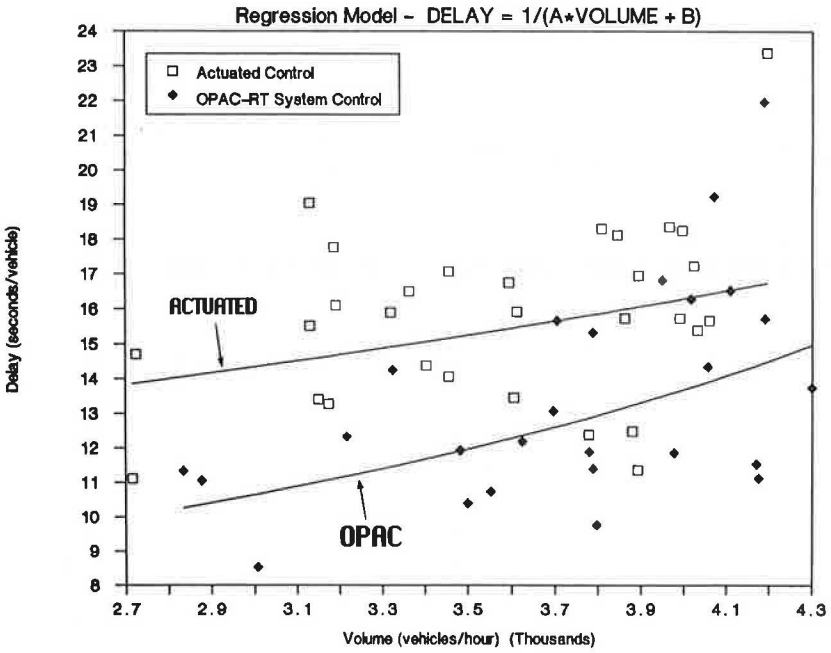


FIGURE 3 Field Test 2 aggregated delay data—regression results.

OPAC-RT system could do much better if certain constraints on the algorithm were removed. Some of these constraints were later removed, and the resulting system was evaluated in Field Test 3.

Percent Stops

Similar to the per cycle delay data, the per cycle stops data (total and stopping vehicles) were also aggregated into intervals of approximately 600 sec. Despite increased volumes and the fact that Version 1.0 of the algorithm does not optimize stops, OPAC-RT increased stops by only 3.9 percent. The increase in stops may be due, in part, to the significantly shorter average cycle length under OPAC-RT control. In general, past research has shown that shorter cycle lengths increase stopping percentages.

Average Cycle Length

On the average, the cycle lengths under actuated and OPAC-RT control were very different. Under actuated control, the average cycle length was 86 sec. Under OPAC-RT system control, the average cycle length was 55 sec. This difference in cycle length was expected because the OPAC algorithm forces the termination of a phase at the calculated optimum time by issuing a FORCE OFF command to the controller. The actuated controller may dwell in a particular phase if there is sufficient demand because the variable green interval will be extended by the passage of each detected vehicle.

Field Tests 1 and 2 Combined

An additional analysis was conducted on the delay data combined from the first and second field tests. To obtain a common basis for the two data sets, the volume/saturation flow ratio was used as the independent variable. Figure 4 shows the aggregated data and the resulting regression equations. Table 3 summarizes the regression results. As indicated in the table, the hyperbolic regression equations were adequate models for the relationship between delay and volume. These models provide a much better fit to the combined data than do the regression models for the individual field tests. This is due primarily to the greater number of data points and the wider range available for the independent variable in this case. The average delay under actuated control for the combined data was 8.13 sec. The average delay under OPAC-RT control for the combined data was 7.41 sec. Thus, there was a 9 percent decrease in average delay under OPAC-RT control.

The combined analysis supports the conclusions from the first and second field tests that the real-time OPAC system has considerable potential for decreasing delay at isolated intersections. As expected, under low volume conditions the benefits of the OPAC-RT system over the actuated controller are small; however, the benefits increase dramatically at higher volumes. As can be seen in the figure, the reduction in delay at volume/capacity ratios of 0.90 and up can be 30 percent or higher.

Field Test 3

The location for the third field test was also at the intersection of Flowing Wells Road and Prince Road in Tucson, Arizona. This intersection was chosen because it was being converted to eight-phase, dual-ring operation and the OPAC detectors required for the OPAC-RT system were already in place. As indicated in the discussion of the second field test, the intersection is part of a computer-controlled network and was placed off-line during the field study.

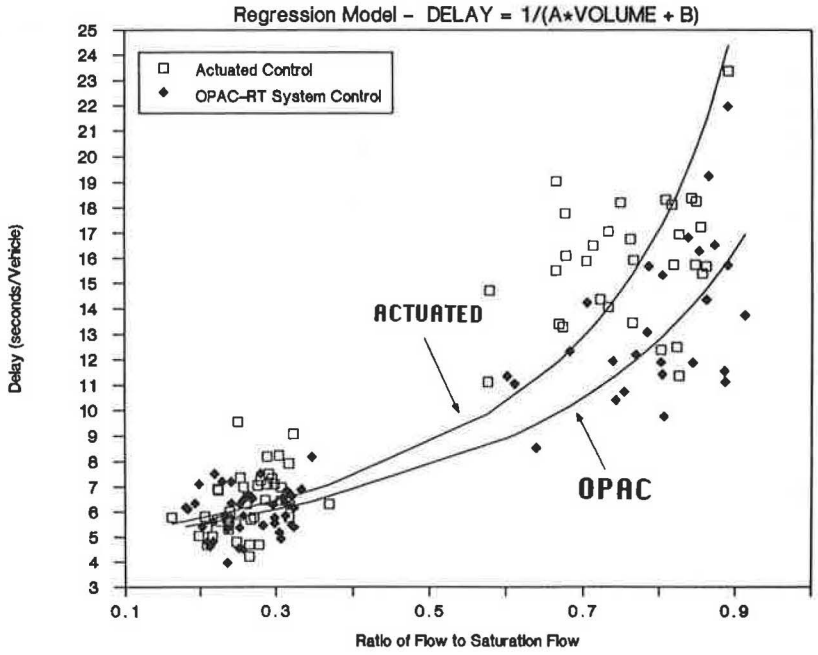


FIGURE 4 Aggregated delay data for Field Tests 1 and 2 combined—regression results.

TABLE 3 SUMMARY OF REGRESSION RESULTS FOR DELAY, FIELD TESTS 1 AND 2 COMBINED, NONLINEAR (HYPERBOLIC) REGRESSION—DELAY = $(A_0 * \text{VOLUME} + A_1)^{-1}$

REGRESSION COEFFICIENTS

VARIABLE	ACTUATED CONTROL	OPAC-RT CONTROL
Volume (A_0)	-4.87×10^{-5}	-4.17×10^{-5}
Constant (A_1)	0.2427	0.2378

ANALYSIS OF VARIANCE - ACTUATED REGRESSION EQUATION

	SUM OF SQUARES	MEAN SQUARES	F	SIGNIFICANCE OF F
Regression	0.16964	0.16964	293.3791	0.0000
Residual	0.04047	0.00058		

ANALYSIS OF VARIANCE - OPAC-RT REGRESSION EQUATION

	SUM OF SQUARES	MEAN SQUARES	F	SIGNIFICANCE OF F
Regression	0.1268	0.1268	219.3086	0.0000
Residual	0.0370	0.00058		

OTHER STATISTICS

	ACTUATED CONTROL	OPAC-RT CONTROL
R SQUARED	0.8074	0.7741
SAMPLE SIZE	72	66

Figure 5 shows the physical layout of the intersection. The operation was eight-phase with lagging left turns on all approaches. The left-turn phases were treated by OPAC-RT as part of the intergreen period. The logic maintains an exponentially smoothed average duration for each minor (typically, left-turning) phase for use in calculating the optimum durations of the major (typically, through) phases. Call-only and OPAC detectors and their locations with respect to the stop lines at the intersection are shown in the figure.

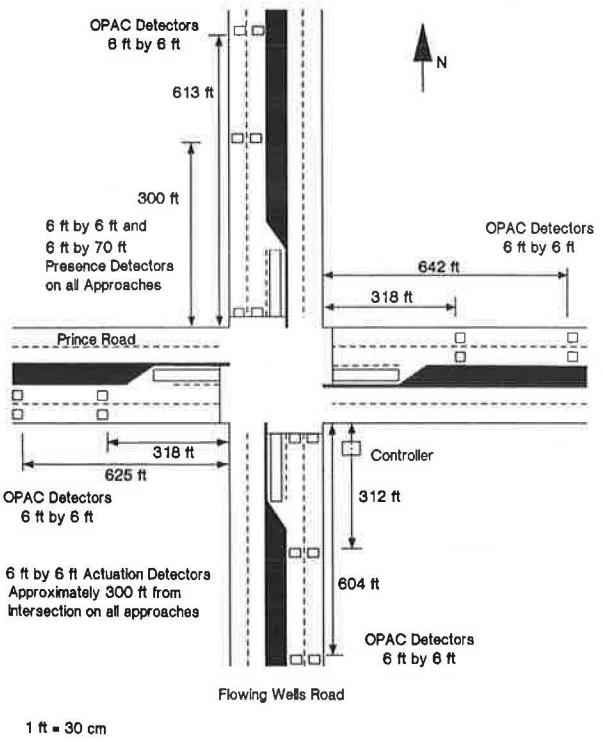


FIGURE 5 Field Test 3 site—Tucson, Arizona.

Saturation flows were recalculated for the third field test. Because left-turning vehicles were permitted during the through phases, it was impossible to determine saturation flows for the turning phases according to the procedures of the test plan. Saturation flows of 1,800 vehicles per hour were assumed for these minor phases. Travel times from the OPAC-RT detectors to the stop lines were all 12 sec. In order to accommodate the new phasing, the horizon length was increased to 20 steps, or 100 sec. The step size was 5 sec. The head period of the projection horizon was 3 steps. Also, the passage time for the through phases was 5 sec. under actuated control. The horizon length was extended to accommodate the new eight-phase configuration.

Delay

Figure 6 shows the aggregated field test data and the hyperbolic curves resulting from the regression analysis. Table 4 summarizes the regression and analysis of variance results. As indicated by the table, the hyperbolic models show weak correlation between delay and volume. As with both Field Tests 1 and 2, this is probably due to the limited range of volume data observed during the field test.

The average delay under OPAC-RT control was 19.23 sec. The average delay under actuated control was 20.83 sec. Operation under OPAC-RT yielded a 7.7 percent reduction in delay overall. The reduction in delay under OPAC-RT control indicates the responsiveness of the algorithm to platooned traffic. As indicated earlier, the intersection is part of a system. Although this intersection was off-line during the field study, the surrounding intersections remained on-line, producing platooned traffic at the intersection of Prince Street and Flowing Wells Street.

The benefits of OPAC-RT, with respect to delay, are not as impressive as those observed during Field Test 2. However, the results do indicate that the enhanced OPAC algorithm has the potential for improving the operation of isolated intersections. As indicated earlier, Version 2.0 uses several averaging functions for information required by the signal timing optimization algorithm. For example, the minor (typically, left-turning) phases are treated as part of the intergreen period. Hence, the algorithm must have estimates of the durations of these phases in order to perform its optimization. The estimates of these phases are made via an exponential smoothing function using a user-input smoothing factor. Errors in these estimates could greatly degrade the performance of the intersection under OPAC-RT control. It is expected that better calibration procedures of the various user-input parameters and smoothing values will further increase the benefits of the OPAC-RT system.

Percent Stops

Although stops were included in the optimization function, OPAC-RT increased percent stops by an average of 9.5 percent. However, the weighting of stops relative to delay was only 1; in reality, this weighting favors delay. A weighting of 15 or 20 should have caused a decrease in stops. If the trends observed during the three field tests are to be taken as valid,

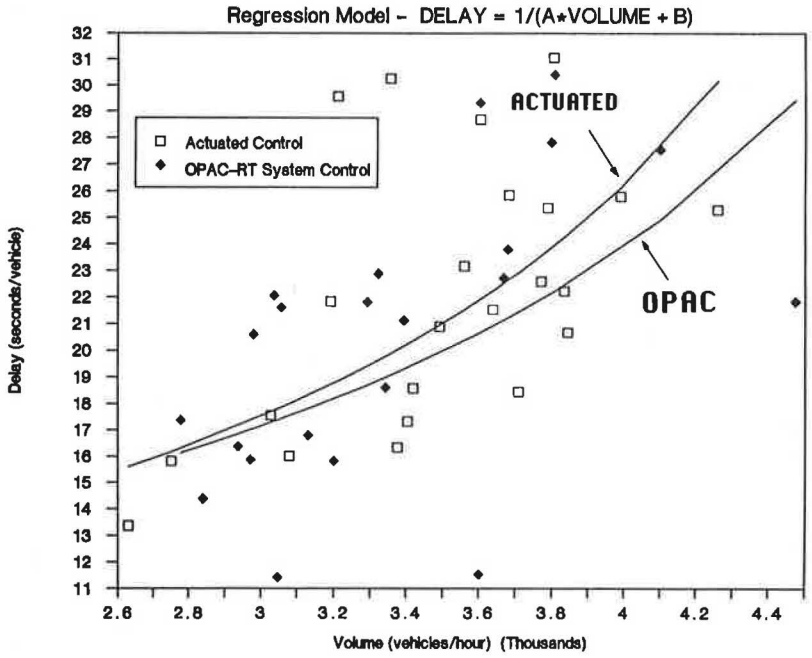


FIGURE 6 Field Test 3 aggregated delay data—regression results.

then shorter cycle lengths reduce delay and increase stops, and longer cycle lengths increase delay and reduce stops. For some combination of weighting factors for delay and stops, both will be optimized. Another factor affecting the performance of the algorithm with respect to stops is the estimation of minor phase (left-turning) volumes. Errors in this estimation may cause errors in the calculation of stops. This could be improved by better calibration procedures.

Average Cycle Length

Again, for the third field test, the cycle lengths under actuated and OPAC-RT control were very dissimilar. Under actuated control, the average cycle length was 110 sec. The average cycle length under OPAC-RT control was only 80 sec. This difference was not unexpected, given the results of the second field test. As indicated, research has indicated that shorter cycle lengths decrease delay and increase stops, and longer cycle lengths tend to increase delay and decrease stops. Because the weighting of stops relative to delay during this field test favored delay, it was expected that the average cycle length under OPAC-RT control would be shorter.

TABLE 4 SUMMARY OF REGRESSION RESULTS FOR DELAY, FIELD TEST 3, NONLINEAR (HYPERBOLIC) REGRESSION—DELAY = $(A_0 * VOLUME + A_1)^{-1}$

REGRESSION COEFFICIENTS

VARIABLE	ACTUATED CONTROL	OPAC-RT CONTROL
Volume (A_0)	-1.90×10^{-5}	-1.65×10^{-5}
Constant (A_1)	0.11397	0.10787

ANALYSIS OF VARIANCE - ACTUATED REGRESSION EQUATION

	SUM OF SQUARES	MEAN SQUARES	F	SIGNIFICANCE OF F
Regression	0.00121	0.00121	15.9983	0.0007
Residual	0.00159	0.00008		

ANALYSIS OF VARIANCE - OPAC-RT REGRESSION EQUATION

	SUM OF SQUARES	MEAN SQUARES	F	SIGNIFICANCE OF F
Regression	0.00108	0.00108	5.81276	0.0256
Residual	0.00372	0.00019		

OTHER STATISTICS

	ACTUATED CONTROL	OPAC-RT CONTROL
R SQUARED	0.4324	0.2252
SAMPLE SIZE	23	22

SUMMARY AND CONCLUSIONS

Three field tests of the real-time OPAC traffic signal control system were conducted. The first two tests evaluated the first version of the OPAC system, which was limited to the control of two phase intersections. Based on the observed performance of this version, various enhancements were identified both to increase the effectiveness of the system and to permit installation of OPAC for control of a broad range of controller phasing configurations. After making the required modifications, the second version was evaluated during the third field test, which was conducted at a site operating with an eight-phase, dual-ring controller.

The principal measures of effectiveness selected for comparison of the performance of an actuated controller and the OPAC systems were delay and percentage of vehicles forced to stop. During the first field test, both delay and percent of stops were decreased when the intersection was under OPAC control. The improvements were modest; on the average, delay was decreased by 3.9 percent and stops were decreased by 1.6 percent. However, the observed volumes during this field test were extremely low and the OPAC algorithm was

handicapped in its operation. More definitive statements regarding the performance of OPAC required the analysis of stops and delays at higher volume levels.

During the second field test, delay was considerably reduced under OPAC control. On the average, delay was reduced by 15.9 percent despite an increase of 4 percent in average volumes. The percentage of vehicles forced to stop, on the other hand, was increased only by 3.9 percent. Because stops were not an OPAC measure of effectiveness in the first version of the system, and because there was also an increase in volume during OPAC operation, this minor increase in stops was to be expected. During the third field test, which was conducted at an eight-phase intersection, delay was decreased on the average by 7.7 percent and the percentage of stopped vehicles was increased by an average of 9.5 percent.

The first version of the OPAC traffic control system was handicapped by several constraints. Despite these limitations, it has demonstrated a potential for significantly improving intersection performance as measured by delay and percentage of stopping vehicles. The results of the third field test indicate that the enhanced OPAC system also improves the operation of multiphase signalized intersections.

The OPAC strategy represents a new dimension in traffic signal control, the potential of which has not yet been fully realized. It carries out sophisticated optimization in real time and adapts to varying traffic conditions. This study has shown that it works well in a field environment and can provide significant benefits over well-timed actuated controllers. A preliminary economic analysis has shown that the incremental costs associated with its implementation as is can be recovered within less than 1 year of operation (*10*). This conclusion was reached notwithstanding the fact that it is only a first implementation of a previously untested method. It can be expected that further enhancements of the method and, especially, the development of improved calibration procedures will help to further improve the performance of the real-time OPAC system.

Because OPAC is a smart controller, it forms a building block for a distributed intelligence traffic control system. Unlike conventional actuated control logic, the OPAC model can communicate with neighboring controllers so as to form a flexibly coordinated traffic control system (*12*). It can similarly be used for critical intersection control within otherwise fixed-cycle coordinated systems. Herein lies the greatest potential of this method. Further research in these areas is now being conducted.

ACKNOWLEDGMENT

The authors wish to acknowledge the assistance and encouragement provided by Stephen L. Cohen of the Office of Research, FHWA, in the conduct of this research.

REFERENCES

1. *Traffic Control Systems Handbook*. Report FHWA-IP-85-11. FHWA, U.S. Department of Transportation, 1985.
2. N. H. Gartner. On-Line and Off-Line Urban Traffic Control. *Computing in Civil Engineering 1981*. ASCE, 1981, pp. 503–513.
3. N. H. Gartner. *A Prescription for Demand-Responsive Urban Traffic Control*. In *Transportation Research Record 881*, TRB, National Research Council, Washington, D.C., 1982, pp. 73–76.
4. N. H. Gartner. Demand-Responsive Traffic Signal Control Research. *Transportation Research*, Vol. 19A, 1985, pp. 369–373.
5. N. H. Gartner. *Demand-Responsive Decentralized Urban Traffic Control, Part I: Single-Intersection Policies*. Report DOT/RSPA/DPB-50/81/24. U.S. Department of Transportation, 1982.
6. N. H. Gartner, M. H. Kaltenbach, and M. M. Miyamoto. *Demand-Responsive Decentralized Urban Traffic Control, Part II: Network Extensions*. Report DOT/OST/P-34/85/009. U.S. Department of Transportation, 1983.
7. N. H. Gartner. Development and Testing of a Demand-Responsive Strategy for Traffic Signal Control. *Proc., 1982 American Control Conference*, June 1982, pp. 578–583.
8. N. H. Gartner. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983, pp. 75–81.
9. H. Chen, S. L. Cohen, N. H. Gartner, and C. C. Liu. Simulation Study of OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation and Traffic Theory* (N. H. Gartner and N. H. M. Wilson, eds.). Elsevier Science Publication, 1987, pp. 233–249.
10. Farradyne Systems, Inc. *Evaluation of the Optimized Policies for Adaptive Control Strategy*. Report FHWA-RD-89-135. FHWA, U.S. Department of Transportation, 1989.
11. D. P. Bertsekas. *Dynamic Programming: Deterministic and Stochastic Models*. Prentice-Hall, 1987.
12. N. H. Gartner. Adaptive Control of Traffic Signal Networks. *Proc., 1st International Conference on Applications of Advanced Technologies in Transportation Engineering*, San Diego, Calif., ASCE, 1989, pp. 373–377.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Knowledge-Based System for Adaptive Traffic Signal Control

S. MANZUR ELAHI, A. ESSAM RADWAN, AND K. MICHAEL GOUL

Signal Control at Isolated Intersection (SCII) is a knowledge-based expert system prototype. It represents an application of expert systems to adaptive signal control. The first generation of the prototype was developed at Arizona State University in 1987. The second-generation SCII can handle three types of intersection geometries. At the end of each signal cycle, SCII determines the performance of the controller operation during that cycle. In case of an unsatisfactory performance, SCII determines the appropriate cycle length, phasing pattern, and split. It also updates the cycle length and phasing scheme based on traffic demand. Different tests validated and calibrated the prototype using a simulation approach. A 20-min traffic volume data set was used to simulate a pretimed controller, an actuated controller, and the operations suggested by SCII. The tests demonstrated the potential of this prototype to reduce delay at isolated intersections.

Existing signal types are mainly classified as pretimed, semiactuated, and full-actuated, including volume-density controllers. In the pretimed operation, the intervals of signal phases are predetermined. In the semiactuated operation, the major road phase is nonactuated and the minor road phase is actuated. In the full-actuated operation, all phases are controlled by actuations created by the detectors. The phasing patterns and their associated intervals can vary to a large extent based on the traffic demand.

Full-actuated traffic operations may have an added feature, volume-density control, which is programmed to operate with variable minimum green time and vehicle extension intervals. In this feature, the phase interval can vary based on a more complex evaluation of traffic conditions.

Another new type of operation, called adaptive control, is still in a research stage. It represents a real-time, demand-responsive traffic signal control and performs the signal operation based on existing traffic conditions. Different strategies have been developed for adaptive signal control.

STATEMENT OF PROBLEM

The conventional type of signal controllers have several limitations. The pretimed controller has a fixed cycle time, split, and phasing pattern over a certain period of time. It cannot respond to fluctuations in traffic demand. Actuated controllers are more flexible in handling traffic fluctuations. However, their performance deteriorates under heavy traffic conditions when some phases reach their maximum duration and the green period per phase does not remain proportional to the traffic demand (*1*). Also, conventional controllers are often preset according to the traffic demand predicted by time of day; hence, some unanticipated demands cannot be handled. To improve signal performance, on-line signal control is desirable because it can adjust the signal plan based on real-time traffic conditions.

There are many good reasons for using artificial intelligence techniques to help solve traffic engineering problems. The design of signalized intersections involves many decision-making processes, for example, what kind of signal operation should be used, what signal phases should be used, and finally, what calculations should be made to find the timing scheme for the cycle. The application of knowledge-based expert system (KBES) technology can be a logical approach to handle this overall problem. The knowledge base of a KBES designed to capture the existing traffic condition, along with historic data, can generate the basis for signal control. With the ability to learn, it can continuously update its knowledge base and adapt to variations in traffic flow with the help of an inference engine.

A KBES has the ability to perform tasks using a human-like decision process in a limited domain. Goul et al. (*2–4*) identified the need for a KBES orientation to real-time signal control operation at isolated intersections and developed an expert system prototype in 1987 at Arizona State University. Radwan et al. (*5*) summarized the experiences from the project, especially during the verification of the system. The expert system prototype, Signal Control at Isolated Intersection (SCII), was developed for a specific intersection geometry and phasing pattern. In this paper, it is called first-generation SCII, or SCII-1. This paper reports on research involved in enhancing SCII. The product of this research is called second-generation SCII, or SCII-2.

LITERATURE REVIEW

Delay Equations

Webster (*6*) pioneered the formulation of a delay model for fixed-time traffic signal operation. This model has been used extensively in computer software. Modifications were made to Webster's formula to estimate stopped-time delay at pretimed signals in the *Highway Capacity Manual* (*7*). The manual describes a step-by-step procedure to find capacity and level of service for signalized intersections. Its measure of

S. M. Elahi, Bureau of Traffic Services, District of Columbia Department of Public Works, 2000 14th St., N. W., 7th Floor, Washington, D.C. 20009. A. E. Radwan, Department of Civil and Environmental Engineering, University of Central Florida, Orlando, Fla. 32816. K. M. Goul, Department of Decision Information Systems, Arizona State University, Tempe, Ariz. 85287.

effectiveness is average stopped-time delay, which is calculated by the following equation:

$$d = 0.38C\frac{[1 - (g/C)]^2}{[1 - (g/C)X]} + 173X^2[(X - 1) + \sqrt{(X - 1)^2 + 16X/c}] \quad (1)$$

where

d = average stopped delay per vehicle for the lane group (sec/veh),
C = cycle length (sec),
g = effective green time (sec),
X = degree of saturation = v/c ratio for the lane-group,
c = capacity of the lane-group (veh/sec), and
v = vehicle flow rate (veh/sec).

The first term in Equation 1 denotes delay for uniform arrivals; the second term denotes incremental delay for random arrivals. Because the vehicle arrival patterns may not really be random, the second term is subject to a correction for the signal progression and other factors. The *Highway Capacity Manual* discusses five different arrival patterns and assigns different correction factors to each of them.

Equation 1 yields reasonable values for values of X between 0 and 1.0. It may be used with caution for values of X up to 1.2.

Queue Equations

Queue length, like delay, is an important measure of effectiveness. Second-generation SCII determines the system effectiveness combining both measures. Different queue equations are available, including Webster's (*6*). Cronje (*8*) conducted research to assess existing formulas for delay, stops, and overflow. He defined overflow as the queue length at the end of the green phase. He took into consideration the formulas given by Webster (*6*), Newell (*9*), and Miller (*10,11*) and compared the results from these formulas with the results generated from a macroscopic computer simulation model. He observed that, among the overflow equations assessed, Miller's second overflow equation produced values closest to the simulated values. This finding is adopted in this study; therefore Miller's second equation was used to calculate the queue length for undersaturated conditions. The equation is as follows:

$$Q_O = \exp[-(4/3) * (\lambda * C * s)^{0.5} * (1 - X)/X]/[2(1 - X)] \quad (2)$$

where

$\exp[z] = e^z$,
λ = G/C ratio,
G = effective green (sec),
s = saturation flow (veh/sec),
v = arrival rate of vehicles (veh/sec),
c = capacity (veh/sec) = $s * \lambda = s * G/C$, and
Q_O = average overflow at the beginning of red.

To calculate the queue length (Q) at the beginning of the green, the number of vehicles that arrived during red (Q_r) should be added to Q_O. For the arrival rate of v and the red interval of r

$$Q_r = v * r = v * (C - G) = v * C * (1 - G/C) = v * C * (1 - \lambda) \quad (3)$$

Finally,

$$Q = Q_O + Q_r \quad (4)$$

The above queue-length model is good for undersaturated conditions only. For near and oversaturated conditions, a deterministic model has been considered (*12*). This model ignores the effects of random variations. This concept is used to develop a method for estimating queue length, referred to as the input-output method. This method assumes a constant rate of vehicle input and output at an intersection. Because vehicles waiting in a queue provide a steady source of input during saturated conditions, this is identified as a suitable method. Let Q_{i-1} be the queue length at the end of a cycle $(i - 1)$. During the Cycle i, the expected overflow in the green interval is $(v - s) * g$ and the additional queue build-up during the red interval is $v * r$. Then, the expected queue length at the end of a Cycle i can be estimated as

$$Q_i = Q_{i-1} + (v - s) * g + v * r \quad (5)$$

where the term $(v - s) * g$ cannot be less than zero.

A combination of Equations 4 and 5 was coded in the prototype SCII-2, where Equation 4 was used for undersaturation and Equation 5 was used for near- and oversaturation.

Expert System

One definition of the term expert system is "An expert system is a computer program that embodies the expertise of one or more experts in some domain and applies this knowledge to make useful inferences for the user of the system" (*13*).

An expert system has two components, namely,

1. Knowledge base, and
2. Inference engine.

The knowledge base contains all facts revealed by the expert for the problem. The inference engine determines the portion of the knowledge base required to solve a particular problem.

The use of expert systems in transportation engineering is relatively new. Work in this field includes CHINA, for highway noise barrier design; DIRECTOR, for urban transportation education; SCEPTRE, for pavement rehabilitation, and TRALI, for traffic signal setting.

Computer Applications and Adaptive Control Strategy

The application of expert systems in adaptive control strategies in designing the operation of signalized intersections is a completely different approach than that used in existing signal control software applications. Most of the available software applications now are applicable to pretimed signal operation. Bullen et al. (*14*) discussed the limitations of a number of available software systems that can be applied for actuated signals. The TEXAS model does not have any optimization capability for signal timing. SOAP84 is capable of providing optimal design but it is highly dependent on Webster's approach (*6*), which is mainly for pretimed signals. NETSIM is another software application that can deal with vehicle-actuated operation, but again it does not have an optimization capability.

VIPAS (*14*) can analyze a wide range of phasing patterns and different types of signals including full-actuated controllers. It is designed for isolated intersection and is able to optimize and analyze a variety of intersections.

Zozaya-Gorostiza and Hendrickson (*15*) developed a KBES prototype, TRALI, to assist traffic engineers in signal timing decision making. TRALI does not have a real-time signal control strategy, but rather provides design parameters. TRALI is coded in the OPS5 programming environment. It uses heuristic rules to determine phase distribution, calculates the optimum cycle length, and estimates delay by Webster's formula.

Much attention is currently being paid to adaptive control strategies. Gartner (*16*) developed a software application called Optimization Policies for Adaptive Control (OPAC) to perform an adaptive control strategy. The first version of OPAC, OPAC-1, uses a dynamic programming approach, which is a mathematical optimization of multistage decision processes. The subsequent version, OPAC-2, uses a simplified approach. The rolling horizon concept was used later. This concept is mainly used by operations research analysts in production-inventory control.

The latest version of OPAC (*17*) is in real time and has been tested in the field. The field results indicated that it performed better than actuated signals, especially at a higher demand level.

Lin et al. (*18*) tested an adaptive control strategy based on predicted data. This strategy did not achieve much success. Lin et al. used three predictors: (a) an exponential smoothing technique, (b) a double exponential smoothing technique, and (c) a pattern search predictor using a heuristic algorithm. None of the predictors consistently produced the smallest prediction error. They selected the exponential smoothing technique for simplicity and compared it with pretimed control. It was found that their strategy did not improve the signal performance.

Lin et al. (*19*) developed another adaptive control strategy, Stepwise Adjustment of Signal Timing (SAST). Its logic divides time into discrete intervals or steps. In each step, a decision is made with available limited future information on whether to terminate the green phase at the end of the step or to extend it beyond the step.

The TRALI expert system (*15*) is identified in assisting a traffic engineer in signal timing decision making. In contrast to this approach, SCII is a tool for signal control at an intersection in real time.

SCII-1 was developed to emulate an adaptive controller using artificial intelligence. Radwan et al. (*5*) described their experiences during the development of SCII and documented different phases of the development framework. Verification of SCII-1 provided important insights about the prototype's performance. Goul et al. (*4*) provided an overview of the prototype.

OVERVIEW OF SCII-2 PROTOTYPE

The first-generation SCII was a milestone for expert system applications in adaptive signal control. The prototype was expanded so that it could perform better and could be applied to a variety of situations. This section describes the modifications and current status of SCII-2.

Modifications

The SCII-1 prototype (*4,5*) was limited with respect to the type of intersection geometry and signal phasing schemes it could accommodate. The major enhancements of SCII-1 are as follows:

1. SCII-1 was developed for a single set of geometric configurations. SCII-2 was made more robust to handle two other configurations of intersection geometries.
2. SCII-1 was designed for adaptive control operation. SCII-2 was expanded to handle both adaptive control and conventional type of signal operations. The adaptive control mode of operation is the subject of this paper.
3. SCII-1 was designed for a fixed four-phase operation. SCII-2 can choose the appropriate phasing pattern for particular traffic demand patterns.
4. In SCII-1, cycle durations are based on total intersection critical volume. In SCII-2, the cycle time logic was modified. A more versatile method was adopted, including saturation flow as another factor.
5. SCII-1 uses vehicle delay as the sole performance measure. In SCII-2, queue length was incorporated as an additional performance measure.
6. The prototype was enabled to perform as a simulator so that SCII-2 can determine the delay and queue length for a given set of data.
7. Initial validation was done to the prototype.

Overall Architecture

SCII-2 has been coded in LISP on a microcomputer. The top level of SCII-2 asks the user whether the mode of operation will be conventional (actuated and pretimed) or adaptive control (Figure 1).

If the user selects the conventional mode, SCII-2 evaluates the signal performance at the end of each cycle and determines the point where signal control needs to be switched from actuated to pretimed and vice versa. SCII-2 uses the methodology outlined in NCHRP Report 233 (*1*) to determine the appropriate mode of signal operation under specific traffic conditions.

The adaptive control operation strategy constitutes the major portion of SCII-2's computer code. SCII-2 evaluates the

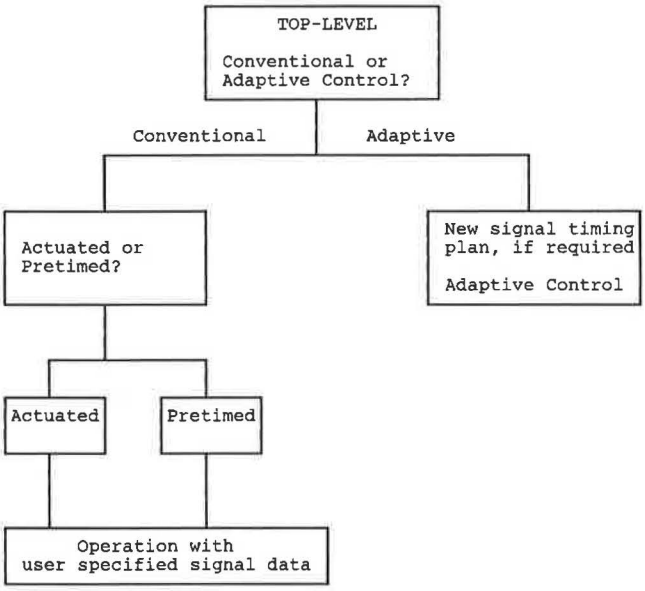


FIGURE 1 Overall architecture of SCII-2.

signal performance at the end of each cycle. If the performance level falls below some user-set threshold value, SCII-2 will determine the parameters for a new cycle.

The knowledge base enables SCII-2 to remember past data that are necessary for critical computations in different modules of this prototype. The heart of the knowledge base is a fusion of multiple data bases; each unit data base represents a specific 15-min interval and stores the related historic traffic volume data broken down by movement type based on signal cycles. SCII-2's knowledge base also includes a data base for signal effectiveness over past cycles. A KBES should possess the capability of "self-adaptation," or learning. In this context, SCII has the ability to adapt its historical data in case of new data.

Adaptive Control Procedure

The basic procedure of SCII-2's operation in the adaptive control mode is described in the following steps:

- *Step 1*: At the end of each signal cycle, this prototype uses the traffic count in that cycle to determine the level of service of each approach and the overall intersection. It follows the procedure described in Chapter 9 of the *Highway Capacity Manual* (7). The prototype also calculates the queue to be expected at the start of the green using Equation 4 or Equation 5, depending on the degree of saturation.
- *Step 2*: The prototype converts the delay and queue length into a performance rating. If this value is higher than the user-set threshold value, it continues with the existing signal timing, skips Step 3, and performs Step 4 directly; if not, it performs Step 3.
- *Step 3*: The prototype forecasts the traffic volume expected in the next cycle based on the "current" cycle traffic volume data and a data base mean of previous volume data. Then it recalculates the signal timing for the forecasted volume. A table look-up procedure is used to determine the cycle length corresponding to the sum of critical volume/saturation flows (*v/s* ratios) on each street. If the user has not predetermined any phasing pattern, SCII-2 will determine a reasonable phasing pattern. Then it calculates the green intervals for the phases based on the *v/s* ratio. It performs the volume, saturation flow rate, and capacity analysis modules using the timing scheme as outlined in Chapter 9 of the *Highway Capacity Manual*.
- *Step 4*: SCII-2 maintains a data base of significant volume data of previous cycles. It checks whether the "current" traffic volume represents a new trend not reflected in the data base. If so, it stores the data in the data base. The prototype loops back to Step 1 to analyze the next cycle.

Intersection Geometries

SCII-2 can handle three different types of four-legged intersections:

1. Each approach with two through lanes and one exclusive left-turn lane,
2. Each approach with three through lanes and one exclusive left-turn lane, and
3. Each approach with one through and one shared left-turn lane.

Phasing Pattern

The second-generation SCII can analyze up to eight different type of phases (Figure 2), of which a maximum of six phases can occur in a single cycle. The user must define whether the left-turn movement is protected or permissive. SCII-2 determines the different lane groups in accordance with the methodology in Chapter 9 of the *Highway Capacity Manual* (7). Then it determines the *v/s* ratio for all the lane-groups. To choose the appropriate phasing pattern, it follows an algorithm that is based on some simple rules of thumb. SCII-2 provides green times to each phase in proportion with the *v/s* ratio. For a particular street (e.g., north-south), if there is a demand for the left-turn phase for both approaches, SCII-2 will select a dual left-turning Phase A with green time proportional to the smaller demand (see Figure 2). If either of the two approaches has demand for the left turn, it will switch to either Phase B or C. Phase D then follows. If both left-turning demands are met simultaneously, SCII-2 will skip Phases B and C and switch directly to Phase D. This is also true for the east-west directions.

Performance Grade of SCII-2

Delay is an important measure of effectiveness (MOE) that has been used in the *Highway Capacity Manual* procedure for signalized intersections. SCII-1 uses delay as the performance measure. To fine-tune a signal setting, an expert may judge the performance of the operation by the visual inspection of the queue length. Because of the significance of queue length as another important MOE, an enhanced performance measure is calculated in SCII-2 combining both MOEs. SCII-2 calculates performance grades for both queue and delay on a 0–100 scale using several heuristics. These two grades are

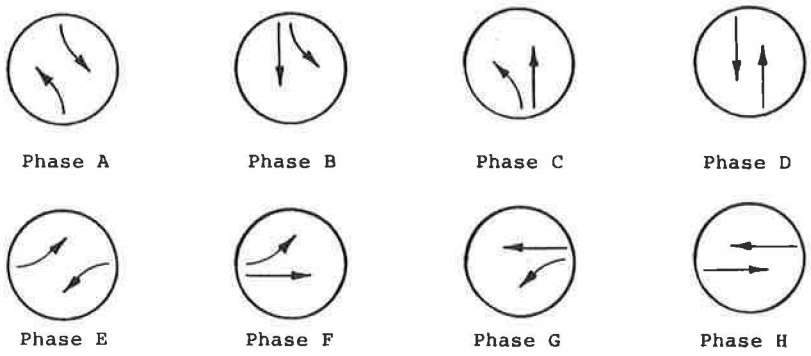


FIGURE 2 Eight-phase signal operation.

then combined to yield one value on the basis of user-chosen weights. The following paragraphs discuss the calculation procedure of queue and delay grades.

Delay Grade

The delay grade is calculated on a 0–100 scale using several heuristics. The 100 points are based on the following components:

1. Approach level of service,
2. Intersection level of service, and
3. Improvements over past cycles.

Queue Grade

Queue length can produce a grade of maximum 100 and minimum 0. A numeric value is calculated for each lane group corresponding to the average queue length per lane prevailing at the start of green period. A maximum value of 100 is achievable for each lane group for a zero queue length. A minimum grade of 0 can occur when queue length exceeds some threshold value. This threshold value is currently set at 25. This value was developed for a downtown area where the signals are 600 to 700 ft apart. Assuming 625 ft for the distance between two signals and 25 ft for the distance between two vehicles front bumper to front bumper, a block can contain 625/25 = 25 vehicles. For this case, a value of 0 is assigned for a queue of at least 25 vehicles and 100 is assigned for a queue length of 0. Intermediate values are calculated by linear interpolation. In the following, QL represents queue length.

$$\text{Rating} = (25 - \text{QL}) * 100/25 \quad \text{for QL} \leq 25$$
$$= 0 \quad \text{for QL} > 25$$

This threshold value can be changed depending on the specific block length.

Combined Grade

Once both the delay grade and the queue grade have been calculated, the combined grade (G) is calculated using user-selected weights W_D and W_Q:

$$G = W_D * G_D + W_Q * G_Q \tag{6}$$

when $W_D + W_Q = 1$.

Cycle Length Logic

In addition, the optimum cycle length is determined from a two-dimensional table as a function of the sum of the critical *v/s* on each road. This is an approximate, but quick and simple method for determining the optimum cycle length. This method was adopted instead of a full-scale optimization procedure, because the prototype needs fast computations for real-time control. Tables were generated using a computer program to search for minimum delay for varying *v/s* on different approaches using the *Highway Capacity Manual* (7) delay equation. Two boundary constraints are placed on the values looked up from these tables. The upper boundary limit for the cycle length is chosen to be 150 sec; a lower limit of 40 sec is selected.

Forecasting Model

The forecasted volume is a weighted combination of the most recent volume and data base mean. A smoothing factor has been introduced to reduce the adverse impact of abrupt rise and fall in the traffic volume. SCII-2 stores a separate data base for each 15-min interval. SCII-2 can keep track of time elapsed and locate the appropriate part of its knowledge base to extract volume information.

Delay Model

SCII-2 calculates delay using the *Highway Capacity Manual* (7) delay equation, documented as Equation 2.

Queue Model

A hybrid model was built to calculate the queue length in the second-generation SCII. Miller's queue model (Equation 4)

is used for undersaturated traffic conditions. For near- and oversaturation, SCII-2 uses the input-output model (Equation 5). The cut-off point for the shift from one method to the other has been determined to be at a degree of saturation $X = 0.98$. A set of calculations was made to determine this boundary point.

Assumptions in SCII-2

Several assumptions were made in the development of SCII-2, as shown in Table 1. These assumptions are not limitations of the prototype, but they can be easily changed in the SCII-2 coding. These assumptions are also realistic for many intersections.

System Requirements

SCII-2 is designed for an IBM-compatible microcomputer. It requires random access memory (RAM) of 512k bytes.

TESTING WITH SIMULATION

Once an expert system prototype is developed, the next logical step is to adjust its parameters. Because SCII-2 is designed to dynamically change the traffic signal settings, a fine-tuning process for the heuristic parameters was needed.

Computer simulation is a cheap and safe tool for numerous "what if" types of analyses. This type of test provides some insights on the model performance. A microscopic computer simulation model (like NETSIM or TEXAS) could be used to test the different settings proposed by SCII-2. Furthermore, the simulation exercise and the results obtained from the simulation runs could provide insight to how well SCII-2 responded to traffic variations and what values to use for particular heuristic parameters. For this study, the TEXAS model was used for testing SCII-2 because this simulation model is developed solely for isolated intersections. Vehicles may be generated using any distribution dictated by the user, such as the shifted negative exponential distribution, the Erlang distribution, or others. The NETSIM computer model uses only a uniform vehicle arrival approach, which may not be suitable for simulating traffic at isolated intersections.

TABLE 1 ASSUMPTIONS MADE IN SCII

Conditions	Elements	Assumption
Geometric Conditions	Area Type	CBD
	Lane widths	12 feet
	Grade	Level
	Parking	Not allowed
Traffic Conditions	Peak hour factor	1.0
	Percent heavy vehicles	0
	Pedestrians	None
	Buses stopping per hour	None
	Arrival type	Totally random
Signalization Conditions	All red	0 sec
	Yellow	3 sec
	Ideal saturation flow	1800 vph

Texas Model

TEXAS is a microscopic simulation model developed at the University of Texas at Austin (*20*). It is a microcomputer-based software application that can make simulation runs for a specified simulation time. Replications of the same run can be made using different random number seeds.

The model can simulate most of the intersection geometric configurations. Pretimed and actuated signal controls can be evaluated.

Initial Fine-Tuning of SCII-2

To perform the simulation tests, traffic data are needed. A 20-min data set previously collected at a local intersection in Phoenix, Arizona, was used for this purpose. Modifications were made to the data to capture the early period of the evening peak when volumes rise to and remain at the peak (Figure 3). These traffic data were used to run SCII-2 under conditions with different parameter settings. The goal behind the simulation test was to check the response of SCII-2 with the varying traffic flow. From experience with this rigorous testing, the following adjustments were made to SCII-2:

1. Separate volume data bases were created for each 15-min period to provide a more appropriate smoothing effect on the forecasted volume.
2. A minimum green time of 9 sec was used for combined left and overlapping phases, rather than individual phases. Another 9-sec value was assigned to each of the two through phases.

Performance of SCII-2

Further tests were done using the same 20-min traffic volume data. Computer simulation was applied to determine the ef-

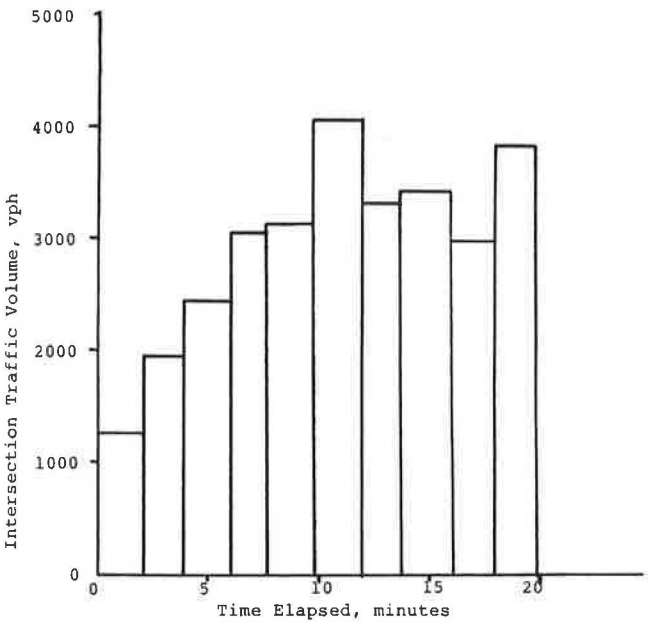


FIGURE 3 Intersection traffic volume.

fectiveness of SCII-2 performance compared with the performances of conventional (pretimed and actuated) controllers.

First, cycle length and splits were determined for a four-phase pretimed operation (dual left, through/right, dual left, and through/right) using basic principles. The cycle length was found to be 110 sec (splits of 19, 38, 10, and 43 sec; yellow of 3 sec; all-red of 1 sec). Therefore, the first run of TEXAS covered a period of 110 sec. The second run covered a period of 110 to 220 sec. Similarly, runs were made for the whole test period with the corresponding traffic volumes. For each run, five replications were made with different random number seeds. Each replication was made for a period of 35 min, with 5 min of warm-up and 30 min of simulation. Traffic delays were noted and the mean was calculated for each run.

Similarly, an actuated setting was simulated with the TEXAS model, using the timing scheme for pretimed operation as the maximum green interval of the actuated operation. Stopped delay over the 20-min period was noted.

A similar approach was adopted for simulating signal settings suggested by SCII-2. Stopped delays were noted for entire 20-min period.

Figure 4 compares pretimed, actuated, and SCII-2 performance for the test case. Comparison between the pretimed controller and SCII-2 indicates that initially the pretimed controller produced a delay of 38 sec/veh, and SCII-2 produced a delay of 10 sec/veh. As the traffic volume gradually increased, the delay increased in both cases. The pretimed controller produced a maximum delay of 63 sec/veh, and SCII-2 produced a maximum delay of 45 sec/veh. SCII-2 was found to perform better than the pretimed controller over the entire 20-min interval.

Comparison between SCII-2 and a full-actuated controller shows that initially SCII-2 performed much better than the actuated controller. With the increase of traffic volume, SCII-2 performance became equivalent to the actuated controller performance.

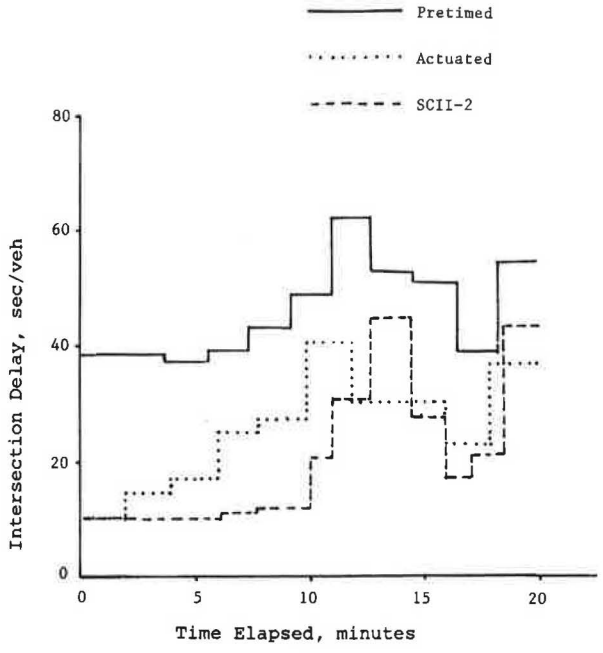


FIGURE 4 Comparison of intersection delays for three controller operations.

CONCLUSIONS

The objective of this research was to modify the initial SCII-1 prototype. The prototype was expanded to operate in both the conventional and adaptive control modes of signal operation. Modifications were made to SCII-1 to operate under different intersection geometric conditions and to determine the appropriate phasing pattern. Most of the SCII-1 heuristics were revised. Queue length was used as a performance measure in addition to stopped delay. Finally, simulations were made to gain confidence in the prototype.

Computer simulation was found to be an effective tool for testing the prototype. The tests using the TEXAS simulation model provided an initial validation of the prototype. It demonstrated the potential of this prototype to reduce delay at isolated intersections using one data set. Rigorous testing of this prototype is required under different scenarios using both simulation and field tests, with further refinement of the prototype as needed in the course of testing.

REFERENCES

1. P. J. Tarnoff and P. S. Parsonson. *NCHRP Report 233: Selecting Traffic Signal Control at Individual Intersections*. TRB, National Research Council, Washington, D.C., 1981.
2. K. M. Goul, K. Moffitt, T. O'Leary, and A. E. Radwan. *The Art of Validating Expert Systems: The Project SCII Experience*. Technical Report 87-2. Decision Systems Research Center, College of Business, Arizona State University, Tempe.
3. K. M. Goul, A. E. Radwan, T. O'Leary, and K. Moffitt. *Project SCII: A Real-Time Knowledge-Based Expert System for Traffic Control*. Decision Systems Research Center, College of Business, Arizona State University, Tempe.
4. K. M. Goul, T. J. O'Leary, and A. E. Radwan. Expert Systems for Traffic Signal Control. *Proc., Microcomputer Applications in Transportation—II International Conference*. ASCE, New York, 1987, pp. 629–638.
5. A. E. Radwan, K. M. Goul, T. J. O'Leary, and K. E. Moffitt. A Verification Approach for Knowledge-Based Systems. *Transportation Research*, Vol. 23A, No. 4, 1989.
6. F. V. Webster. *Traffic Signal Settings*. Road Research Technical Paper 39, Road Research Laboratory, Her Majesty's Stationery Office, London, England, 1958.
7. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
8. W. B. Cronje. Analysis of Existing Formulas for Delay, Overflow, and Stops. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983.
9. G. F. Newell. Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. *SIAM Review*, Vol. 7, 1965.
10. A. J. Miller. Settings for Fixed-Cycle Traffic Signals. *Operational Research Quarterly*, Vol. 14, 1963.
11. A. J. Miller. *The Capacity of Signalized Intersections in Australia*. ARRB Bulletin 3, Australian Road Research Board, 1968.
12. G. F. Newell. *Applications of Queuing Theory*, 2nd ed. Chapman and Hall, London, England, 1982, pp. 287–300.
13. F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. *Building Expert Systems*. Addison-Wesley Publishing Company, Inc., Reading, Mass., 1983.
14. A. G. R. Bullen, N. Hummon, T. Bryer, and R. Nekmat. EVIPAS: A Computer Model for the Optimal Design of a Vehicle-Actuated Traffic Signal. In *Transportation Research Record 1114*, TRB, National Research Council, Washington, D.C., 1987.
15. C. Zozaya-Gorostiza and C. Hendrickson. Expert System for Traffic Signal Setting Assistance. *Journal of Transportation Engineering*, Vol. 113, No. 3, March 1987.
16. N. H. Gartner. OPAC: A Demand-Responsive Strategy for Traffic

Signal Control. In *Transportation Research Record 906*, TRB, National Research Council, Washington, D.C., 1983.
17. *Evaluation of the Optimized Policies for Adaptive Control Strategy*. Report FHWA-RD-89-135. FHWA, U.S. Department of Transportation, 1989.
18. F. Lin, D. Cooke, and S. Vijayakumar. Use of Predicted Vehicle Arrival Information for Adaptive Signal Control—An Assessment. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987.
19. F. Lin, N. Wang, and S. Vijayakumar. Development of an Intelligent Adaptive Control Logic. Presented at the Engineering Foundation Conference on Management and Control of Urban Traffic, New Hampshire, June 1987.
20. C. E. Lee, T. W. Rioux, and C. R. Copeland. *The TEXAS Model for Intersection Traffic—Development*. Research Report 184-1. Center for Highway Research, The University of Texas, Austin, 1977.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Algorithm for Estimating Queue Lengths and Stop Delays at Signalized Intersections

HUEL-SHENG TSAY, JHY-FU KANG, AND CHIEN-HUA HSIAO

Queue length is a basic element of urban traffic control for advanced analysis or applications. An algorithm for estimating queue lengths and stop delays at signalized intersections has been developed. The algorithm can be used to predict the queue length and the number of queuing vehicles on each approach or block to reflect actual traffic conditions by every second or other assigned time interval depending on the requirement of each traffic control center. A simulation program based on this algorithm is developed to calculate these values by considering the status of real-time traffic lights, the number of queuing vehicles left, and the location of a designated vehicle detector. Field and video measurements were made at four lanes of three intersections in Taiwan to test the predictions of this time-dependent queuing model. The preliminary results of comparisons between observed and estimated queue lengths and stop delays are encouraging and interesting. This simulation program has been installed at five TRUSTS (Traffic Responsive and Uniform Surveillance Timing Systems) in Taiwan for controlling urban traffic effectively.

When the lights are red, queues build up as a result of turning movement into the arterial at the previous intersection before the appearance of green. These values include not only turning vehicles from the previous intersection but also the vehicles that do not pass through the arterial at the end of the last green time. The phenomena are quite obvious and should not be neglected at any signalized intersection during the entire day. Queue length on each approach or block is a basic and important element of urban traffic control for advanced analysis and applications. For example, the most important variable for solving the maximum progression bandwidth is to calculate the time needed to clear the average number of vehicles standing in the queue on each block under the analysis period, such as a 15-min or 1-hr traffic flow. In other words, the incoming through-band vehicles cannot cross the intersection unless all queues in front have cleared. Therefore, it is necessary to obtain a more accurate queue length estimation to reflect traffic conditions. Through these values, the operator understands the degree of traffic congestion on each block and evaluates the suitability of current signal timing plan. The values can be further applied to find the actual shortest or second shortest routes between any origin and destination, detect the incident, or develop the adaptive control strategy.

H.-S. Tsay, Taipei Rapid Transit Systems Company, 10 Chung Hsiao E. Rd., Section 5, Taipei, Taiwan, R.O.C. J.-F. Kang, National Cheng Kung University, Tainan, Taiwan, R.O.C. C.-H. Hsiao, Taiwan Signal Company, Taipei, Taiwan, R.O.C.

A new type of urban on-line traffic control system, TRUSTS (Traffic Responsive and Uniform Surveillance Timing System), has been successfully developed in Taiwan. The system involves several personal computers (PCs) that are connected by a NOVELL network. Each PC can be replaced immediately, without system breakdown, if it malfunctions or becomes functionally obsolete. It offers the user the choice of on-line timing plan generation, on-line timing table selection, or a time-of-day timing plan (*1*).

The TRUSTS wall map depicts city streets and administrative boundaries. Information on the wall map is provided by the wall map PC. The wall map capability includes displays of multiple-phase green, degree of congestion, flashing intersections, malfunctioning intersections, locations of detectors, and the 10 most congested approaches. It has several 5-cm-square polycarbonate boxes. Each box has nine lights showing the real-time traffic lights for each intersection with eight lights, including leading and lagging phases, and one-way streets. One flashing light in the center of the box represents the malfunction. The wall map PC can calculate queue length to show the degree of congestion for each approach or block. It is displayed in four colors for three indicators: occupancy, speed, and queue length. This display provides useful information to the user, and helps the user to select a suitable control strategy to cope with the current traffic condition. The relationship between the display colors and the three indicators used in TRUSTS are shown in Table 1. The occupancy and speed can be directly obtained from vehicle detectors. Queue length, however, has to be estimated from a formula or simulation program. The value used in Table 1 may vary from city to city. The large TRUSTS wall map, with its dynamic display of traffic lights and flow conditions at various threshold levels and its status display, fulfills area-wide needs. This type of wall map is very different from those used in other countries.

The PC graphic shows the real-time traffic signal system data on a color monitor. The monitor displays information on traffic lights, volume, speed, occupancy, and shortest routes. This information helps the user to understand the actual traffic conditions of a designated area or intersection. The graphics can be designed to focus on small areas, displaying increasing levels of detailed information. The monitor shows the shortest routes between two intersections, with or without considering the queue length on each approach. In addition, the graphics can provide up-to-date road and traffic information, such as the train schedule, the location and causes of a closed road,

TABLE 1 RELATIONSHIP BETWEEN DISPLAY COLORS AND THREE INDICATORS IN WALL MAP

Color	Occupancy(%)	Speed(km/hr)	Queue Length(m)
No display	≤ 10	≥ 40	≤ 20
Green	11-20	39-30	21-50
Yellow	21-30	29-15	51-100
Red	> 30	< 15	> 100

and traffic jams. The PC graphics give the user in-vehicle information through a combination of sensors and a communication system with the traffic control centers. This part, called the Advanced Driver Information System (ADIS), has been defined as one of four major intelligent vehicle-highway system (IVHS) areas in the United States.

This paper develops an algorithm for estimating queue lengths and stop delays at signalized intersections. Several variables related to this queuing model are discussed. In order to perform the sensitivity analysis and practical applications, a simulation program based on the algorithm considering the current signal timing plan is also developed. The output prints estimated queue lengths and the number of queuing vehicles by lane at the end of each 15-sec counting interval. Field and video measurements of queue lengths and stop delays test the accuracy of the predictions of this time-dependent queuing model. Field observations are made at four lanes of three intersections for a number of counting intervals at each site.

PREVIOUS FORMULAS FOR ESTIMATING QUEUE LENGTHS AND DELAYS

Catling (*2*) developed formulas to estimate mean queue lengths and delays under both undersaturated and oversaturated conditions for an interval with a stationary mean arrival rate and starting with zero queuing length. Later, he extended the work to cover variable demand levels and non-zero initial lengths. Branston (*3*) investigated the formulas for observations in three traffic peaks at sites in London and concluded that reasonable results could be obtained. The estimated mean queue length is valid only at discrete intervals of time, namely, the beginning of successive red periods. Mean queue lengths at other times during a cycle must be calculated from these values with a knowledge of the arrival rate and saturation flow. Shawaly et al. (*4*) concluded that the arrival flow delays have a significant effect on the resulting queuing length and delays even though the departure pattern remains unchanged. Kimber and Daly (*5*) found that queue length and delay predictions are particularly difficult when demand reaches capacity. Steady-state approximation no longer holds, and time-dependent stochastic methods must use approximation to cope with what is in reality a complex time development of the queuing states involving difficult sampling problems. The observed data indicated a big variation about the estimated mean profile of queues and delay.

TRANSYT is one of the most popular computer programs for optimizing the signal timings of a network with coordinated intersections. Because TRANSYT simulates traffic for a single cycle, the calculated queue length represents only that which would occur because of traffic arriving during that cycle (*6*). Queues do not build over time. The queuing model assumes that all vehicles travel the full length of the link before joining a stationary queue, thus forming vertical queues at the stop line. Although this model is not realistic, it deals adequately with the delay imposed on traffic at intersections. That is, TRANSYT does not consider queues spatially, nor does it limit the length of queue at any stage relative to the length of the link. The total number of departures using the saturation flow rate equals the number of vehicles in the vertical queue, including the vehicles joining after the start of green. The TRANSYT queue model is shown in Figure 1. This type of estimated queue length at the stop line usually gives rise to an overestimate of queue length. The estimation becomes more serious with the increase in queue length (*7*).

TIME-DEPENDENT QUEUING MODEL

The arrival pattern of the algorithm is established so that the traffic joins the back of the queue with the consideration of actual vehicle lengths. It can avoid an overqueuing situation due to the block-length constraint. After the beginning of green, a vehicle in the queue remains stationary until it is reached by the start wave. Before the queue is cleared, vehicles in the queue discharge at a speed associated with the saturation flow. The queue length will be continually added up from the incoming traffic until the start wave reaches the back of the waiting queue. At this moment, the queue length immediately becomes zero. Although conceptually it may not be reasonable that the queue length becomes zero after the last vehicle in the queue is reached by the start wave, in field tests the queue length estimation on each approach or block still represents the actual value in most cases. After the total queue disappears, vehicles move off at the same speed as the arriving traffic before the traffic light turns to yellow and all-red. It is hoped that queues can be cleared from the most distant detectors during green phases. If traffic flow increases still further, queue lengths may extend beyond the most remote detectors and the computation of the queue length will be terminated or need more assumptions.

The queue length defined here is the number of vehicles from the stop line of a signalized intersection to the back of the last stationary vehicle in the queue, irrespective of whether vehicles other than the last vehicle are moving. Queuing vehicles, however, represent the number of vehicles that are actually stopped in order to estimate the stop delay. These two major items, queue lengths and queuing vehicles, are considered separately in this time-dependent queuing model.

The dynamics of this time-dependent queuing model are shown in Figure 2. The queue length in this figure increases from one to two units of vehicle length at the end of 5 sec

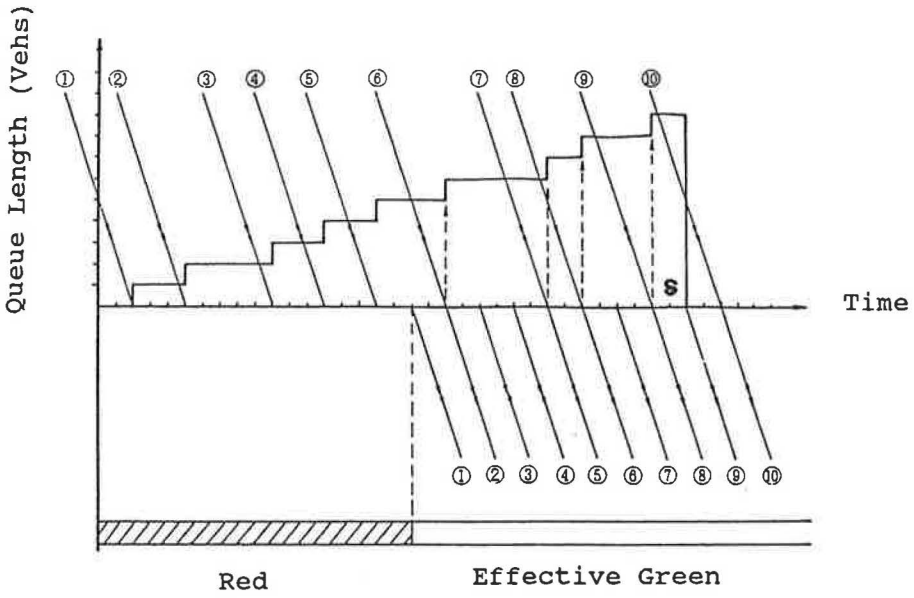


FIGURE 1 Queuing model for TRANSYT.

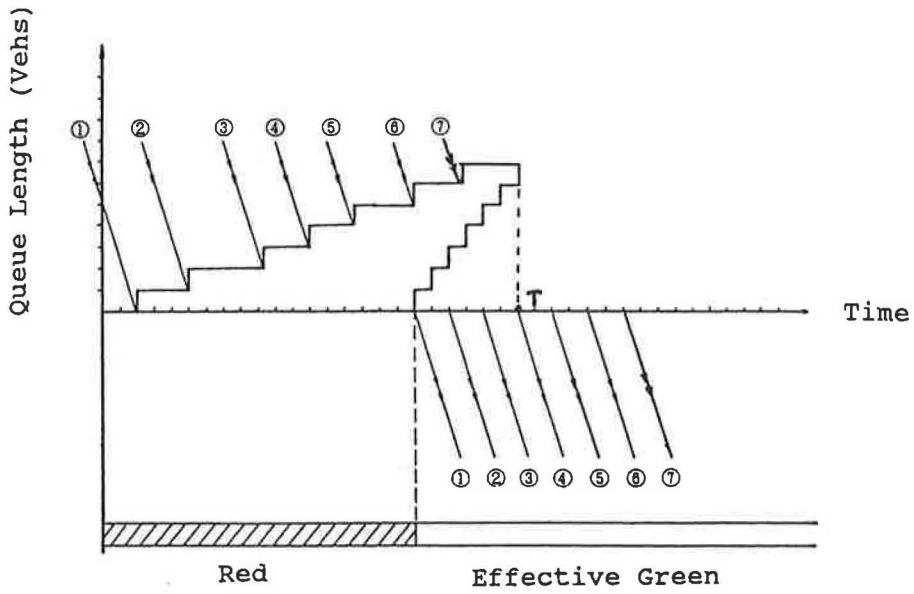


FIGURE 2 Dynamics of time-dependent queuing model.

(one step equals 1 sec) after the beginning of the red. Then the queue length becomes three units at the end of 9 sec and so on, until it reaches the maximum length of six units at 18 sec in this example. As the light turns green, the queue length still increases to seven units, and this last stationary vehicle is reached by the start wave at 24 sec (symbol T in the figure). The estimated queue length defines zero after that time. In the meantime, the seven vehicles in the queue dissipate with a saturation flow rate of 2 sec per vehicle and clear at the end of 30 sec.

In a real situation, the number of vehicles queuing varies continually from cycle to cycle. This time-dependent queuing model provides a measure of the estimated queue length and queuing vehicles at the end of every second or a longer defined interval. It not only shows the number of vehicles stationary in the queue but also indicates the actual waiting distances. At a signalized intersection, traffic delays usually result in queues. Vehicles in such a queue are likely to remain stationary and can be used to calculate stop delays through the number of queuing vehicles instead of the estimated queue length. Therefore, the actual stop delay for each queuing vehicle is simply determined by subtracting the stop time after its arrival from the next moving time. These stop delays are then summed up over each time period and averaged for the total number of vehicles that arrived during the same time interval. The delay equation in the 1985 *Highway Capacity Manual* (HCM) (*8*) was developed from both uniform delay and random delay through deterministic queuing models. With

real-time traffic inputs, the average queue size cannot be estimated through this delay equation second by second or for a 15-sec time interval. Therefore, the 1985 HCM delay equation considers the number of queuing vehicles and stop delay through the macroscopic concept accompanying a complete cycle length. The proposed time-dependent queuing model, however, considers the microscopic movement in an assigned short period.

Queue length represents the result of a complex function of vehicle moving speed, length and type of arriving vehicles, location of vehicle detector, start wave speed, geometric shape, and lane-change behavior. Some variables are discussed here. Vehicles, after passing the vehicle detector, usually maintain a constant speed for a few seconds, then decelerate gradually to stop and join the last queue vehicle. This is a nonlinear procedure and is difficult to trace accurately because there are not enough detectors to provide the needed information along each block. One simple way to solve this problem may be to use a lower constant speed. The variable for length and type of each incoming vehicle can be directly estimated from the vehicle detector and classifier. It is easy to use this value in the time-dependent model through the TRUSTS equipment. The variation in the start wave speed during the effective green stage has a varying degree of importance, depending on the number of vehicles queuing. If a higher start wave speed is assumed, then the time needed to clear the estimated queue length will be decreased, and vice versa. A better way to obtain this value may be observing the videotape of field tests based on different locations of vehicle detectors. For the lane-change behavior, the model becomes very complex if each vehicle is traced, and advanced electronic equipment is required to record the movement of all vehicles. Therefore, for simplicity, the lane-change variable is not considered at this stage.

In order to obtain the estimated queue lengths and stop delays from the real-time traffic lights, a simulation program, written in C language, was based on the above concept. The flow chart of this program is shown in Figure 3. The program first checks the status of the signal timing plan. If the light is red, the program continues to calculate the accumulated queuing vehicles and stop delays by every second or other assigned time interval. After the light turns green, the program automatically investigates the queue length whether or not it has been cleared. If the answer is negative, the program not only accumulates the number of vehicles in the queue but also traces the reaching location of start wave in order to estimate queue length, queuing vehicles, and stop delay. If the answer is positive, the queue length remains zero until the next red phase. The program continues to compute the queue length and the number of queuing vehicles at the end of each assigned time interval and prints the time of queue length cleared.

This program allows the user to perform sensitivity analyses by setting different values of variables. The input items include the average approaching speed, distance from the detector to stop line, current signal timing plan, and time of each vehicle passing through the detector. Figure 4 shows the portion output of estimated queue lengths and queuing vehicles for a 15-sec interval and average stop delay throughout test periods. This example assumes that the average vehicle length has a magnitude of 6 m with no lane-change behavior. The start wave occurs in the beginning of the green phase

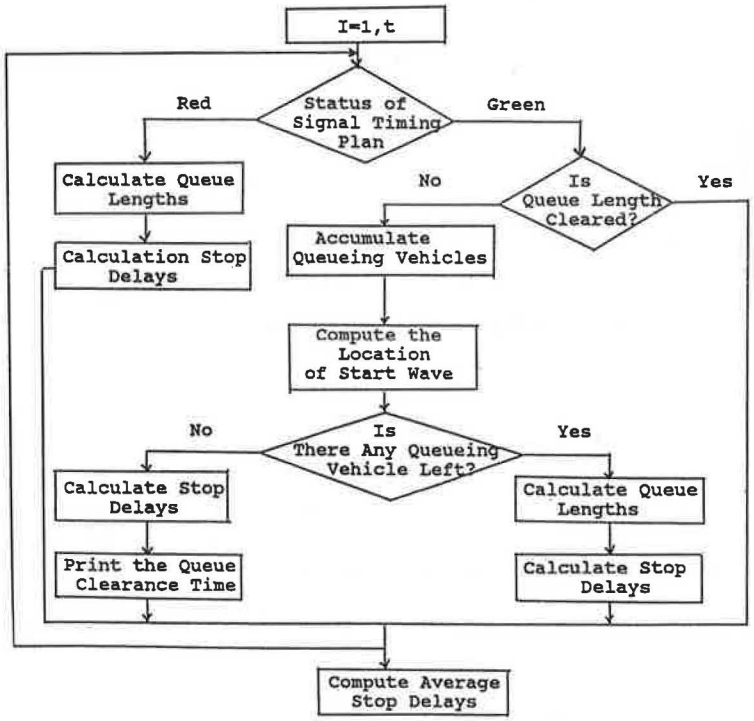


FIGURE 3 Flow chart of simulation program based on time-dependent queuing model.

Time	Estimated Queue Lengths		The Location of Start Wave		Queueing Vehicles
35:00	0	-	0	=	0
35:15	0	-	0	=	0
35:30	0	-	0	=	0
35:45	4	-	0	=	4
36:00	9	-	0	=	9
36:15	14	-	0	=	14
36:30	15	-	9	=	6
Queue lengths were cleared at 36:36					
36:45	0	-	0	=	0
37:00	0	-	0	=	0
37:15	0	-	0	=	0
37:30	0	-	0	=	0
37:45	2	-	0	=	2
38:00	8	-	0	=	8
38:15	10	-	0	=	10
38:30	12	-	9	=	3
Queue lengths were cleared at 38:33					
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48:45	0	-	0	=	0
49:00	0	-	0	=	0
49:15	0	-	0	=	0
49:30	0	-	0	=	0
49:45	6	-	0	=	6
50:00	12	-	0	=	12
50:15	15	-	0	=	15
50:30	17	-	9	=	8
Queue lengths were cleared at 50:38					
50:45	0	-	0	=	0
Total vehicles: 192					
Average stop delay per vehicle: 18.97 sec/veh					

FIGURE 4 Portion output of estimated queue lengths, queuing vehicles, and average stop delays.

with a speed of 1.1 veh/sec based on various observations of videotapes. The assumed values of these variables may vary from city to city.

COMPARISONS OF OBSERVED AND ESTIMATED VALUES

Queue length estimated by the proposed comprehensive time-dependent queuing model is compared with queue lengths actually observed during six different time periods. To obtain more accurate and reliable field data, the minimum time of observation was 15 min for each counting period. Measurements were made at four lanes in three intersections in Taichung City, Taiwan. Lanes A and B were located in the same approach; Lanes C and D, however, represented two different approaches. The selected lanes allowed manual and video measurements of the physical length of road occupied by the queue. The field measurement included the observed queue length and stop delay at the end of a 15-sec interval during total counting periods. The developed simulation program and videotapes allow the user to calculate and investigate the queue length second by second or for a longer time interval depending on the requirement of a local traffic control center. The observed stop delays were obtained through a point sample procedure that was found to be the most practical method for measuring the intersection delay in the field. A video system with two cameras collected the required data. The first camera focused on the vehicles passing through the vehicle detector, and the second camera recorded the entire process of vehicle movements from the upstream to downstream intersections with a time sequence on the screen. The two video cameras and observers' watches were synchronized to a common time base before the start of observations.

The preliminary comparisons of observed and estimated queue lengths by lane during six counting periods are shown in Table 2. This table includes the mean queue length and its standard deviation, the distance from the vehicle detector to stop line, the sample size of each counting period, and the percentage of accurate estimation for four different lanes at various periods. Some discrepancies were found between observed and estimated results. From this table, the estimated values from Lanes A and B were more accurate than those obtained from Lanes C and D by over 20 percent. This discrepancy was probably due to the fact that the travel distance to stop line for Lane D was 100 m longer than for Lane B after vehicles passed through the vehicle detector. This finding implies that vehicles may overtake one another and distort the sequence of recorded queue vehicles during the arriving process from the location of existing vehicle detector to the last queue vehicle. Therefore, the estimated results are sensitive to this value, which is a strongly site-dependent parameter.

The relationship between estimated and observed values during test periods for Lanes B and D are shown in Figures

TABLE 2 COMPARISONS OF OBSERVED AND ESTIMATED QUEUE LENGTHS FOR FOUR LANES

Lanes	Counting periods	Distance from Detector to Stop-Line (Meters)	Estimated Queue Length (Vehs)	Estimated Queue Length Standard Deviation (Vehs)	Observed Queue Length (Vehs)	Observed Queue Length Standard Deviation (Vehs)	Total Sample	Same Sample	Percent (%)
A	1	65	0.32	0.69	0.49	0.98	96	80	83.3%
	2	65	0.67	1.11	0.81	1.18	135	105	77.8%
B	1	65	0.60	1.07	0.70	1.20	96	75	78.1%
	2	65	0.98	1.40	0.83	1.15	135	99	73.3%
C	3	100	2.60	3.11	2.47	2.86	176	92	52.3%
	4	100	1.10	1.48	1.34	1.49	62	32	51.6%
D	5	165	3.70	4.66	3.38	4.48	64	38	59.4%
	6	165	4.76	5.84	3.09	5.10	89	49	55.1%

5 and 6. The irregular trend of actual queue length during test periods is shown. Comparisons of estimated and observed values by considering one vehicle difference are shown in Table 3. This table shows that at least 70 percent of estimation from the simulation program is reasonable. That is, the estimated queue lengths can represent actual values in most cases for a 15-sec interval.

Table 4 shows the differences between observed and estimated stop delays. The estimated stop delays are directly obtained from queuing vehicles instead of the estimated queue length. The observed stop delays are collected through the point sample method. The raw value for stopped time is multiplied by 0.92 to represent the observed stop delay. This multiplier factor applied to the raw field data achieves a better estimate of the true value. The coefficient 0.92 was recommended by Homburger and Kell (*9*) and can be varied from different locations if sufficient field data are available. The differences shown in Table 4 range from 3 to 28 percent. From these values, it can be concluded that the estimated stop delay, as well as the queue length calculated through this proposed time-dependent queuing model, apparently represents the actual value.

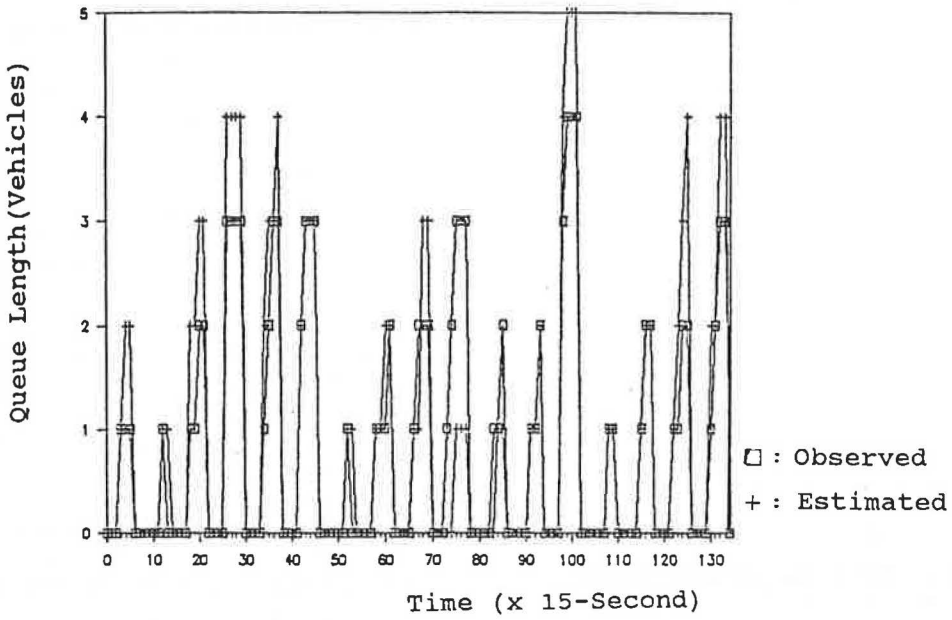


FIGURE 5 Relationship between estimated and observed values during test periods for Lane B.

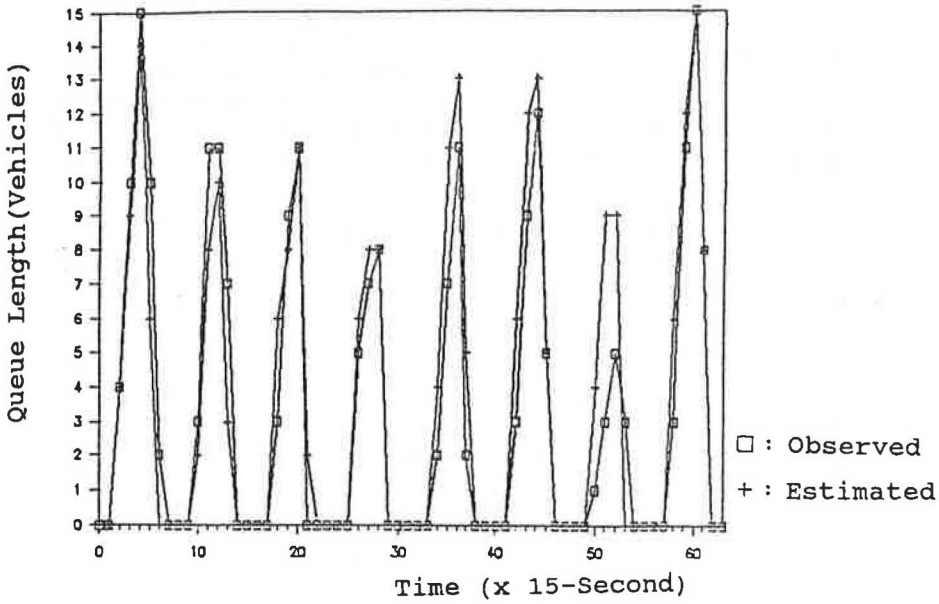


FIGURE 6 Relationship between estimated and observed values during test periods for Lane D.

TABLE 3 COMPARISONS OF ESTIMATED AND OBSERVED QUEUE LENGTHS CONSIDERING ONE-VEHICLE DIFFERENCE

Lanes	Counting Periods	Less One Vehicle (%)	Same (%)	One Vehicle More (%)	Total (%)
A	1	14	83	1	98
	2	15	78	5	98
B	1	8	78	5	91
	2	3	73	20	96
C	3	13	52	12	77
	4	19	52	8	79
D	5	8	59	6	73
	6	9	55	6	70

TABLE 4 COMPARISONS OF ESTIMATED AND OBSERVED AVERAGE STOP DELAYS FOR FOUR LANES WITH DIFFERENT COUNTING PERIODS

Lanes	Counting Periods	Estimated (sec/veh)	Observed (sec/veh)	Differences (%)
A	1	2.46	3.08	-13.6%
	2	4.54	5.25	-13.5%
B	1	3.89	4.02	- 3.2%
	2	6.07	4.74	28.1%
C	3	22.51	28.19	-20.1%
	4	10.72	13.84	-22.5%
D	5	18.97	16.96	11.9%
	6	20.49	19.82	3.4%

CONCLUSIONS

An algorithm for estimating queue lengths and stop delays at signalized intersections has been developed. Based on the information of incoming traffic and the current signal timing plan, the algorithm can be used to predict queue lengths and the number of queuing vehicles for every second or other time interval assigned according to the needs of a traffic control center. Preliminary comparisons of observed and estimated queue lengths and stop delays are encouraging. The proposed time-dependent queuing model considers several related variables but some are simplified in the process of estimation. More research on certain variables, such as lane-change logic and approaching speed, is needed.

Queue length on each approach or block is a basic element of urban traffic control for advanced analysis or applications. This information is used in the wall map and computer graphics of TRUSTS to represent the degree of congestion for various blocks. Through this estimated queue length, TRUSTS can provide in-vehicle information to drivers, such as the actual shortest route or second shortest route between any origin and destination. The next step for this study is to add the estimated queue length into the performance index function of each approach in order to develop a new type of adaptive control that is different from the concept used in OPAC (*10*) or SCOOT (*11*). Some interesting results have been obtained and will be presented later.

Finally, a simulation program based on this algorithm has been installed at five TRUSTS in Taiwan. The display of degree of congestion on different blocks shows the operator the actual traffic conditions of the entire area. It allows the user to quickly spot the problem through the large wall map and computer graphics. The operator can change the signal timing plan or immediately alert the patrolling police to the congested areas from the control center. So far, TRUSTS controls urban traffic flow effectively in Taiwan.

REFERENCES

1. H. S. Tsay. A Microcomputer-Based On-line Traffic Control System. *ITE Journal*, Dec. 1989, pp. 29–36.
2. I. Catling. A Time-Dependent Approach to Junction Delays. *Traffic Engineering & Control*, Nov. 1977.
3. D. Branston. A Comparison of Observed and Estimated Queue Lengths at Oversaturated Traffic Signals. *Traffic Engineering & Control*, July 1978, pp. 322–327.
4. E. A. A. Shawaly, R. Ashworth, and C. J. D. Laurence. A Comparison of Observed, Estimated, and Simulated Queue Lengths and Delays at Oversaturated Signalized Junctions. *Traffic Engineering & Control*, Dec. 1988, pp. 637–643.
5. R. M. Kimber and P. N. Daly. Time-Dependent Queuing at Road Junctions: Observation and Prediction. *Transportation Research*, Vol. 20B, No. 3, 1986, pp. 187–203.
6. *TRANSYT-7F User's Manual*, Release 5.0. FHWA, U.S. Department of Transportation, 1987.
7. M. C. Bell. A Queuing Model and Performance Indicator for TRANSYT 7. *Traffic Engineering & Control*, June 1981, pp. 349–354.
8. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
9. W. S. Homburger and J. H. Kell. *Fundamentals of Traffic Engineering*, 12th ed. Institute of Transportation Studies, University of California, Berkeley, 1988.
10. Evaluation of the Optimized Policies for Adaptive Control Strategy. FHWA, U.S. Department of Transportation, 1989.
11. P. B. Hunt, D. I. Robertson, R. I. Bretherton, and R. I. Winton. *SCOOT—A Traffic Responsive Method of Coordinating Signals*. TRRL Report 1104, U.K. Transport and Road Research Laboratory, 1981.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

True Distributed Processing in Modular Traffic Signal Systems—San Antonio Downtown System

RICHARD W. DENNEY, JR., AND MICHAEL J. CHASE

In San Antonio, limited funds precluded the traditional approach to a consultant-designed Urban Traffic Control System (UTCS) for traffic signal control in the downtown area. By applying principles now common in other parts of the computer industry, San Antonio engineers were able to formulate an alternative to the traditional approach that not only provides very substantial improvements in cost-effectiveness, maintainability, and reliability, but also provides the end user with complete access to the inner workings of the system. Emphasis on conforming to computer-industry standards of system design and commitment to open hardware and software architecture allowed full portability of software. Highly distributed processing greatly reduced the communications overhead while improving operation compared with other large-scale systems currently in operation in the United States. The San Antonio system has full-featured capabilities, including planned traffic-responsive operation, and uses one-tenth of the usual communications overhead without using a machine larger than a microcomputer. A comparison is made with a recent UTCS project.

The development of traffic signal systems has moved in two directions since the introduction of the Urban Traffic Control System (UTCS) in the early 1970s. The UTCS approach (*1*) is followed in most large system applications (more than 32 intersections); because of manufacturer specificity, approaches with more complete distribution of processing have evolved in order to provide a more cost-effective product for smaller systems.

Despite the obsolescence of the original UTCS approach to processing and communications, it is a boon to public-sector traffic engineers who must rely on consultants for design and who need a product allowing open competition for local controllers. Public employees could go to their city council or commission and present the design of the system as a consultant project (therefore not requiring low-bid purchasing in most areas) without being locked in to a specific manufacturer for future maintenance supplies. Unfortunately, UTCS is highly oriented to central processing and, even in its more recent manifestations, requires a large central computer and a dense communications network. Later implementations reduce the need for central computing and communications, but require customized programming, usually by the consultant designing the system. Therefore many of the significant modules of the UTCS software end up looking not at all like the UTCS original, thus undermining the purpose of public-domain software.

R. W. Denney, Jr., City of San Antonio, P.O. Box 839966, San Antonio, Tex., 78283-3966. M. J. Chase, Boulder Software Group, P.O. Box 14200, Boulder, Colo. 80308-4200.

More up-to-date systems are available from a number of manufacturers. Many of these products boast excellent performance, but all require a commitment to a particular manufacturer for maintenance and expansion materials. Also, many of the manufacturers are justifiably reluctant to expose their communications methods and protocols, preventing a city from improving or enhancing the system without going back to the manufacturer. Confidentiality of the protocols also makes it impossible for cities willing to experiment with different control technologies to get their ideas on the street without significant cost. Many signal system experts in the public sector are presented with a choice between the high cost of trying new methods with the manufacturer making the modification, or the even higher cost of developing an in-house system so that the public agency will own the program code. The latter choice is available only to very large organizations, for example, the Texas Department of Highways and Public Transportation's development of the Flexible Advanced Computer Traffic Signal System (*2*).

Most of the manufacturer-specific systems are designed for smaller applications. They are often referred to as closed-loop systems because they "close the loop" with the traffic engineer by providing remote access to the system. Larger cities find it easier to purchase these systems because they can still purchase other systems from other vendors for other parts of the city.

A common thread among many traffic engineers is the consideration of these systems primarily as traffic signal systems rather than as computer systems. A natural result of this thinking is that the traffic control features are mixed in with computer system features, and often traffic control considerations overshadow computer system necessities. Also, traffic engineers find themselves in the uncomfortable position of trying to toughen up (supposedly) competitive specifications in order to get the features they need.

This paper presents an example of a different approach to the conception of a large system. In San Antonio's case, this system was brought about by necessity. The bond program approved for the downtown signal system did not provide enough funds for a UTCS approach, nor were there adequate funds to pay for an outside design. In 1984, voters in San Antonio approved $1.5 million in general obligation bonds for a new traffic signal system in the downtown area. Because of lack of engineering resources, the project was not begun until August 1987. The central business district in San Antonio includes about 150 traffic signals, which were part of a PR system originally installed in 1957. San Antonio is the ninth

largest city in the United States, with a population of over 1 million and approximately 1,150 traffic signals.

The different approach to the San Antonio system can be summed in one phrase: Design a computer system as a computer system, and then get the software it needs to run traffic signals. We believe that this thinking has resulted in the most cost-effective system in the United States, comparing favorably with recent independent developments in Taiwan and elsewhere, without sacrificing any capabilities that are important to the task of traffic control, especially access to the operation of the system by the owner to allow experimentation with various control strategies.

This paper first describes the general design of the San Antonio system, and then compares the resulting "computer-system" features to a recent implementation of UTCS. To emphasize the treatment of signal systems as computer systems, we make frequent comparisons to a familiar part of the computer industry—the standard desktop microcomputer.

DISTRIBUTED PROCESSING IN SAN ANTONIO

When setting out to purchase a microcomputer, the first-time buyer usually thinks only generally of the intended use, and concentrates primarily on processing power, user access, and other computing features. Most first-time users have too little experience with the available software packages to make an informed decision about which hardware to use. Generally, people using computers for artistic purposes, such as visual or musical art, purchase a Macintosh, and those needing more business-oriented applications use an IBM-type system. The distinction between the suitability of these different environments for various purposes has, however, blurred significantly over the years, and now most any kind of work can be done on either machine. The point is that many people select a hardware platform based more on suppositions of intended use rather than on actual experience with the software, and then create a demand for the software to run, say, Macintosh-like applications on their IBM, and vice versa. Very few first-time buyers can predict the applications to which they will put their computers, and most applications are discovered after the machine is available.

Most people, therefore, in some way violate the traditional approach to buying computers: Define specific uses, and then select the software and hardware to provide those capabilities. Rather, they buy the most powerful hardware they can afford, and then obtain software as the need arises. This approach may seem less "systematic," but it makes more sense in the long run. Using traditional thinking, users run the risk of having machines that cannot grow with their improved understanding, and they limit themselves to a premature concept. For example, many owners of off-brand microcomputers now look longingly at the IBM machines their competitors are using. The off-brand system was the most effective provider of their original concept, but now cannot do the other things their users now realize they need. In many cases, the off-brand computer costs as much or more than the more widely used IBM-based counterpart.

Purchasers of traffic signal systems have tended to make the mistake that most microcomputer purchasers have avoided (even if by accident). No specifications are more tightly written in the traffic engineering industry than those written to purchase signal systems. However, little detailed thinking is done about the computer hardware itself.

In San Antonio, we first committed to the hardware and then wrote a specification for software. This approach seems simplistic, and we emphasize that detailed thinking went into the hardware decision. When considering the hardware, we established a basic application of distributed processing by determining the level at which each function in the system would be performed. The basic rule of distributed processing is to place the processing power as close as possible to the processing need. The closer it is, the simpler and more economical the communications task becomes. In San Antonio, this approach led us to the following conclusions:

- All local intersection timing should be done by the local controller. Cycle length, offsets, splits, and even time-of-day and pattern scheduling should be handled at the point where it is used. As any operator of a time-based coordinated system knows, most controllers on the market already have these capabilities.
- All decisions made at the control-group level should be made by the zone master. Most systems on the market today use zone masters to control groups of local intersections. If local controllers can make all decisions about local timing, then the zone master must only handle traffic-responsive pattern selection and broker communications between the central computer and the local intersection. From a computer system standpoint, the latter feature is critical. If the central computer is to talk to individual intersections directly, as in the traditional UTCS system, the communications task is enormous, requiring large-scale multiplexing to allow messages for hundreds of locals to come into relatively few communications channels, or requiring a very expensive central computer to allow a channel for each local. Large systems no longer require the latter method.

From a traffic control perspective, the use of control groups and zone masters may, at first, seem to get in the way of the more subtle requirements of traffic control systems. For example, one significant feature of the UTCS software is the ability to redefine control-group boundaries by time of day. This feature is useful, though not implemented often, but it is an outgrowth of the basic UTCS operation. The original UTCS system used the central computer to advance the interval of each controller on the system, and all cycle length and split decisions were directed at the central point. The software was designed to break up the system into subsystems; within each subsystem the pattern selected would be the same. In a system where all signal timing is done at the local level, the zone master does not even have to know what the cycle length is, and control-group boundaries can strictly be a function of the way the patterns are designed in the local controller.

- Supervision of zone masters and provision of user access should be done at the central computer. The central computer has the capability to store information about the whole system, and should maintain a copy of the timing parameters, which are contained in each local and zone master to minimize the need for communications to provide the user with day-

to-day information. In San Antonio, access to the system is needed from different locations, including the maintenance facility, the control center, the traffic engineering office, and the homes of staff members required to monitor the system after hours (to troubleshoot a problem, for instance). This need requires a multiuser access capability, such as that provided by minicomputers and microcomputers running a multiuser operating system, such as UNIX.

- Graphics displays and the user interface should be run on the user terminal. An operating system such as UNIX provides many of the capabilities for graphic display and user access to the system data base; however the substantial recurring graphics and screen display information would move very slowly over a modem or other long-distance serial link, and competition with other users would decrease some response. Following the rule of distributed processing, one realizes that the *pictures* are always the same, only the *data* presented in those pictures need be communicated from the central computer. By using terminals that are standard microcomputers, the user interface can be programmed on the terminal, thus freeing the central computer of the task of making and communicating pictures. Thus, remote terminals accessing the system via dial-up telephone links will be able to enjoy the same graphics and user interface as terminals with a direct link, with no significant loss in response time.

San Antonio selected the Type 170 controller for use as the local controller. The Type 170 is a microcomputer (albeit not a very powerful one), and its open architecture allows separate purchase of the hardware and software (*3*). The development of the software for the Type 170 controllers was included along with the development of the central computer software, and the hardware was purchased with the annual supply contract for controllers. This allowed us to specify the desired features from the software developer in a negotiated consultant contract (which is not based on low bid) and purchase the hardware separately and competitively. This arrangement would not have been possible with controllers following the National Electrical Manufacturers Association (NEMA) specification (*4*), where hardware and software are linked in a closed architecture. The result is that software features may be sacrificed in order to allow a competitive hardware bid.

The advantage of separately purchased software cannot be overemphasized. While in the NEMA sphere, features must be common to several manufacturers before a competitive specification can be formulated. The ability to purchase software separately allows features that are not even available to be specified, because the features are in the software and are not competitively bid. Therefore the tools of innovation are back in the hands of the system buyers, rather than hardware manufacturers who at times must have divergent concerns. Again, this is analogous to the personal computer industry. When IBM introduced the personal computer (PC), it opened the books on its architecture, bus, and peripheral design. Consequently, a foundation now exists for providing applications not dreamed of by the original designers. Innovative software, however, is only developed in a competitive environment where software developers are working from a common and open hardware standard. Such a standard is provided by the Type 170 specification, but not by the NEMA specification.

Based on these design guidelines, which were detailed in a 13-page preliminary design report, the city of San Antonion awarded a $163,000 consultant contract to develop the software and obtain the computer equipment.

SYSTEM DESCRIPTION

The multilevel hierarchy of the system architecture allows independent development of the various components of the system. The mechanism that directs the interaction of these various processes is called a kernel; in this case the kernel is part of the central control program (CCP) software running on the central computer under UNIX. The other components of the software communicate with each other by sending a series of messages to the kernel to be passed on to other processes. By defining these messages first, the basic function of each component of the software can be defined before the program is actually written, and all program modules are written with a firm understanding of how they fit into the whole. This system description begins with a discussion of the kernel and the central computer, and then describes some of the features of the local controllers, zone masters, and terminals.

Kernel and Central Hardware

The CCP kernel is written in the C programming language under the UNIX operating system, following the AT&T System V Interface Definition and ANSI X3.159 programming conventions. As such, it will run without modification when compiled on any computer running AT&T standard System V UNIX. Therefore, the size of the central computer then becomes strictly a question of needed hardware horsepower. Because the central computer is not required to direct individual intersection operation or generate presentation graphics, the task of simply brokering communications of the various elements of the system is not large, even on a large system. Processing power is not as critical as communications capability. Using intelligent, multiple serial port add-on boards made by one of several manufacturers, a standard desktop PC can have up to 24 serial ports. Each serial port is used to talk to a zone master, a terminal, or a modem. In San Antonio's design (see Figure 1), two directly connected terminals use two of these ports. Other users access the system via a dial-up modem, which uses another two ports (to allow two dial-up users at one time). The remaining 20 ports in a PC environment can be used to talk with 20 zone masters. As shown later, each zone master can operate up to 32 locals, allowing a theoretical maximum of 640 intersections on a PC-based system. In actual practice, running the system at capacity would inhibit the flexibility to configure some zone masters with less than 32 intersections, which would prohibit future in-filling.

The advantage of using C and UNIX becomes apparent when the system outgrows the 24 serial ports allowed on standard microcomputers. The central computer can then be replaced with a small and inexpensive minicomputer, which can easily accommodate up to 64 serial connections. No modifi-

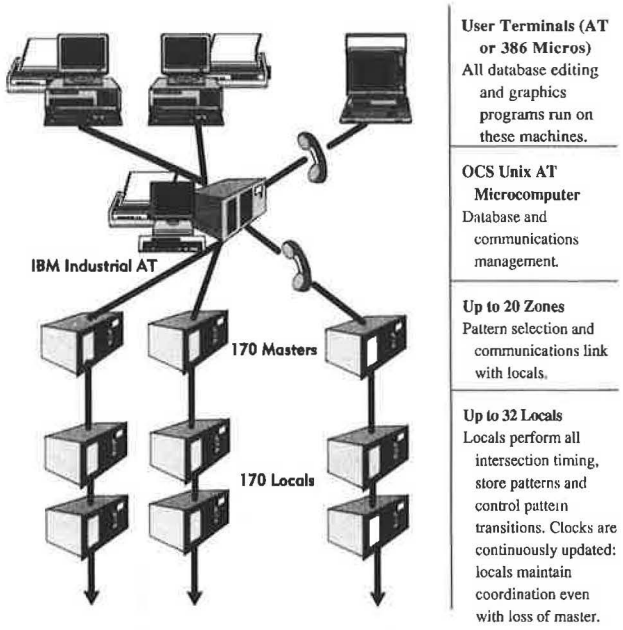


FIGURE 1 Organization of San Antonio CBD signal system.

cation of the software will be required, only recompiling the C programs on the new machine (C compilers are readily available in all UNIX implementations). Because UNIX provides complete access to communications hardware in the machine without machine-dependent instructions, the software is completely portable across all machines running UNIX.

The downtown system in San Antonio contains only 150 intersections; therefore, the PC environment was chosen. The hardware itself is an IBM Industrial AT, which is a rack-mounted version of the standard desktop AT. The rack-mounted case offers better control and filtration of cooling air, providing a cleaner environment for the hard disk. Otherwise, the components of the Industrial AT are no different from the components in a standard desktop PC-AT. The AT included a 30-megabyte hard disk, monochrome monitor, 4 megabytes of random access memory, and three plug-in cards with eight serial ports each. Also purchased was a 750-watt uninterruptable power supply, which also includes line conditioning and filtration. All these components can be easily and cheaply bought in any computer store, and no component of the system costs more than a few hundred dollars.

Because UNIX is a multiuser, multitasking operating system, many programs can run simultaneously. Each program running under UNIX is called a process. A process is invoked for each serial port, whether it is a terminal or a zone master; in addition, processes are run to maintain information about the current status of the system and to provide real-time supervision of the zone masters.

All of the traditional information reporting the status and integrity of the computer system is provided as in any large-scale system, including alarm status of zone masters and locals, on-line status, local controller status (current plan, conflict flash, etc.), and so on. Also added to these features are the standard features found on all UNIX systems, such as multiple-level security, electronic mail, user maintenance, and non-CCP programs.

The final user of the system is encouraged to make modifications to enhance the system; the user can write programs for independent processes that communicate with the CCP via the messages originally defined. For example, the San Antonio system specification requires the ability for the city staff to experiment with different means of coordinating the effort of adjacent traffic-responsive zone masters. Several schemes for coordinating control groups have been formulated, but few have been experimented with because implementing the proposals traditionally requires costly modifications to proprietary software. The result is that most systems do not grow with the improved understanding of their operators, as described previously. The city of San Antonio explicitly avoided this drawback by requiring an open architecture at the outset. City staff will be able to write a small program implementing a proposed algorithm for subsystem coordination by merely asking for the appropriate messages from the CCP kernel, making the necessary calculations, and sending the appropriate messages back to the kernel for subsequent direction to the zone masters. This independent program can be developed in any computer language running under UNIX (e.g., C, Pascal, FORTRAN). This example illustrates the need for providing mechanisms for future enhancements without necessarily knowing what those enhancements might be. Indeed, the algorithm used in this example is not yet formulated, and no research has been undertaken to shed light on how it might work.

The ability of the end user to add or modify control algorithms is one of the fundamentally important benefits of standard programming practices and open software architecture. The following scenario illustrates how this would work using the above example of traffic-responsive zone master control. The user would write a program called, say, Zone_Control. Zone_Control would be a continuously running process under UNIX. At the beginning of every control period, say, every 5 min or so, Zone_Control would send a message to the kernel to get the choice made by Zone Master 6. If Zone Master 6 traffic-responsively selected Plan 5, for instance, Zone_Control would discover this by sending the message Get_The_Selected_Plan(Master_6). The kernel is programmed to respond to such a message by sending a message to the process communicating with the zone master, which then queries the master and returns the result back to the kernel. The kernel then sends the return message Selected_Plan(Master_6, Plan_5) back to Zone_Control. Zone_Control then makes the decision to confirm or deny this selection by looking at the adjacent zones, and perhaps specific detector data (which would be obtained by sending a message to the kernel, etc.). At the conclusion of this algorithm, Zone_Control would send the message Run_This_Plan(Master_6, Plan_4) if the algorithm determined that Plan 4 would better suit the situation. Zone_Control could be easily programmed in Pascal, for example, and would use the standard UNIX techniques for talking with other processes. The programmer of Zone_Control would not have to know anything about the mechanics of communicating with signal controllers on the street, nor about how the kernel works. Only the messages (Get_The_Selected_Plan) and the system constants (Master_6) must be known.

As the technology of traffic control proceeds, these programming techniques will become more important. A gap already exists between computer programmers and most traffic engineers; a message-based control system allows the engineer to reasonably map out the logic of traffic control features without understanding the details of programming the communications and data base protocals. Many practicing engineers will be able to go even farther by actually writing the simple programs involved.

Zone Masters

As previously indicated, the zone masters will control up to 32 intersections. Because the locals already maintain time of day, it is not necessary for the zone master to provide any time-critical information to the local, such as interval advance or synchronization pulse. The zone master is therefore free to poll the intersection continuously while the communications capacity is not otherwise used. For example, the zone master polls each of the 32 intersections once each 1 or 2 sec, unless the user wants to look at a graphic display of an intersection. The needs of a real-time graphic display are too large to allow polling every second, and the zone master, while maintaining the high flow of information required by the graphic display; may only be able to poll the other intersections once every 5 or 6 sec. This is not a problem, because nothing being communicated is time-critical. Consequently, the demand on the communications network is greatly reduced, and only two twisted pairs are required.

The zone master sends and receives a small packet of information from each local. The outgoing packet includes the address of the local being accessed, the selected pattern, and the time of day. The returning information includes the status of the intersection, the green return for the coordinated phases, and any system detector information for system detectors attached to that local controller.

Because the time of day is sent on each poll, the system clocks are always updated and synchronized. In the event of a failure, the locals are all synchronized, and automatic time-based coordination proceeds.

These features were already available in the zone master software running on Type 170 controllers at the time the project began. For this project, we added a further requirements that each zone master was to be capable of operating up to three completely independent control groups. This allows the user to define small control groups without a corresponding increase in the communications load. Each control group can traffic-responsively select its own pattern using the standard volume plus weighted occupancy information from system detectors, a capability used by most systems in the United States. Further subdividing of the system is possible, however, by merely defining the patterns in such a way as to be compatible with the patterns in an adjacent control group, whether or not it is part of the same zone master. For example, a control group may call for Plan 6, Offset B in a particular case. In some controllers, Plan 6 is a 50-sec cycle, while in others, Plan 6 is a 70-sec cycle. If an adjacent control group is running a 70-sec cycle, then its locals will be coordinated with the 70-sec cycle controllers in the first group. The zone master only sends the current pattern to use; the corresponding cycle length in the local controller is of no consequence to the zone master. This provides complete flexibility to the signal timing designer. At system capacity, each control group will average 10 or 11 locals, but patterns can be further arranged to be compatible across control-group boundaries as necessary. With fewer controllers on each zone master, as will usually be the case, the flexibility is even greater.

Local Controller Software

The city of San Antonio has standardized on the Type 170 controller since the early 1980s. Several times during that period, systems were designed and constructed using these controllers. Each time a new system was purchased, the hardware remained the same, but the demands on the software were increased. Each time new software was purchased, features not then available were included in the specification. The result has been that the software for Type 170 controllers has been upgraded based on the direct leadership of practitioners in the field working in a noncompetitive situation with the software developers. The downtown system is a further example of this process.

The local controller software being implemented on this system includes most of the operational features found in the latest proprietary NEMA systems and far exceeds the operational requirements of the NEMA specification. For example, the controller software includes built-in time-based coordination, several interconnect alternatives (including seven-wire, NEMA-coordinator, single-pair modem, and two-pair, two-way modem communications), two levels of railroad preempts, four levels of emergency vehicle preempts, a feature to allow the coordinated phase to gap early, and the ability to allow controlled accommodation of pedestrian timing when infrequently called. Also included is a pretimed mode, very useful in the downtown system, which allows the pedestrian walk intervals to expand to use all the time available. The software also includes features not usually found in NEMA-plus controllers, such as the ability to collect and store data from all local detectors and the ability to monitor the length of actuated phases.

In addition to these features, the specification for the downtown system has added an improved method of pattern transition. Each phase has a defined timing parameter known as the transition minimum. This interval is longer than the initial time for the phase, which is (and should be) a function of the type and layout of the approach detectors. During a transition, however, these transition minimums will not be violated. When the controller receives a directive to change patterns from the zone master, it will calculate the time necessary to hold the coordinated phase until it is in step with the new pattern. The controller will then add up the transition minimums of successive phases to see if, by timing these minimums, the controller can get in step in one cycle. If not, the controller adds one cycle length to the transition time, and prorates the force-offs (splits) over that transition time. The resulting transition cycle provides the same percentage split to each phase as before, and still gets in step within a single cycle. With such rapid transitions, traffic-responsive operation can reasonably

and effectively use much shorter control periods than is practical in a UTCS system, say, 5 min instead of 15 min.

Terminals

The terminals used in the system are standard desktop IBM-type microcomputers. The specification required the software consultant to obtain these computers from local San Antonio vendors in order to maintain nearby service capabilities. A PC-AT (intended as a hardware spare for the rack-mounted AT), two AT portables, and two desktop machines using the 80386 microprocessors were purchased. Also included were a plotter, a laser printer, and various off-the-shelf software packages. All desktop units have large (71-megabyte) hard disks and very-high-resolution (VGA) graphics displays. As with the central computer, all components are standard items readily available from any local computer store.

The terminal user interface (TUI) is designed to run under MS-DOS on a standalone PC workstation. It provides all the basic user editing and graphics display functions typical of the best closed-loop systems currently on the market. When logging into the system, the users enter the terminal mode of the TUI to allow direct communication with UNIX and the CCP kernel. From there, the users can perform any of the functions available directly from the CCP. Once logged in, they return to direct control of TUI, which then communicates with UNIX transparently to the users. At that time, they can upload, modify, download, and store any signal timing information in the system. They can also review the system status and observe an individual intersection using real-time graphics. The individual intersection display includes a graphics representation of the particular intersection, the current time of day, pattern, local and master cycle timers, offset, and real-time displays of each green, amber, "walk," "don't walk," and vehicle detection at the intersection. Each zone master can also be observed graphically, showing the intersection status and the green return for the coordinated phase.

DISTRIBUTED PROCESSING AND TRADITIONAL APPROACHES COMPARED

Three key results of process distribution at the level implemented in San Antonio are the reduced demand on the communications network, the reduced demand on any one component of the system, and dramatically increased reliability. In the traditional UTCS approach, the central computer is responsible for intersection timing. Even with powerful and expensive communications multiplexing, this real-time load is very demanding of central hardware, and requires a large minicomputer even for systems of moderate size. In one recent implementation of such a system, the cost for the central computers to allow a theoretical build-out of 800 intersections was $377,000, not including any ancillary computer equipment such as PCs and printers. The cost of equipment in San Antonio, for all of the PCs, including plotter, etc., and including 20 zone masters, is less than $100,000 for a theoretical capacity of 640 intersections. Software cannot be directly compared because the UTCS software, which had to be extensively (and expensively) customized, was part of the overall design contract which exceeded $500,000, and the software remains in escrow with access by the purchaser restricted. The portion of San Antonio's contract with the software developers that did not include the above hardware was less than $90,000, and includes full availability (under license) of source code, including the training necessary to know how to modify it. In another situation, the cost to update a UTCS computer to use a larger hard disk required over $10,000 in modifications to the software. In San Antonio's system, such a modification would require only the cost of the hard disk (a few hundred dollars) and the time necessary to move the system files (less than a day).

The major cost difference, however, lay in the required communications network. Because only critical data are actually transferred from one level of the system to another, and because the processing is performed at the location where it is used, the load on the network is very light. The central computer communicates with each zone master at 1,200 bits/sec on a direct serial link. This link may alternatively consist of only two wire pairs using the inexpensive short-haul serial modems now available, or dial-up modems (though real-time zone supervision and monitoring is not practical with dial-up links). In San Antonio, all Type 170 controllers running zone master software are located in the control center with the central computer, and their serial ports are directly connected to the serial connectors on the communications add-on boards.

Summarizing the cost of the system, we are paying approximately $8,000 per intersection, including the local and zone controllers, central equipment, control center remodeling, software development by consultant, and in-house engineering and construction. We plan to add as much as $1,000 to this cost for the future installation of system detectors to allow traffic-responsive operation, for a total projected cost of about $9,000 per intersection. This cost does not include communications cable, which was from the previous system. Fully expanded (640 intersections), the system would cost, excluding cable, about $6,760 per intersection, including about $6,000 for local controllers (installed).

Each zone master talks to up to 32 locals using two pairs. The standard Model 400 modems purchased with every controller are used.

Because of this very light demand on the communications plant, we were able to continue using the twisted-pair cables originally installed in 1957 for the PR system. The PR system cable includes eight circuits with nine usable pairs in each circuit for 150 intersections. We are currently using only about 40 percent of the pairs available in the cable plant. Even including the spares, the system has only 72 usable pairs entering the control center, which is sufficient to accommodate the maximum capacity of the system (40 pairs). By contrast, the UTCS system previously mentioned had over 400 pairs entering the control center, at very high cost, even considering that the system allowed local intersections to time themselves with only once-per-minute polling by the central computer.

Reliability is a key advantage to a fully distributed system. Because processing power is distributed to the lowest level possible, the criticality of key components further up the line is substantially reduced. For example, only traffic-responsive operation and continuous time-clock updating are lost when the communications network or the zone master fails, and even then they are lost only for the zone suffering the prob-

lem. The controllers revert automatically to time-based coordination on a time-of-day basis, using the timings already contained in the local hardware. If the central computer fails, the zones continue to operate independently and traffic-responsively (if so programmed). All that is lost is system-level supervision. No mechanism in the system can cause a general failure of all intersections. Such failures are, however, not unknown in the UTCS world. One of the authors saw a recent report of a UTCS system which, because of a failure in the central computer, placed hundreds of intersections on the system on conflict flash, and each local controller had to be manually reset. These kinds of failures are not possible in a system where no one element has global control at the detailed level. Most failures in the San Antonio system are monitored and repaired before any detrimental operation is seen by drivers. So far, all of the failures have been either in the local controller (to the same extent as in any other microprocessor-based controller) or in the communications network caused by the extensive construction currently under way in downtown San Antonio.

Another benefit of a high distribution of processing is the decreased physical size of the components. For example, the San Antonio system requires, including all the zone masters, three full-height racks. The rack reserved for the computer is mostly empty, and would be large enough for the small minicomputer should expansion become necessary. The UTCS example used a large minicomputer requiring six racks for the computer alone.

We hasten to affirm that the UTCS example works very well, even though future modification will be expensive and therefore not readily available to the system operators. The point in citing this example is to illustrate the potential cost savings that result by spreading the workload over enough machines so that no machine need be larger than a microcomputer, and by not communicating data to remote processors.

WHAT DOES THE FUTURE HOLD?

Traditional research into traffic signal systems has concentrated solely on the algorithms of traffic control, and little attention has been paid to the architecture of systems from a computer standpoint. Of course, traffic operations are the reason signal systems are installed in the first place, and there the emphasis must be. More study, however, of how the computer mechanisms interrelate with traffic control needs will allow us to ask more and better questions about traffic control methods. An open architecture in which many independent software developers can provide *features* for standardized hardware would result in more powerful systems with more design input by the practitioners who must implement them. Once a completely open system, such as PC users enjoy, is generally available, practitioners will be able to work with researchers to experiment with different control strategies; then the real questions will be able to be asked.

The San Antonio system is not particularly innovative, from either a traffic control standpoint or a computer system standpoint. What is innovative, we believe, is the system design approach that emphasizes good computer system thinking within a solid foundation of traffic control experience. By synergistically taking the best from each technology, the San Antonio design has illustrated the huge cost savings and other advantages of fully distributed processing, while giving the end user unprecedented access to the mechanics of the system.

We are convinced, however, that this step is only the beginning. As local intersections increase in power, more and more thinking will be done at the local intersection level. Once practitioners and researchers can program the operation of individual controllers at the algorithm level, then basic calculations (e.g., volume/capacity or residual queuing) can be done locally, greatly reducing the task of real-time signal timing optimization. Thus the door will be opened for widespread research in parallel processing and neural network technologies now causing excitement in other parts of the computer business.

Researchers in particular need these capabilities as they move more deeply into adaptive control systems and intelligent vehicle-highway systems. With control systems based on open hardware and software, researchers will be able to develop, one piece at a time, the complex control algorithms that will be required of systems that learn.

The experience in San Antonio, despite the smallness of the step in that direction, leads to the conclusion that these new technologies will only flourish under the open architecture now so important in most of the computer industry.

REFERENCES

1. R. Wilshire, R. Black, R. Grochoske, and J. Higginbotham. *Traffic Control Systems Handbook, Revised Edition—1985*. Report FHWA-IP-85-11. FHWA, U.S. Department of Transportation, 1985, pp. 7.25–7.30.
2. H. E. Haenel. *Texas Stand-Alone Arterial Systems*. Texas Department of Highways and Public Transportation, 1981.
3. *Type 170 Traffic Signal Controller System—Hardware Specification*. Report FHWA-IP-78-16. FHWA, U.S. Department of Transportation, 1978.
4. *Traffic Control Systems*. Standards Publication TS1-1983. National Electrical Manufacturers Association, Washington, D.C., 1983.

Publication of this paper sponsored by Committee on Traffic Signal Systems.

Development of a Self-Organizing Traffic Control System Using Neural Network Models

TAKASHI NAKATSUJI AND TERUTOSHI KAKU

A multilayer neural network model is introduced in order to realize a self-organizing traffic control system. The neural model inputs split lengths of signal phases and outputs measures of effectiveness such as queue lengths or performance indexes. The operation is separated into two processes, a training process and an optimization process. In the training process, iterations of the training operation by the backpropagation method were effective in forming a steady input-output relationship between splits and measures of effectiveness. In the optimization process, a stepwise method combining the Cauchy machine with a feedback method was proposed. The Cauchy machine is a sort of Monte Carlo method and gives the adjustments in a statistical way. This machine was introduced to urge the convergence and avoid the entrapment into local minimums. The feedback method is based on the steepest descent method and gives the adjustments in a deterministic way. This method has a self-organization ability because it can make adjustments that are closely related to traffic situations. The neural model was applied to a road network consisting of three intersections, and split lengths were optimized in order to minimize the squared sum of queue lengths on inflow links. The neural network model was able to give approximated splits and queue lengths that were in good accordance with analytical ones.

Today, most large cities in industrialized countries are confronted with chronic traffic congestion. With regard to this problem, the Organization for Economic Cooperation and Development (OECD) (*1*) issued a report on traffic management systems in urban areas. It states that future traffic systems should be operated on the self-organizing principle, in which the system would alter the basic form of the control law to respond not only to variations in traffic conditions but also to changes in transportation policies. Moreover, it says that applications of artificial intelligence techniques such as knowledge-based expert systems and fuzzy logic would be effective tools for realizing such intelligent traffic management systems. Because neural network models are also characterized by the ability of self-organization, they would serve to develop future traffic control systems.

Although neural computers have not yet been put into practice, neural network models, which are fundamental concepts of neural computers, have the potential of being able to compute in parallel and being able to learn from past experience. In particular, the self-organization ability is expected to have great effect on future traffic management systems because neural models are able to learn without any knowledge of the system and any logic such as if-then operations in expert systems. In other words, they are able to establish a characteristic input-output relationship without any preliminary information of the system. Therefore, they seem to be applicable even to nonlinear, nonstationary, or nonlogical problems. We are developing a macroscopic traffic simulation program using these characteristics of neural network models. So far, we have applied them to traffic control problems such as short-term prediction of traffic variables, traffic-responsive selection of prestored timing plans, traffic assignment, and split optimization for an isolated intersection under a criterion of the minimum queue length (*2*,*3*).

With regard to optimization of signal parameters, entrapment into local minimums is a serious and inevitable difficulty. In the hill-climbing method adopted in TRANSYT (*4*,*5*), a traffic optimization program used throughout the world, escape from local minimums is the major problem. Because some neural network models have the ability to escape from local minimums by introducing some stochastic techniques, they are expected to be effective in overcoming this difficulty. Furthermore, in optimal traffic control, application to a large-scale network is another difficulty because it takes great computation time. A hierarchical technique, first proposed by Singh and Tamura (*6*), is a superseding approach to overcome this difficulty. This method, however, is difficult to understand because it requires some mathematical knowledge. Because neural computers, if they are to be realized in the near future, have the ability of parallel processing, they are potentially applicable to large-scale networks.

This paper is mainly concerned with applications of a neural network model to optimize splits of signal phases. First, we briefly introduce the fundamental ideas of a multilayer neural model and the corresponding training algorithm, the backpropagation method. Second, we formulate optimal traffic control problems using the neural network model. In this formulation, we adopted two kinds of optimization criteria: the minimum queue length and the minimum performance index, which is a weighted sum of delays and stops. To avoid entrapment into a local minimum and urge the convergence to a global minimum, we proposed a stepwise method that combined the Cauchy machine with a feedback method in sequence. Finally, based on numerical analyses, we conclude that the neural network model has a good possibility for the development of future traffic control systems.

Civil Engineering Department, Hokkaido University, Kita 13, Nisha 8, Kita-ku, Sapporo, 060, Japan.

NEURAL NETWORK MODEL

Artificial Neurons

Artificial neurons are designed to emulate the basic mechanism of biological neurons. Figure 1 (left) shows a model that implements this function. A set of outputs (y_{i1}, y_{i2}, . . ., y_{iN}) from other neurons and a bias input (I_i) from itself are applied to a neuron (i). Each output is multiplied by synaptic weights (W_{i1}, W_{i2}, . . ., W_{iN}) and summed up algebraically:

$$x_i = \sum_{j=1}^{N} W_{ji} y_j + I_i \tag{1}$$

The signal x_i is activated by a function, which is called an activation function or a response function, as shown in Figure 1 (right):

$$y_i = F(x_i) \tag{2}$$

We adopted here a sigmoid function, $F(x) = 1/[1 + \exp(-x)]$, as the activation function. This nonlinear function prescribes the fundamental capability, as well as synaptic weights, of neural network models. Details of artificial neurons can be found in Wasserman (*7*).

Multilayer Neural Network

A multilayer neural network model was used in this analysis, as shown in Figure 2. The neural system consists of several layers: an input layer, some hidden layers, and an output layer. Assume that the neurons in the input layer serve only as distributors. The original input signals are normalized there and transmitted to the next layer. Therefore the first hidden layer, Layer B, has the same number of neurons as the input layer. Neural operations take place at the hidden layers and the output layer. The output layer produces the objective signals.

To obtain precise output signals, the synaptic weights must be adjusted. This adjustment is called the training. The backpropagation method (*8*) is used for training of multilayer networks. The method is based on the steepest descent method, in other words, the delta rule: Synaptic weights are adjusted so as to minimize the error between the output signals and the target signals, which are desired results determined externally. Letting y_k be the output signal and z_k be the target signal at the kth neuron in the output layer, and letting W_{ij} and W_{jk} be the synaptic weights between the layers shown in Figure 2, the error function is defined:

$$E = \frac{1}{2} \sum_k (y_k - z_k)^2 \tag{3}$$

Differentiating this error function with respect to W_{jk} and W_{ij} in sequence, we obtain the following expressions for adjusting synaptic weights:

$$\delta w_{jk} = \eta(z_k - y_k) y_j y_k (1 - y_k) \tag{4}$$

$$\delta W_{ij} = \eta \sum_k \delta W_{jk} W_{jk} y_i y_j (1 - y_j) \tag{5}$$

where η is the training rate coefficient in the range of 0 to 1. In actual computations, some constants are introduced to smooth the adjustments and urge the convergence. Noting that the error $\Sigma \delta W_{jk} W_{jk}$ in Equation 5 corresponds to $z_k - y_k$ in Equation 4, we can derive the adjustments for the upper layers in sequence.

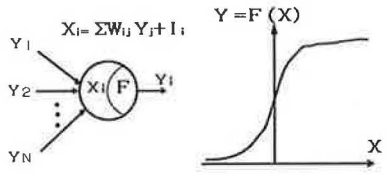


FIGURE 1 Neural network model: *left*, artificial neuron; *right*, activation function.

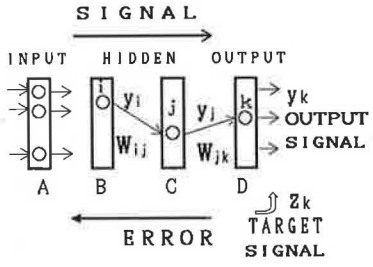


FIGURE 2 Multilayer neural network.

OPTIMAL TRAFFIC CONTROL PROBLEM

Neural Network Model for Estimating Optimal Splits

Figure 3 shows a neural network model for estimating optimal splits. It consists of four layers of neurons, Layers A to D. This neural network model describes the relationship between control variables (splits of signal phases) and objective variables (traffic variables such as queue lengths or performance indexes). That is, it inputs splits into Layer A and outputs

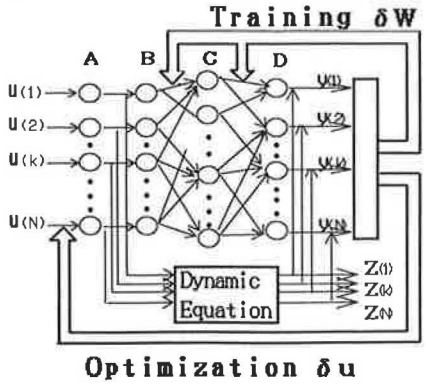


FIGURE 3 Multilayer neural network model for estimating optimal splits.

traffic variables on inflow links from Layer D. Although it is not shown in Figure 3, the traffic volumes on inflow links are also given to the neural system externally. As mentioned, because neurons in Layer A serve only as distributors to Layer B, the number of neurons in Layer B is equal to that of Layer A. The number of neurons in Layer D is the same as the number of inflow links. However, the number of neurons in Layer C used to be determined from numerical manipulation. In this case, we found that equal numbers of neurons in Layers C and D produced acceptable results. Furthermore, it should be noted that by using the time sequence of splits and traffic variables, it is possible to optimize the splits varying with time. For example, suppose an isolated intersection with four arms that is operated by two signal phases. By estimating the splits that vary every cycle, the number of neurons of the input layer is $2 \times N$ and that of the output layer is $4 \times N$, where N is the number of cycle periods.

Dynamic Equation

As mentioned, the neural network model in this analysis requires iterative trainings to adjust synaptic weights. Training signals are given by dynamic equations that are defined by objective variables and control variables. The internal dynamic model in Figure 3 produces those training signals. We formulate two kinds of dynamic equations for a simple road network system: one for queue length and the other for performance index (PI), which is used in the TRANSYT program (*4*,*5*). In this analysis, we assume for simplicity that the cycle length is common over the network and does not vary with time. Furthermore, we assume that there are no offsets between adjacent intersections.

First, we present the dynamic equation with respect to queue length. Assume a road network that consists of several intersections. Each intersection has inflow links of n_i and signal phases of p_i. We denote the split and the queue length at cycle time k by $\mathbf{y}(k)$ and $\mathbf{u}(k)$, which are column vectors of $N = \Sigma\, n_i$ and $P = \Sigma\, p_i$, respectively. The dynamic equation is given by

$$\mathbf{y}(k+1) = \mathbf{y}(k) + B_0\mathbf{u}(k) + B_1\mathbf{u}(k-1) + \ldots + B_M\mathbf{u}(k-M) + \mathbf{q}(k) \quad (k = 0, 1, \ldots, K-1) \tag{6}$$

where $\mathbf{q}(k)$ is the input flow vector of N, and B_m is the control weighing matrix of $N \times P$, which is defined by saturation flow rates on inflow links. In this analysis, we adopted an optimization criterion that minimizes the squared sum of queue length:

$$J = \sum_{k=1}^{K} \sum_{i=1}^{N} y_i(k)^2 \tag{7}$$

As an example, we suppose a simple road network consisting of two intersections as shown in Figure 4. Each intersection has two phases, one for the eastbound traffic movement and one for the southbound movement. Moreover, we denote the inflow rate at the stop line on link i by $q_i(k)$ and

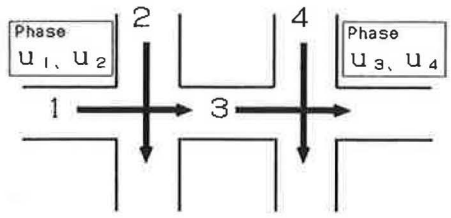


FIGURE 4 Entrance link and internal link.

the saturation flow rate by s_i. For entrance links, such as Links 1, 2, and 4, the dynamic equation is given by

$$y_i(k+1) = y_i(k) + q_i(k) - s_i u_i(k) \qquad (i = 1, 2, 4) \tag{8}$$

For internal links, in this case Link 3 alone, the inflow at stop line depends on outflows from the upstream links and splits. Denoting the inflow at entrance by $p_3(k)$, we can derive the dynamic equation as follows:

$$y_3(k+1) = y_3(k) + \sum_{m=0}^{M} \gamma_{3,m} p_3(k-m) - s_3 u_3(k) \tag{9}$$

$$p_3(k-m) = s_1 u_1(k-m) \tag{10}$$

where $\gamma_{3,m}$ is the dispersion coefficient. The value of M is determined from a correlation analysis between the upstream flows and the downstream flows. Assembling these equations, we obtain a dynamic equation that is identical to Equation 6. For details, refer to Singh and Tamura (*6*).

Next we present the dynamic equation with respect to the performance index. In this case, we have to divide each cycle period into steps of equal duration and formulate the dynamic equation, which is identical to Equation 6, for each time step t. By integrating all of those traffic profiles for each step, we can define some measures of effectiveness, such as delay and stops, for each cycle time. As defined in TRANSYT, the performance index on inflow Link i for Cycle Time k is calculated as follows:

$$PI_i(k) = \mathrm{DLY}_i(k) + \kappa_i \mathrm{STP}_i(k) \tag{11}$$

where

$\mathrm{DLY}_i(k)$ = total delay on Link i for Cycle k,
$\mathrm{STP}_i(k)$ = number of stops on Link i for Cycle k, and
κ_i = stop penalty coefficient.

We took this performance index as the objective variable, $y_i(k)$. Also in this case, the same optimization criterion as Equation 7 was used. The TRANSYT users manuals (*4*,*5*) provide details of the definition of the delay and stops.

Both the objective and the control variables are subject to constraints for every Cycle Time k:

$$\mathbf{0} \le \mathbf{y}(k) \le \mathbf{Y}\mathrm{max} \tag{12}$$

$$\mathbf{U}\mathrm{min} \le \mathbf{u}(k) \le \mathbf{U}\mathrm{max} \tag{13}$$

$$u_{i,1}(k) + u_{i,2}(k) + \ldots + u_{i,p_i}(k) + ls_i = 1 \tag{14}$$

where $u_{i,r}(k)$ is the rth split at intersection i, and ls_i is the ratio of loss time to cycle length.

Computational Procedures

Referring to Maeda (*9*), we separated the operation of this neural model into two processes, the training process and the optimization process. In the training process, synaptic weights are adjusted so that the output signals from the output layer coincide with those from the internal model as much as possible. This adjustment can be done by the direct use of the backpropagation method. On the other hand, the optimization process performs iterative adjustments of splits to minimize the objective function under given constraints.

Figure 5 shows the block diagram for estimating optimal split lengths. First we have to perform initial training. After preparing a set of traffic volumes that arrive at entry links and scores of split patterns that are randomly generated, we adjust synaptic weights of the neural network model. We repeat the backpropagation operations until the squared sum of the deviations between the output signals and the target signals becomes sufficiently small. We iterate initial training until the neural models satisfy the convergence condition for all split patterns.

Next, we predict traffic volumes on entry links for several cycle periods. There are many prediction methods; however, because the discussion on the methods is beyond the scope of this paper, we assume that precise traffic volumes are already being predicted. Because those traffic volumes are different from those in the initial training process, we have to adjust synaptic weights again. However, traffic volumes do not change drastically, so we can adjust them through several iterations of the backpropagation method.

Establishment of a steady relationship between splits and objective variables makes it possible to estimate optimal splits properly. To do this, we proposed a combined, stepwise method. Theoretically, by repeating the procedures from prediction to optimization in sequence, it might be possible to estimate optimal splits in real time. However, under the present circumstances, the neural approach takes more computation time compared to the conventional analytical methods.

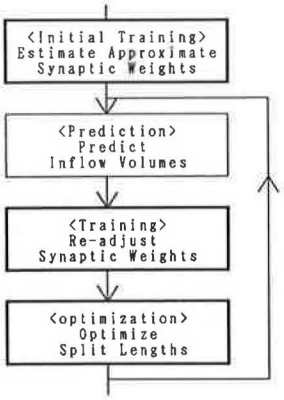


FIGURE 5 Computational procedures for estimating optimal splits.

In addition, because the modeling of offsets is being left unresolved, we analyze the splits for a set of traffic volumes.

Stepwise Method

To optimize split lengths, it is necessary to adjust them iteratively in order to minimize the objective criterion under given constraints. Referring to Wasserman's technique (*7*), we proposed another combined method consisting of a two-step process. First, we adjust splits based on the Cauchy machine to avoid entrapment into a local minimum and urge the convergence. Next, we adjust the splits using a deterministic technique similar to the backpropagation method. This combined algorithm is called the stepwise method. Entrapment into a local minimum is a serious and inevitable difficulty in some minimum-seeking problems. To overcome this difficulty, some stochastic methods, such as the Boltzman machine, the Gaussian machine and so on, have been proposed in neural network analyses. Szu [Wasserman (*7*)] developed a stochastic method, called the Cauchy machine, for steady convergence to a global minimum. It is a sort of Monte Carlo method; by adding small changes, which follow the Cauchy distribution, into the present split values, we accept those changes if they improve the objective function, and abandon them otherwise. The probability density function of the Cauchy distribution is given by

$$p(x) = T(t)/[T(t)^2 + x^2] \tag{15}$$

$$T(t) = T_0/(1 + t) \tag{16}$$

where $T(t)$ is the artificial temperature, and T_0 is the initial temperature. Integrating the density function, we obtain the following distribution function:

$$P(x) = \arctan[x/T(t)] \tag{17}$$

Then, resolving for x yields

$$x = \rho T(t)\tan[P(x)] \tag{18}$$

where ρ is a coefficient in the range of 0 to 1. Regarding x in the above equations as split change δu, we can find the change as follows:

1. Select a random value from a uniform distribution over the interval $(-\pi/2, \pi/2)$.
2. Substitute it into $P(x)$ in Equation 18 and calculate the change.
3. Retain it if the adjustment improves the objective function, and return it to the previous value if otherwise.
4. Decrease the deviation of the Cauchy distribution and go back to Step 1 and repeat again.

This algorithm can drastically reduce the computation time because it adopts an annealing scheme in which the temperature is decreased inversely linearly, rather than inversely logarithmically as in the Boltzman machine.

Next we adjust the splits in a deterministic way similar to the backpropagation method in the training process. The steep-

est descent method is used again. By differentiating the objective function of Equation 7 with respect to $u_i(k)$, we can easily derive the following expression for adjustments of the splits:

$$\delta u_i = \eta \sum_k y_k^2(1 - y_k) \sum_j W_{ij}W_{jk}y_j(1 - y_j) \quad (19)$$

where η is a coefficient ranging 0 to 1. Because those adjustments in this optimization process are not backpropagated as in the training process, we call such a process the feedback method. The adjustments in Equation 19 are related to synaptic weights, which vary with traffic situations. This means that we are able to alter the parameters for adjusting splits automatically corresponding to the change of traffic situations. This self-organizing ability is a promising feature of neural network models. Furthermore, although in this analysis we adopted the optimization criterion given by the form of Equation 7, we can derive similar expressions to Equation 19 for any criteria only if they are differentiable with respect to $u_i(k)$.

NUMERICAL EXPERIMENTS

Training

The ability of the neural model depends on how precisely the synaptic weights are adjusted. The initial training process requires scores of training operations for each split pattern. Using an isolated intersection as an example, we explain how the synaptic weights were adjusted by the backpropagation method. As shown in Figure 6, the intersection has eight inflow links and is operated with three signal phases. We assume that the cycle length is 120 sec and the simulation period consists of four cycles. Furthermore, we assume for simplicity that inflow rates and split lengths are constant over the simulation period. This assumption is not requisite; a problem for time-variant splits is also presented. Detailed information on the inflow links is shown in Table 1.

First, we discuss the problem of the minimum queue length. We build up a neural network model, shown in Figure 3, in which the neuron in the input layer corresponds to the split length of each signal phase and the one in the output layer to total queue length on each inflow link. That is, the number of neurons in the input layer is three, and that of the output

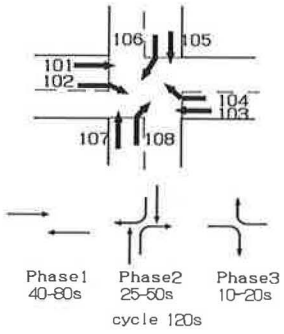


FIGURE 6 Isolated intersection.

TABLE 1 LINK DATA FOR ISOLATED INTERSECTION EXAMPLE

Link No.	Saturation Flow Rate veh./cycle	Inflow Volume veh./cycle	Initial Queue veh.
101	113	26.67	50
102	53	1.53	10
103	113	30.20	100
104	53	3.67	10
105	113	17.67	50
106	53	1.60	10
107	113	11.13	50
108	53	2.87	10

layer is eight. To perform initial training, we prepared in advance 20 randomly generated sets of split patterns that satisfy the constraint conditions. We then calculated the total queue lengths on inflow links for four cycles and made them the training signals for each split pattern.

Figure 7 shows how the estimation error of the synaptic weights would decrease with iterative operations of the backpropagation method for some split patterns, Patterns 1, 2, 11, and 20. Here, the error was calculated by the root mean squared (RMS) value of the deviation between queue lengths by the neural system and those by the dynamic system. We truncated the iteration when the error became less than 10. Roughly speaking, this means an error of 2 percent because both the output and the target signals were normalized by a number of 500. Figure 7 shows that once synaptic weights had been adjusted for the first split pattern, they were easily adjusted for the other ones. However, it also shows that the completion of adjustments for a split pattern brings the deterioration of synaptic weights for the other patterns. Therefore, we have to repeat scores of training operations until the RMS error becomes less than the threshold for all split patterns. Figure 8 shows the variation of the maximum and the average RMS error with iterations of training operations. The average RMS error represents the root mean squared value of the RMS error for each split pattern. The figure shows that the synaptic weights are improved gradually but certainly. In this case, it took 357 iterations to complete the training, and the final average RMS error was 4.97, nearly half of the truncation threshold.

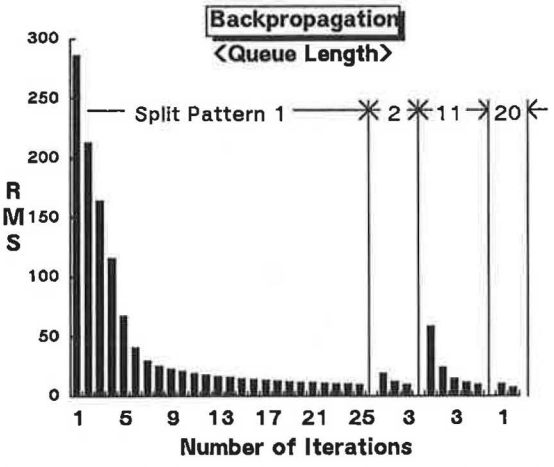


FIGURE 7 Backpropagation operations in the initial training process for some randomly generated split patterns (output variable is queue length).

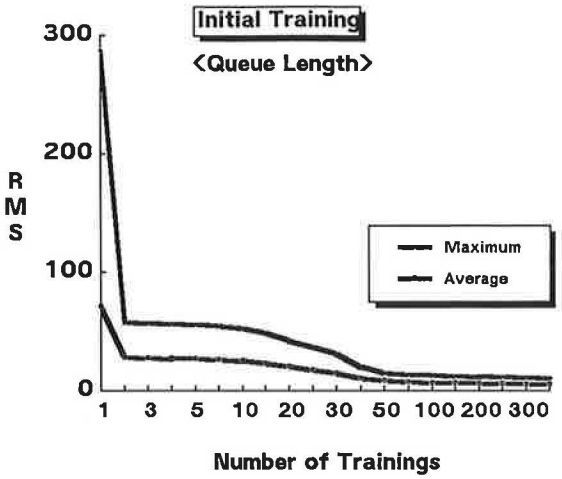


FIGURE 8 Adjustment of synaptic weights by iterations of the training. A training consists of iterative operations of backpropagation method for all split patterns (output variable is queue length).

To examine the ability of the neural system that completed the adjustment of synaptic weights, we prepared other split patterns. We then calculated the total queue lengths on the links using the neural system and compared them with analytical ones, which were given by the dynamic model. Figure 9 is the histogram of the RMS error for 100 sets of split patterns. It shows that the RMS error was less than 5.0 for more than 60 split patterns. For only three patterns, it exceeded the threshold of 10.0. The maximum RMS error was 10.66. This means that the initial training by 20 split patterns was sufficient.

Similarly, for the problem of the minimum performance index, we can build up another neural system that has a steady input-output relationship between split lengths and the corresponding performance indexes on inflow links. The difference lies only in the dynamic model for estimating target signals. We performed the initial training for the same intersection, shown in Figure 6, with the same split patterns as in the previous problem. In this analysis, we divided a cycle length of 120 sec into 60 steps of 2 sec. Parameters to calculate the delay time were the same as those in TRANSYT-7F. The stop penalty of five was used for all links. The output and target signals were normalized by a value of 600. It took 206 iterations of training operations to adjust synaptic weights completely. Figure 10 shows the distribution of the RMS errors for 100 sets of untrained split patterns. Although results in Figure 10 are not as good as in Figure 9, 36 split patterns had RMS errors less than 5.0, and only 6 patterns exceeded 10.0. The average and the maximum RMS errors were 5.06 and 13.42, respectively.

Optimization

Figure 11 shows how the stepwise method worked in the optimization process. We took the same problem in Figure 6. Figure 11a is for the minimum queue length, and Figure 11b is for the minimum performance index. We compared the stepwise method with the feedback method, in which no Cauchy operations were applied. The x-axis represents the number of iterations and the y-axis represents the values of the objective function, the squared sum of queue lengths for Figure 11a and that of performance indexes for Figure 11b. Figure 11 shows that there is little difference between the two methods. The feedback method also reaches the global minimum without being entrapped into a local minimum because the intersection is isolated and operated with simple signal phasing. However, the stepwise method was effective to urge the convergence, particularly for the performance index. Next, we present another example in which the stepwise method was effective to avoid local minimums.

Practical Simulation

As a practical example for real intersections of complicated geometry and phasing, we chose a road network that consists of three intersections, which was analyzed by Singh and Tamura (*6*). The configuration of those intersections and inflow links is given in Figure 12. Every intersection is operated with two phases. That is, the road network has 12 inflow links and six phases in total. The simulation period is three cycles. Detailed information on saturation flow rates and inflow volumes are given along with the initial values in Table 2. In this problem, split lengths are optimized every cycle period so as to minimize the squared sum of queue lengths on the inflow

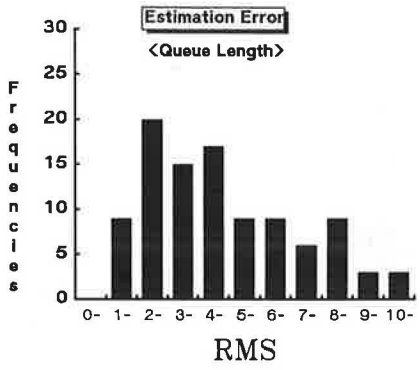


FIGURE 9 Distribution of RMS errors for 100 sets of split patterns (output variable is queue length).

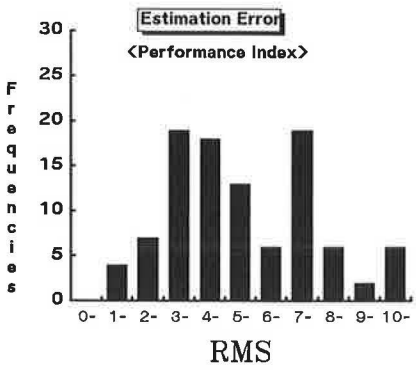
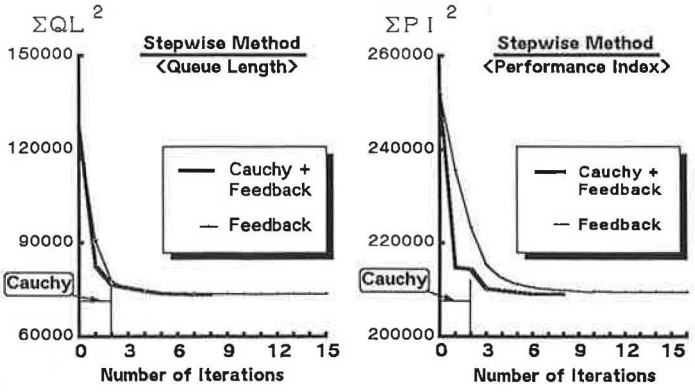


FIGURE 10 Distribution of RMS errors for 100 sets of split patterns (output variable is performance index).



(a) Queue Length (b) Performance Index

FIGURE 11 Optimization process for an isolated intersection.

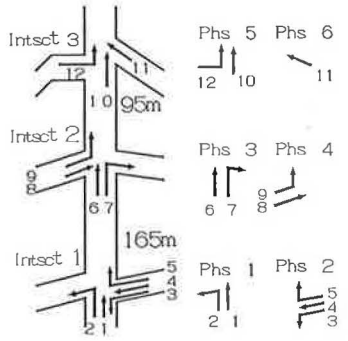


FIGURE 12 Road network for practical simulation (*6*).

TABLE 2 LINK DATA FOR ROAD NETWORK EXAMPLE (*6*)

Link No.	Link Length m	Saturation Flow Rate veh./cycle	Inflow Volume veh./cycle	Initial Queue veh.
1	-	65	7.6	30
2	-	25	3.0	30
3	-	34	23.7	70
4	-	31	21.7	70
5	-	4	27.0	70
6	165	64	Intrnl	40
7	165	26	"	40
8	-	132	15.0	30
9	-	34	4.0	30
10	95	96	Intrnl	20
11	-	90	2.4	30
12	-	25	2.0	30

links. Referring to Singh and Tamura (*6*), the dynamic equations for this problem reduce to

$$y_1(k+1) = y_1(k) + q_1(k) - s_1u_1(k)$$

$$y_2(k+1) = y_2(k) + q_2(k) - s_2u_1(k)$$

$$y_3(k+1) = y_3(k) + q_3(k) - s_3u_2(k)$$

$$y_4(k+1) = y_4(k) + q_4(k) - s_4u_2(k)$$

$$y_5(k+1) = y_5(k) + q_5(k) - s_5u_2(k)$$

$$y_6(k+1) = y_6(k) + 0.7s_1u_1(k-2) + 0.7s_2u_2(k-2) - s_6u_3(k)$$

$$y_7(k+1) = y_7(k) + 0.3s_1u_1(k-2) + 0.3s_2u_2(k-2) - s_7u_3(k)$$

$$y_8(k+1) = y_8(k) + q_8(k) - s_8u_4(k)$$

$$y_9(k+1) = y_9(k) + q_9(k) - s_9u_4(k)$$

$$y_{10}(k+1) = y_{10}(k) + s_6u_3(k-1) + s_9u_4(k-1) - s_{10}u_5(k)$$

$$y_{11}(k+1) = y_{11}(k) + q_{11}(k) - s_{11}u_6(k)$$

$$y_{12}(k+1) = y_{12}(k) + q_{12}(k) - s_{12}u_5(k) \qquad (20)$$

where

$y_i(k)$ = queue length on Link i at cycle k,
$q_i(k)$ = inflow rate,
s_i = saturation flow rate, and
$u_r(k)$ = split length for signal phase r.

The values of 0.7 and 0.3 represent the dispersion coefficients. All splits were constrained to lie between 0.2 and 0.7 and to satisfy the conditions of $u_1(k) + u_2(k) = u_3(k) + u_4(k) = u_5(k) + u_6(k) = 0.9$.

We built up a neural network model as shown in Figure 3. However, distinct from the one in the previous discussion, it inputs the time sequence of the split lengths and outputs that

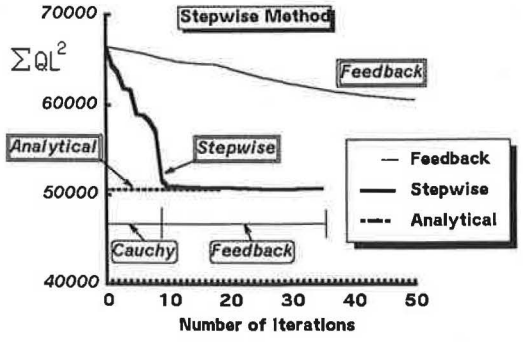


FIGURE 13 Optimization process for a road network.

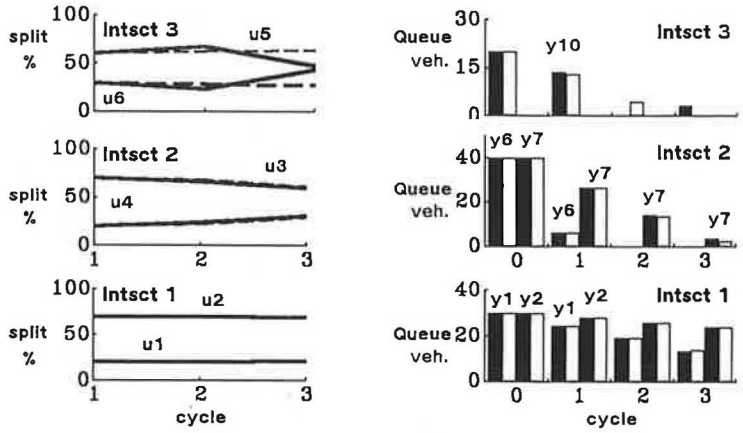


FIGURE 14 Optimal splits and queue lengths.

of the queue lengths because we have to estimate the splits that vary with cycle. Therefore, the number of neurons in the input layer is 6 × 3 and that of the output layer is 12 × 3, where the value of 3 is the number of cycle periods.

Figure 13 shows how the stepwise method effectively optimizes those split lengths. We compared three methods: an analytical method by a hierarchical approach, the feedback method without the Cauchy machine, and the stepwise method. We directly referred to results by Singh and Tamura (*6*) for the analytical method. The x-axis represents the number of iterations and the y-axis represents the squared sum of queue lengths. The comparison shows that although the feedback method was entrapped into a local minimum and took a large number of iteration values, the stepwise method succeeded in reaching the global minimum. Figure 14 shows the optimized control sequence and the corresponding queue lengths on main inflow links, Links 6, 7, and 10. The real lines are for the stepwise method and the dotted ones are for the analytical solutions. They show that solutions by the stepwise method were in good agreement with those of the analytical method.

CONCLUSIONS

Presuming applications to future traffic control systems, we introduce a neural network model, which is characterized by its self-organizing ability, for split optimization problems. First, we built up a multilayer neural network model that inputs split lengths of signal phases and outputs objective variables. We adopted two kinds of control criteria, the minimum queue length and the minimum performance index. Next, we divided the problem into two processes, the training process and the optimization process. In the training process, the backpropagation method was effective to adjust the synaptic weights. We established a steady input-output relationship by scores of iterations of training operations. In the optimization process, we proposed a stepwise method, combining the Cauchy machine and the feedback method, to urge the convergence and avoid entrapment into local minimums. Through numerical analyses, we showed that this method improved the convergence into a global minimum and that solutions by this method were in good accordance with analytical ones.

This paper is only the first step for realization of a self-organizing traffic control system. Many problems must be solved before a neural network model can be applied to an actual road network. One problem is the optimization of offsets. Without modeling the parameters, it is impossible to realize real self-organizing traffic control. The modeling of dispersion phenomena of vehicle platoons is another problem. This modeling is requisite for sophisticated traffic flow simulation. The improvement of computation time is also important. Because we used a conventional digital computer, the neural models presented here required much more computation time than the corresponding analytical method. Some emulation machines that have several parallel processors and are able to realize particular neural algorithms have already been developed. However, the application of a neural network model to an actual road network system would require the development of a neural computer with thousands of parallel processors. We confirm that such neural computers will be realized in the near future.

ACKNOWLEDGMENT

The authors wish to express their thanks to Dr. Tamura of Osaka University for allowing reference to the examples in his paper.

REFERENCES

1. *Dynamic Traffic Management in Urban and Suburban Road Systems*. Organization for Economic Cooperation and Development, Road Transportation Research, 1987.
2. T. Nakatsuji and T. Kaku. Application of Neural Network Models to Traffic Engineering Problems (in Japanese). *Proc., Infrastructure Planning*, Vol. 12, 1989, pp. 297–304.
3. T. Nakatsuji and T. Kaku. Application of Neural Network Models to Traffic Engineering Problems, *Proc., 11th International Symposium on Transportation Traffic Theory*, Yokohama, Japan, July 1990, pp. 291–306.

4. R. A. Vincent, A. I. Mitchell, and D. I. Robertson. *User's Guide to TRANSYT*, Version 8. U.K. Transport and Road Research Laboratory LR888, 1980.
5. *TRANSYT-7F Self-Study Guide*. FHWA, U.S. Department of Transportation, 1986.
6. M. G. Singh and H. Tamura. Modeling and Hierarchical Optimization for Over-Saturated Urban Road Traffic Networks. *International Journal of Control*, Vol. 20, No. 6, 1974, pp. 913–934.
7. P. D. Wasserman. *Neural Computing*. Van Nostrand Reinhold, New York, 1989.
8. D. E. Rumelhart et al. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing*, Vol. 1. MIT Press, Cambridge, Mass., 1986.
9. Y. Maeda, M. Kawato, Y. Uno, and R. Suzuki. *Multi-Layer Neural Network Model Which Learns and Generates Human Multi-Joint Arm Trajectory*. Japan IEICE Technical Report MBE87-133, 1988, pp. 233–240.

Publication of this paper sponsored by Committee on Traffic Signal Systems.