TRANSPORTATION RESEARCH
RECORD

No. 1457

Highway Operations, Capacity, and Traffic Control

PART 1

1994 TRB Distinguished Lecture

ADOLF D. MAY

PART 2

Traffic Flow and Capacity

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1994

Transportation Research Record 1457
ISSN 0361-1981
ISBN 0-309-06100-8
Price: $45.00

Subscriber Category
IVA highway operations, capacity, and traffic control

Printed in the United States of America

Sponsorship of Transportation Research Record 1457

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

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

Committee on Highway Capacity and Quality of Service
Chairman: Adolf D. May, University of California at Berkeley
Secretary: Wayne K. Kittelson, Kittelson & Associates Inc.
Rahmi Akcelik, James A. Bonneson, Werner Brilon, Kenneth G. Courage, Rafael E. DeArazoza, Richard Dowling, Daniel B. Fambro, Ronald K. Giguere, Fred L. Hall, Douglas W. Harwood, Michael Kyte, Joel P. Leisch, Douglas S. McLeod, John Morrall, Barbara Katherine Ostrom, Ronald C. Pfefer, James L. Powell, William R. Reilly, Roger P. Roess, Nagui M. Rouphail, Ronald C. Sonntag, Alex Sorton, Dennis W. Strong, Stan Teply, Pierre-Yves Texier, R. Troutbeck, Thomas Urbanik II, John D. Zegeer

Committee on Traffic Flow Theory and Characteristics
Chairman: Hani S. Mahmassani, University of Texas at Austin
Siamak A. Ardekani, Gang-Len Chang, Nathan H. Gartner, Fred L. Hall, Michael Kyte, Henry C. Lieu, Carroll J. Messer, Abbas Mohaddes, Kyriacos C. Mouskos, Markos Papageorgiou, Ajay K. Rathi, Nagui M. Rouphail, Michael A. P. Taylor, Michel Van Aerde

Transportation Research Board Staff
Robert E. Spicher, Director, Technical Activities
Richard A. Cunard, Engineer of Traffic and Operations
Nancy A. Ackerman, Director, Reports and Editorial Services
Susan E. G. Brown, Editor

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1993.

Transportation Research Record 1457

Contents

PART 1

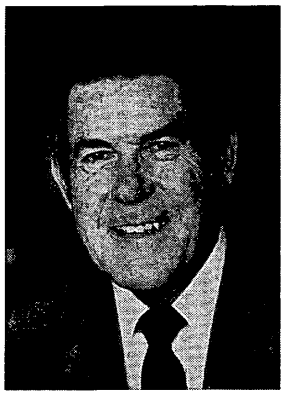
1994 TRB Distinguished Lecture

Part 1
Foreword

In 1990 the Transportation Research Board Executive Committee approved the establishment of the Distinguished Lectureship Series to recognize the career contributions and achievements of an individual in one of four areas covered by the Board's Technical Activities Division: transportation systems planning and administration (Group 1); design and construction of transportation facilities (Group 2); operation, safety, and maintenance of transportation facilities (Group 3); and legal resources (Group 4).

Those selected are provided a forum at the TRB Annual Meeting to present an overview of their technical areas, including evolution, present status, and prognosis. Adolf D. May, Professor Emeritus of the University of California, Berkeley, is the third to be honored with the TRB Distinguished Lectureship. His lecture, entitled *Traffic Management from Theory to Practice: Past, Present, Future*, was sponsored by Group 3 and presented at the 1994 Annual Meeting. It is published in this Record.

Adolf D. May, 1994 TRB Distinguished Lecturer



ADOLF D. MAY

Adolf D. May is Professor Emeritus of Civil Engineering at the University of California, Berkeley.

May, who was elected to the National Academy of Engineering in 1990, has made significant contributions to nearly every aspect of transportation involving traffic facilities and operations. He was the first director of the Chicago Area Freeway Surveillance and Control Project, a forerunner of current activities in intelligent vehicle-highway systems.

Former Vice Chair of the Department of Civil Engineering at the University of California, Berkeley, May has also taught at Northwestern University and Michigan State University. He holds a bachelor's degree from Southern Methodist University, a master's degree from Iowa State University, and a doctorate from Purdue University, all in civil engineering. Active in TRB since the early 1950s, May has served on many committees and panels. He is currently a member of TRB's Executive Committee and chairs the Highway Capacity and Quality of Service Committee. May is the author of more than 300 technical books and articles and the recipient of many awards, including the Theodore M. Matson Memorial Award of the Institute of Transportation Engineers in 1992.

Traffic Management from Theory to Practice: Past, Present, Future

ADOLF D. MAY

It is proposed that traffic management will be most successful when theory and theoreticians work closely with practice and professionals. The past, present, and future are discussed because observing the path of traffic management to its current position provides insights into future possibilities. The scope is limited to the major urban road system. After a brief overview of traffic management fundamentals and the recognition of the important contributions of TRB, emphasis is given to important topics related to freeways and freeway systems, including capacity analysis, speed-flow relationships, simulation models, and traffic management strategies. Final observations are presented in the context of "bridges between."

It is indeed an honor to have been selected to give this third annual TRB Distinguished Lecture. It is of particular pleasure to me because over the years I have held TRB in such high esteem. Many individuals and organizations have guided me throughout my career, but my association with many of you through various TRB activities during the past 40 years has been most important to me.

My contact with the Highway Research Board began in the early 1950s when the annual meetings were held at the National Academy of Sciences building on Constitution Avenue and Roy Crum was the Executive Director. HRB (now, of course, TRB) has been fortunate over the years to have the fine leadership of Fred Burggraf, Grant Mickle, Bill Carey, and Tom Deen. I would also like to acknowledge the excellent support and encouragement that those of us in traffic and operations have received from TRB staff engineers, including K. B. Johns, Dave Witheford, Bob Spicher, Dan Rosen, and Richard Cunard.

During this period my involvement with TRB has increased from committee member to committee chairman to Group Council to Executive Committee, and each new opportunity has been a rewarding experience. For those of you who have over the years participated within the TRB family, I am sure that your experiences have been equally rewarding. For those of you who are just beginning your association with TRB, I encourage you to become more involved, for TRB needs your participation and your professional and academic contributions—the more you become involved, the more you will be rewarded with an enriched experience.

The title of my presentation is *Traffic Management from Theory to Practice: Past, Present, Future*. The selection of the topic for me was easy. I deeply believe that one of the most important areas in transportation—particularly today—is managing traffic operations on the system. Further, I believe that traffic management will be most successful when theory and theoreticians work closely with practice and professionals. I have added "past, present, future" because, as it is for the surveyor, observing the path to the current position provides insights into future possibilities.

JHK & Associates, 2000 Powell Street, Suite 1090, Emeryville, Calif. 94608.

Preparing the contents of the presentation, however, was very difficult, for traffic management is a very broad topic and there are many subjects that could be covered. After considerable thought, it was decided to limit the presentation to the major urban road system; after a brief introduction to suggested fundamentals, emphasis will be given to freeways and freeway corridors. If I have not included subjects that you feel are equally important, their omission is not because they are not recognized as being important, but because time was limited.

The outline of the presentation is now covered. After introducing an overview and some fundamental concepts of traffic management, a historical perspective of a TRB committee involved in a particular aspect of traffic management will be presented. Then attention will be focused on freeways and freeway corridors, with particular emphasis given to capacity and speed-flow relationships, simulation models, and traffic management strategies. The presentation will conclude with some final observations.

TRAFFIC MANAGEMENT OVERVIEW

This overview of traffic management of the major urban road system will be presented in three parts. First, it will be suggested that the urban road system be structured as consisting of traffic operating environments that can be studied individually or in combinations to represent more complex traffic operating environments. Then a generalized analytical framework will be presented that is applicable to the various traffic operating environments. Finally, a demand-supply relationship is proposed that can be used to identify traffic operation problems and possible sets of potential solutions.

Traffic Operating Environments

It is proposed that the major urban road system be considered to consist of individual traffic operating environments or of a combination of such environments. Figure 1 identifies these individual environments and their various combinations leading to more comprehensive traffic operating environments. The urban area, for example, is made up of the arterial network and the freeway system. The arterial network consists of arterials that in turn are connected to individual unsignalized and signalized intersections. The freeway system consists of individual freeways, which in turn are connected straight-pipe sections, ramp junctions, and weaving sections. A unique traffic operating environment receiving much attention today is the freeway corridor, which integrates the arterial network with the freeway. The freeway corridor will serve as an example of traffic management theory and practice in this paper, and each topic will provide a historical perspective of development to date and anticipation for the future.

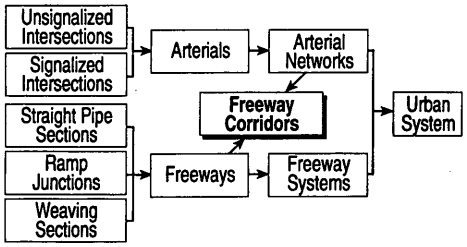


FIGURE 1 Traffic operating environments for major urban road system.

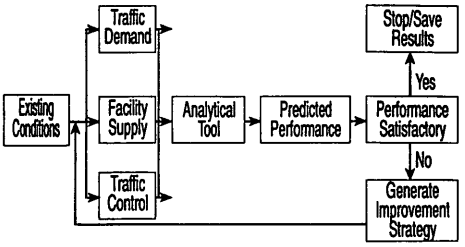


FIGURE 2 Generalized analytical framework for major urban road system.

The traffic operating environment concept for the major urban road system is suggested for several reasons. Experience has shown that larger systems cannot be managed without a full understanding of the individual components and their interactions. Therefore, it is imperative that the traffic phenomena in the smaller elements be understood completely, and then the traffic interactions between the smaller elements that make up the larger traffic operating environments, before the traffic management of the larger traffic operating environment can be addressed.

Analytical Framework

An analytical framework is proposed for the traffic operating environments identified previously. An attempt has been made to generalize the analytical framework so that it applies to all of the traffic operating environments. The proposed generalized analytical framework is shown in Figure 2.

The analytical framework begins with assembling traffic demand, facility supply, and traffic control for the existing or base conditions in the traffic operating environment being investigated. This information provides the input to the analytical process, which in turn provides an output in terms of predicted performance. If the performance is satisfactory, the analysis can terminate and the results be saved as a base for future analysis. If the performance is unsatisfactory, an improvement strategy is generated that results in the modification of the initial demand, supply, or control (or all three), and the analytical process is repeated until performance is deemed satisfactory.

More details of this analytical process will be presented later in the paper as the freeway corridor is addressed as an example of a traffic operating environment.

Problem Identification and Potential Solutions

There is commonality between traffic operating environments in terms of approaches to identifying problems and potential solutions. Although operational problems may occur when undersaturated conditions exist, such as at unsignalized and signalized intersections, the most severe operational problems in all traffic environments are those associated with oversaturation. Figure 3 suggests a way of identifying the oversaturated condition and alternative potential solution approaches.

Consider a time-space matrix in which an estimate is made of the demand and capacity in each cell of the matrix. The term D_{it} represents the demand in space domain i and time domain t. Correspondingly, the term C_{it} represents the capacity in space domain i and time domain t. The demand and capacity values are compared in each cell, and the problem is identified—that is, oversaturated condition, when demand exceeds its corresponding capacity. When this occurs the performance of the system deteriorates greatly, with inferior traffic performance, inefficient use of the road system, and harmful effects to the environment in terms of air pollution, energy consumption, and vehicle noise.

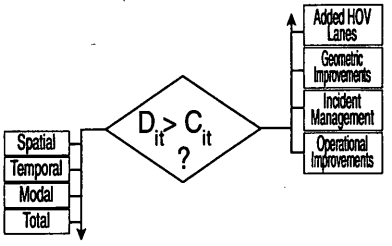


FIGURE 3 Problem identification and potential solutions for major urban road system.

What are the alternative solutions? There are three possible approaches: increase capacity, reduce demand, or combine the two. There are a number of ways to increase the capacity and reduce the demand. Capacities can be increased in several ways: operational improvements, incident management, geometric improvements, and added high-occupancy vehicle (HOV) lanes. Operational improvements might range from intersection control improvements to capacity increases by maintaining smooth, steady flow. Incident management attempts to reduce the duration and magnitude of capacity reductions due to incidents. Geometric improvements vary from providing special turn lanes at intersections to adding auxiliary lanes along the freeway. Introducing HOV lanes on arterials and freeways is another way to increase the capacity side of the equation.

There are a number of ways that demand can be reduced on critical portions of a traffic operating environment. The intent of these demand management strategies is to reduce demand on critical sections through spatial response, temporal response, modal response, and total response. Traffic diversion to parallel underused alternative routes is an example of spatial response. Spreading the demand to pre- and post-peak congested periods is an example of temporal response. Encouraging increases in vehicle occupancy and use of public transit are examples of modal response. Finally, reducing the vehicle miles traveled is what is meant by total response; examples include combining trips, changing the destination of trips, or simply no longer making trips.

HISTORICAL PERSPECTIVE OF A TRB COMMITTEE

Early Committee Activities

My involvement in TRB activities dates to the early 1950s, with attending Annual Meetings, presenting technical papers, and participating in committee meetings. Those of you who have had similar long-term association with TRB are aware of the significant efforts of individuals who volunteer their time to the work of TRB committees and the effect that these committees have had on pushing forth the frontiers of knowledge. Those of you who have more recently become involved in TRB activities might not have seen the significant long-term positive impacts of a TRB committee. Perhaps tracing the history of one TRB committee's activities and contributions will, when integrated over the several hundred TRB committees, give an appreciation of the important role that TRB has played over the years.

The TRB Committee on Highway Capacity was established in the mid-1940s. The first major accomplishment of the committee was the publication of the 1950 *Highway Capacity Manual* (HCM), which became the principal guide for capacity analysis not only in the United States but throughout the world. O. K. Normann and William Walker of what was then the Bureau of Public Roads served as chairman and secretary, respectively, of the 18-member committee.

Recognizing almost immediately the need to continuously update and expand the HCM, the committee began work on assembling new material for the revision. A comprehensive nationwide intersection study program was initiated in 1954. By 1957, the committee began detailed planning for a new edition. Progress was gradual until 1963, when a five-man task group was assigned by the Bureau of Public Roads to work with the committee on writing the new HCM. The second HCM was published in 1965. One of the major contributions of the 1965 HCM was the introduction of the concept of "level of service." O. K. Normann continued to serve as committee chairman until his death in 1964. Carl Saal served as chairman during the final publication stages of the 1965 HCM, and Art Carter served as committee secretary and played a major role in the completion of the manual. More than 30,000 copies of the 1965 HCM were printed, and it has been translated into 27 languages.

After a few years, efforts began again toward developing the third HCM. Bob Blumenthal, Jim Kell, and Carlton Robinson led the committee during this period. In the spirit of transforming the Highway Research Board to the Transportation Research Board, the committee expanded the scope of the HCM to give greater emphasis to transit capacity and to include capacity analysis of pedestrian and bicycle ways. Added attention was also given to a systems approach both for freeways and arterials. The third HCM was published in 1985.

Current Committee Activities

The committee took on new leadership in the late 1980s; its specific goals included the development of a research program in highway capacity and the revisions of several key chapters. In addition, the committee gave greater attention to the international community by involving a number of world experts on capacity and by holding midyear meetings in Germany in 1991 and Australia in 1994.

Seven of the 14 chapters of the 1985 HCM have been revised; they were published in 1994. These chapters for the most part were undertaken by the various subcommittees of the Capacity Committee; they were prepared by committee members working almost completely on a voluntary basis. The revision of each chapter was the responsibility of a subcommittee, which presented it to the total committee for approval. Subcommittee chairpersons played a very important part in this effort and should be recognized: Stan Teply, Roger Roess, Ron Pfefer, Ken Courage, Mike Kyte, and Dan Fambro. Their efforts have included field studies, analyses, and the preparation of individual chapters.

A program of research in highway capacity was developed and published as *Transportation Research Circular 371* in June 1991. The program recommended the conduct of 21 research studies as a means to address deficiencies in the present HCM and to upgrade future editions. The total cost of the research program was estimated to be $3.55 million over the next 6 years. The research program with the support of research sponsoring agencies has been very successful in that 12 of the 21 studies are either under way or were expected to begin in 1994.

Future Committee Activities

Efforts will continue to ensure that all research studies needed for updates and expansions will be undertaken and completed. The target is to complete the new HCM in the year 2000. In the Matson Award paper presented by the Institute of Transportation Engineers in 1992, I attempted to indicate what the year 2000 manual might look like. Five key phases were identified to describe the direction for the next HCM: multimodal, systems, oversaturation, computerization, and user interface. The paper concluded with this summary:

> In summary the year 2000 capacity manual may have an entirely different look than previous capacity manuals. It may not be printed on paper but be part of the computer software. It will most likely include multimodal analysis on a systems basis and handling the oversaturated condition. The user interface with the computer program will be much improved with extensive diagnosis to aid the user by providing information as analysis is undertaken and warning the user when difficulties can be foreseen.

FREEWAY CAPACITY AND SPEED-FLOW RELATIONSHIPS

Speed-flow relationships are fundamental to understanding traffic flow phenomena on freeways. These relationships have been studied for more than 60 years, and it is important to note the significant changes that have taken place over this period with anticipation of what these relationships will look like in the years ahead.

Early Speed-Flow Relationships

In the 1930s Greenshields studied traffic flow relationships and proposed that speed was a linear function of density that results in parabola-shaped relationships between flow and density as well as between speed and flow. These stream flow relationships are shown in Figure 4.

The speed-density relationship is shown in the upper left-hand corner of the illustration, with speed shown as a linear function of density. Although this relationship is of greater interest to the theorist than the practitioner, it defines the other two relationships as shown in the illustration.

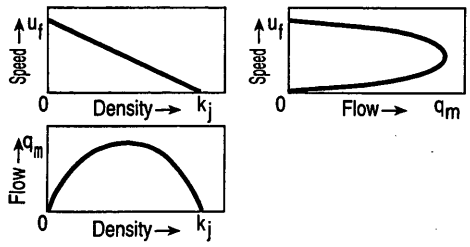


FIGURE 4 Basic stream flow diagrams.

The flow-density relationship is shown in the lower left-hand corner of the illustration, with flow being shown as a function of density. Theorists and practitioners who are engaged in freeway control often base their control algorithms on this relationship. Density (or occupancy) is used as the control variable and flow as the measure of productivity. It is important to note that maximum flow (or capacity) is obtained at a midrange value of density: density values lower than this midrange are indicative of free-flow conditions and higher levels of service; density values higher than this midrange are indicative of congested or oversaturated conditions. The portion of this illustration under high-density conditions reveals that this flow regime is characterized by both poorer level of service and lower productivity.

The speed-flow relationship is shown in the upper right-hand corner of the illustration, with speed as a function of flow. This relationship is most important to the practitioner whether involved in planning, design, or operations. Like the flow-density relationship, there are two portions of the curve (free flow and congested flow), but in this case they are separated on the basis of a midrange speed value. This relationship is used in three ways. The upper portion of the curve is used to estimate level of service for the users under free-flow conditions, the right maximum values are used to estimate capacity, and the bottom portion of the curve is used to calculate resulting upstream congestion patterns through shockwave analysis.

1950 HCM Speed-Flow Relationships and Predicted Capacities

A major accomplishment in speed-flow relationships occurred in 1950 with the publication of the first HCM. This HCM contained procedures for calculating capacity (then called possible capacity) and what was then called practical capacity [occurring at an approximate volume/capacity ratio (*v/c*) of 0.75]. Although some speed-flow relationships were included, the methodology was directed at the calculation of possible and practical capacity. It is interesting to note that the capacity under ideal conditions was proposed to be 2,000 passenger cars per hour per lane (pcphpl) and that this value was changed only in the 1990s with the revised HCM chapters. The highest maximum observed hourly volumes were reported for such facilities as US-1 near the Newark (New Jersey) International Airport [2,275 vehicles per hour per lane (vphpl)], Grand Central Parkway in New York (2,194 vphpl), and the Outer Drive in Chicago (1,958 vphpl).

1965 HCM Speed-Flow Relationships and Predicted Capacities

One of the key advancements with the 1965 HCM was the introduction of level of service based on the *v/c* integrated with the determination of capacity. The proposed speed-flow relationships in the 1965 HCM are shown in Figure 5. Note that the horizontal scale has been normalized and is shown as the *v/c*. The procedure was to calculate the capacity, then calculate *v/c*, and determine the operating speed from this illustration. The value for the capacity under ideal conditions continued to be 2,000 pcphpl. More than 10 sites were reported to have hourly lane flows of more than 2,000 vphpl, and the highest maximum observed hourly volumes were reported for such facilities as Lake Shore Drive in Chicago (2,236 vphpl), US-99 in Seattle (2,189 vphpl), and Hollywood Freeway in Los Angeles (2,190 vphpl).

1985 HCM Speed-Flow Relationships and Predicted Capacities

The 1985 HCM continued the level-of-service concept integrated with the determination of capacity. Greater emphasis was given to density, and it was used as the traffic parameter on which level of service was determined. The proposed speed-flow relationships in the 1985 HCM are shown in Figure 6. The value for capacity under ideal conditions continued to be 2,000 pcphpl. The shape of the upper portion of the speed-flow curve continued to be parabolic, and capacity was expected to occur at speeds of about 35 mph. Many sites were reported to have hourly lane flows of more than 2,000 vphpl.

Multiregime Speed-Flow Relationships

During the period between the publication of the 1965 and the 1985 HCMs, several researchers proposed and developed multiregime

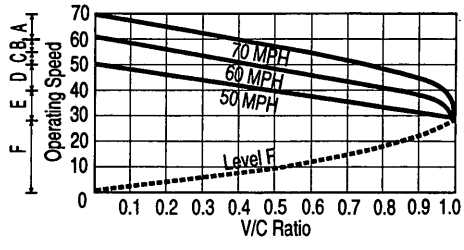


FIGURE 5 1965 HCM speed-flow relationships.

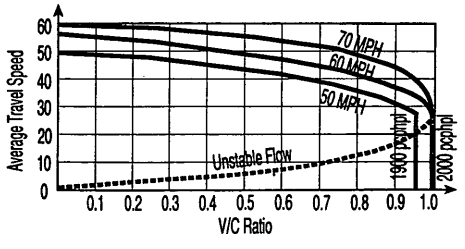


FIGURE 6 1985 HCM speed-flow relationships.

speed-flow relationships. An example of such a relationship is given in Figure 7. The concept was that there was a discontinuity between free-flow conditions and congested-flow conditions, and models that could be used to represent one condition were not necessarily the best for the other.

1994 HCM Speed-Flow Relationships and Predicted Capacities

The chapter on freeway capacity and level of service has been revised and was published in 1994 with the other updated chapters. The level-of-service concept is continued, with emphasis on density, but two significant changes have occurred. First, the capacity under ideal conditions has been increased by 10 percent, from 2,000 to 2,200 pcphpl. Improvements in the vehicle fleet and driver capabilities are thought to be the reason for the increase. The other significant change is the shape of the upper portion of the speed-flow relationship as well as the beginning recognition of a multiregime relationship. The proposed speed-flow relationships in the revised freeway chapter are shown in Figure 8. It was common to find sites with hourly flows of more than 2,000 vphpl, and a number of sites were reported to have hourly flows of more than 2,200 vphpl.

Current NCHRP Project 3-45 on Speed-Flow Relationships and Capacity Predictions

A new research project specifically directed at developing speed-flow relationships and estimating freeway capacity for the anticipated year 2000 HCM is under way as part of NCHRP Project 3-45. The hypothesis to be tested is that the speed-flow relationship is a multiregime relationship consisting of three regimes, or segments, as shown in Figure 9. This hypothesis is based of the work

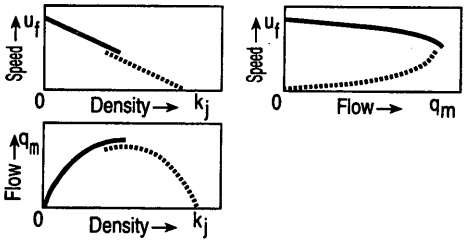


FIGURE 7 Multiregime speed-flow relationships.

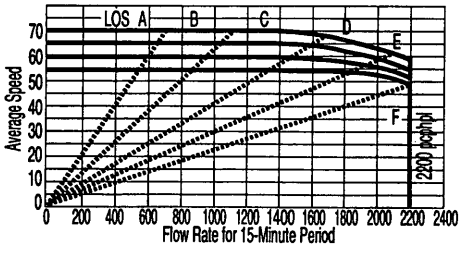


FIGURE 8 1994 HCM speed-flow relationships.

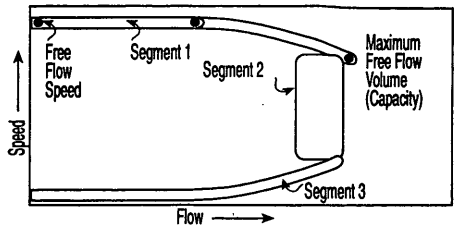


FIGURE 9 NCHRP Project 3-45 hypothesized speed-flow relationships.

of many individuals, including Hall, Banks, Roess, Reilly, and Urbanik.

The upper portion of the speed-flow relationship is assumed to be essentially horizontal up to a flow of about 1,400 vphpl and then to continue to capacity with only a slight decrease in speed on the order of 5 to 10 mph. Capacities would be observed at bottlenecks just before upstream congestion commences. Because of the small decrease in speeds over this complete range, the use of density (percentage occupancy) and *v/c* for identifying levels of service becomes more apparent.

The second flow regime would be represented by an almost vertical line in which near-capacity flows would be recorded but with speeds ranging from approximately 50 to 30 mph. Density would increase in the range from approximately 45 to 70 vehicles per mile per lane. There is some evidence that the capacity is slightly higher—on the order of 2 to 5 percent—before congestion occurs just upstream of the bottleneck. This would mean that if free-flow conditions could be maintained and the bottleneck capacity fully used, a slightly higher capacity could be obtained.

The third flow regime is represented by the lower portion of the speed-flow relationship, which occurs under congested flow conditions. Less is known about this flow regime, yet it strongly affects the degree of upstream congestion. If, for example, the speed at a flow of 1,000 vphpl was 10 mph instead of the indicated 8 mph, the resulting density would be 100 instead of 125 vehicles per mile per lane. Thus, the travel time rate within the congested portion of the freeway would be decreased by 1.5 min/mi of travel, but the length of the congested portion would be 20 percent more.

Future Speed-Flow Relationships and Capacity Predictions

In the short term, as the vehicle fleet and drivers' capabilities continue to improve and as in-vehicle driver aids increase, it would not be surprising if the capacity value under ideal conditions were to grow to 2,400 pcphpl by the time the year 2000 HCM is published. The shape of the upper portion of the speed-flow curve should remain unchanged, with the curve extending to the right to the higher capacity value. Additional and more precisely measured traffic flow characteristics will probably show that capacity flow prior to congestion will be about 5 percent higher than when congestion forms. The bottom portion of the speed-flow relationship will be much better defined for use in predicting shock-wave phenomena and resulting congestion patterns.

A possible speed-flow relationship by the year 2000 is superimposed on typical speed-flow relationships as contained in the 1965, 1985, and 1994 HCMs (Figure 10). Over time, the upper portions

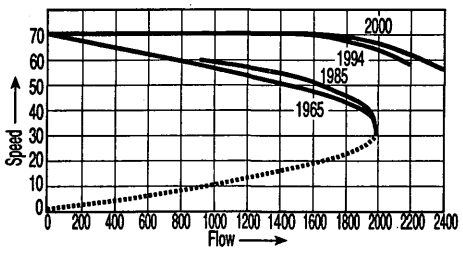


FIGURE 10 Past-present-future speed-flow relationships.

of the relationships have continuously moved upward in the diagram and extended farther to the right.

It is more difficult to predict speed-flow relationships beyond the year 2000. With fully automated vehicle control systems, higher speed-flow relationships could be expected along with the corresponding higher capacity values. The limitations will be determined by the fail-safe system. Capacities in excess of 2,400 vphpl (average time headways of 1.5 sec/veh) will be required to exceed anticipated freeway capacities without fully automated vehicle control systems.

FREEWAY SIMULATION MODELS

Professional engineers and researchers are accepting and recognizing more readily the important role that simulation models can have in assessing problem areas, generating potential solution approaches, and evaluating the traffic and environmental impacts of implementing advanced traffic management systems (ATMS) and advanced traffic information systems (ATIS). The applications of simulation models have expanded because of the development of improved and more comprehensive models, models that can better represent real-life situations, with user-friendly input and output interfaces and much improved computer capabilities. A new frontier in the use of simulation models is in the area of on-line surveillance, control, and information systems, which will be given attention later in this paper when freeway entry control systems are described.

Three classes of simulation models for freeway corridor-type traffic operating environments will be described: the FREQ model, the INTEGRATION model, and a newly developed traffic-planning model.

FREQ Simulation Model

The FREQ model is a deterministic macroscopic model that includes simulation and optimization submodels and permits a time-stream evaluation of freeway corridor performance under design or control traffic management strategies. An illustration of this time-stream evaluation is shown in Figure 11. Its development and enhancements have evolved over the past 20 years, and the current versions are referred to as FREQ10 and FREQ11. It operates on a personal computer, and one of its strengths is its input and output user interface with diagnostics when problems are encountered. FREQ on-hands computer laboratory workshops have been held in a number of states, most recently California, Texas, and Washington.

As shown in Figure 11, the inputs to the model are the time slice traffic counts and facility design features. A synthetic origin-

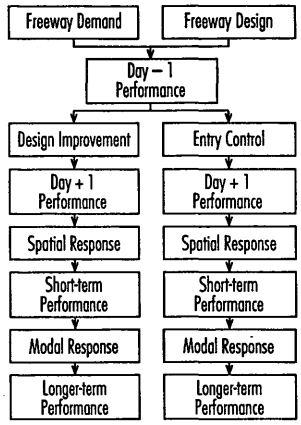


FIGURE 11 Freeway simulation model.

destination (O-D) procedure is incorporated into the model that converts the traffic counts into time slice O-D tables. The user provides subsection capacities based on the facility design features. The simulation model is employed to predict traffic performance in the freeway corridor for the time before implementation of a traffic management strategy (Day – 1). Design improvements such as added HOV lanes and mixed-flow lanes can be incorporated into the model, and the simulation model will predict the effect of these improvements without traveler responses (Day + 1). Optimized entry control strategies can be generated either with or without the design improvements, and the simulation model will predict the effect of these improvements without traveler responses (Day + 1).

The FREQ model also includes spatial and modal response submodels that predict and reassign users to alternative routes or multioccupancy vehicles. The simulation model can be used to predict the short- and long-term traffic performance of implementing the design/control improvement with spatial or modal responses.

INTEGRATION Simulation Model

The INTEGRATION model was developed in the 1980s and has been enhanced and expanded since then. The original work was done at Waterloo University; work has been done more recently at Queens University by Van Aerde. This development has been sponsored in part by the Ontario Ministry of Transportation, and its chief applications have been to the Trav-Tek Project in Orlando, Florida, and to the SMART corridor in Los Angeles. It is currently being used by the intelligent vehicle-highway system (IVHS) architecture contractors in assessing IVHS strategies.

The model is a deterministic macroscopic model explicitly developed for IVHS applications in freeway corridors; it has the following features:

- Models freeways and arterials simultaneously,
- Uses individual vehicles with self-assignment capabilities,
- Includes five vehicle types with varying levels of information,
- Optimizes traffic signals, and
- Simulates multiple incident and construction scenarios.

An auxiliary program is available called QUEENSOD that will generate time slice traffic demand O-D tables based on traf-

fic counts. The O-D tables, physical network, and intersection and ramp signal control serve as inputs to the model. The traffic assignment routines and the traffic performance predicted by the simulation model operate concurrently to load the traffic onto the network and thus predict the traffic performance in the freeway corridor.

ATMS strategies such as on-freeway HOV lanes, incident management, entry control, and intersection signal control can be investigated. ATIS strategies such as highway advisory radio (HAR), changeable message signs (CMSs), and in-vehicle information systems can be simulated. Up to five vehicle types can be simulated with varying levels of information. Various vehicle types can obtain current travel time information either precisely or with noise at every node in the corridor, at HAR/CMS locations, or not at all.

The output to the model is very comprehensive and includes off-line geographic maps of the entire corridor or subparts with superimposed input/output data and with the option of on-line vehicle animation on the computer screen for any portion of the freeway corridor. Vehicles can be color-coded to represent various situations: their existence in free- or congested-flow conditions, or perhaps their status as HOV vehicles or vehicles receiving updated traffic travel time information.

Traffic-Planning Simulation Model

The traffic-planning simulation model integrates a regional planning model with traffic simulation models (FREQ and TRANSYT) in order to predict traffic and environmental impacts. The use of the regional planning model provides a broad view of travel in the region for current and future scenarios. The use of the traffic simulation models predicts freeway and arterial traffic performance in specific freeway corridors on the basis of the regional planning model's prediction of demand.

This new type of simulation model was developed by JHK & Associates for the Transportation Systems Center and the California Department of Transportation. It is being tested and evaluated on the I-880 freeway corridor in the San Francisco Bay Area.

A simplified flow chart of the traffic-planning model is shown in Figure 12. The planning submodel portion is shown in the five boxes in the upper-left portion of the figure, and the traffic submodel portion is shown in the five boxes in the upper-right portion of the figure. The outputs of the model are shown in the two boxes at the bottom of the figure and the box on the far right of the figure.

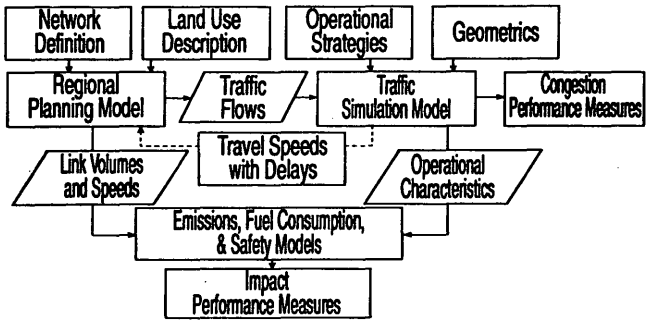


FIGURE 12 Traffic-planning model integration.

The input to the regional planning model is the network definition and the land use description. The regional planning model estimates the traffic flows on each link of the freeway corridor. The traffic simulation model uses these traffic flow estimates with operational strategies and geometrics to predict travel speeds with delays. The travel speeds with delays from the traffic simulation model are compared with similar measures from the regional planning model, and if they are found to be significantly different, the regional planning model is rerun using adjusted travel speeds and new predicted traffic flows. This process is continued until equilibrium is reached between the two models. Once this occurs, the traffic congestion performance measures and the environmental impact performance measures can be predicted.

FREEWAY ENTRY CONTROL

Freeway entry control is one of the major strategies within the ATMS program. Referring to Figure 3, which identifies possible solution approaches when demand exceeds capacity, freeway entry control is one strategy that incorporates almost all of the possible solution approaches. On the capacity side,

- HOV bypass lanes can be provided for the on-ramps,
- Geometric improvements can be made a part of the entry control strategy,
- The entry control system can be designed for recurring congestion as well as for incident congestion, and
- If free-flow conditions can be maintained, slight increases in freeway capacity can be expected.

On the demand side, freeway entry control can result in a variety of traveler responses. Excess demand may be

- Diverted to underused parallel facilities,
- Spread to pre- and post-congested time periods,
- Shifted to carpools, vanpools, and buses due to priority treatment of higher-occupancy vehicles, and
- Reduced by trip consolidation, altered destinations, and trip reductions.

Although implemented freeway entry control systems have been successful in many locations in North America and abroad, some such systems have not been as successful. Lessons have been learned over the years, and key features for the more successful projects have been identified:

- Recognize the corridor solution approach,
- Institute partnerships with government agencies,
- Enhance entry control systems of ATMS/ATIS,
- Implement comprehensive surveillance systems first,
- Consider spot improvements on freeway and arterials,
- Select the most promising implementation,
- Take advantage of public information and marketing,
- Monitor implementation closely, and
- Fund operations and maintenance adequately.

Earlier Periods of Freeway Entry Control

Freeway and tunnel entry control began in the early 1960s with tunnel entry control systems in some of the New York tunnels and

with manual police officer control on the Red Feather Expressway in St. Louis. Soon thereafter, surveillance and control projects were initiated in Chicago and Detroit, and the first automatic entry control system was implemented in Chicago in 1963.

Freeway entry control systems grew in terms of both number of geographic areas and number of ramps controlled in the geographic areas. Two types of freeway control systems were implemented: local traffic responsive control and multiramp time-of-day control based on historical traffic information.

TRB's Committee on Freeway Operations was started as a task force in the late 1950s and has continued to serve as a national central point for discussions and publications related to freeway operations. One of its continuing activities is to provide a freeway operations project summary periodically. The most recent summary was published in 1991 and indicated that more than 2,000 ramps were being metered in more than 30 geographic locations.

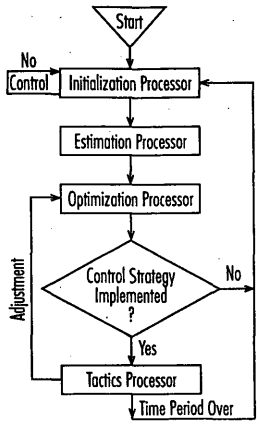


FIGURE 13 Overview of systemwide traffic responsive control.

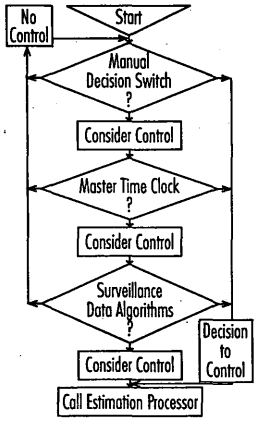


FIGURE 14 Initialization processor.

Working Toward Coordinated Traffic Responsive Entry Control

There has been a continuous improvement in freeway entry control systems in terms of educational programs, enforcement procedures, priority entry control, use of detectors, and control strategies. One of the more technically challenging efforts has been to work toward coordinated traffic responsive entry control.

When an individual controlled ramp reaches its minimum metering rate or its maximum permitted queue length, it cannot further restrict entry onto the freeway. Without some form of coordinated traffic responsive control, freeway congestion will occur but further restriction to entry to the freeway from upstream ramps will not occur until sometime later, when the queues extend through the next upstream ramp.

The purpose of coordinated traffic responsive entry control is to reduce the metering rates at upstream ramps as soon as any ramp reaches either its minimum metering rate or its maximum permitted queue length. In this way action is taken immediately and the entry control strategy begins to take on a systems approach. Very recently, several researchers have proposed a more systemwide traffic responsive freeway entry control strategy approach. "Helper" ramp implementations include those in Chicago, Denver, Minneapolis-St. Paul, Seattle, and Europe; examples of on-line simulation models include the Cornerstone, European, and Minneapolis-St. Paul approaches.

Systemwide Traffic Responsive Freeway Entry Control

One approach to systemwide traffic responsive freeway entry control is to integrate a freeway optimization and simulation model into the on-line entry control system. The approach presented in this paper incorporates four processors that will be described in the following; their interrelationships are shown in Figure 13.

The control strategy begins with the first processor, which is called the initialization processor; a flow chart of this processor is shown in Figure 14. One of three decisions is made: the choices are to (a) not control and therefore not call the next processor, (b) consider control if a feasible control strategy can be generated, or (c) require an implementation of control. The selection is based on the system manager's instructions, the system time clock, and traffic performance.

If the decision is made to consider control or to implement control, the second processor, referred to as the estimation processor, is called; its flow chart is shown in Figure 15. The purpose of this processor is to assemble the necessary input data for the freeway optimization and simulation model. The input data consist of an array of subsection capacities along the freeway and a freeway traffic demand O-D table for the next period. The nonincident capacity of each freeway subsection is stored in the computer memory but

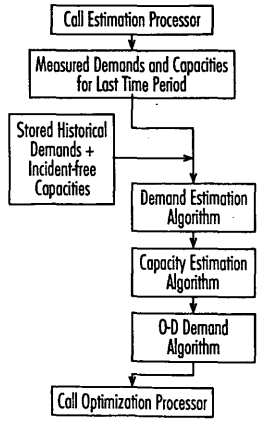


FIGURE 15 Estimation processor.

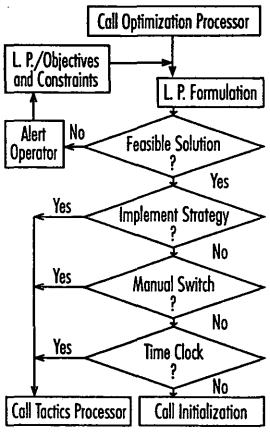


FIGURE 16 Optimization processor.

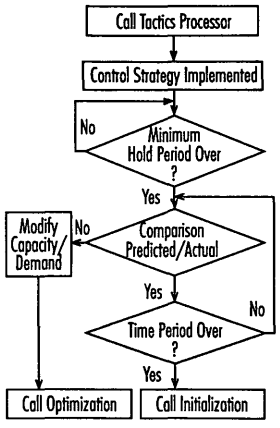


FIGURE 17 Tactics processor.

would be replaced by on-line estimates of capacity if they were found to be lower. The traffic demand O-D table would be obtained on the basis of entry and exit flow measurements obtained during the previous time period and historical data. Demand predictions would be made for the next time period and the O-D demand table developed through synthetic O-D techniques. The demands and capacities would be transferred to the next processor.

The next processor is called the optimization processor, which determines the systemwide traffic responsive control strategy based on the O-D demand table and the subsection capacity array. A flow chart of the optimization processor is shown in Figure 16. A linear program would be employed in which the objective function would have been preselected and the constraints would include queue limits, minimum and maximum permitted metering rates, and desired maximum *v/c*'s. The operator would be informed of the resulting control plan and either could modify the objectives and constraints or could modify any specific element of the control plan.

The fourth and final processor is referred to as the tactics processor, which is called once the control plan for the next period is implemented. A flow chart of the tactics processor is given in Figure 17. A short-term new set of demands, capacities, and system performance is measured and compared with predicted values. If there are no significant differences, the control plan is maintained, and another short-term new set of demands, capacities, and system performance is measured and compared, and the process is repeated. If there is a significant difference between any of the new measurements and predicted values, the optimization processor is recalled, a new control plan is developed, and the tactics processor is called again. This process is continued until the end of the period, at which time the initialization processor is recalled and the four-processor procedure is repeated.

Placing the traffic optimization and simulation model on-line using predicted demands and capacities has many advantages. First, the entry control strategy is determined on the basis of *anticipated*—not historical or previous—traffic measurements and is optimized on a system basis. Second, anticipated travel time and ramp delay information is available for ATIS-types of information systems. Third, unusual traffic conditions are identified early on in terms of modifying traffic signal systems and alerting traffic incident management activities. As traffic conditions change, the traffic management strategy is reassessed and a more appropriate one is selected.

SOME FINAL OBSERVATIONS: "BRIDGES BETWEEN"

This presentation has attempted to propose certain fundamentals of traffic management that have been learned from research and practice. Then the freeway and freeway corridor traffic operating environments were selected as an example with special emphasis given to TRB committee contributions, capacity and speed-flow relationships, simulation models, and traffic management strategies.

The discussion closes with some final observations that are presented in the context of "bridges between." Traffic management is a complicated issue and requires many bridges between people, organizations, disciplines, and approaches. No one person, organization, discipline, or approach can solve the problem, but by working together we can make a difference and give the traveling public the very best use of the urban road system. Let me close with a few examples.

Educators and engineering professionals need to work together to encourage the very best candidates to enter the transportation educational process. Educational programs must be designed carefully to provide the fundamental education for those who wish to enter professional service. Education is a continuing process and requires continuing interaction between the educational communities and the profession through seminars, workshops, short courses, and more formal educational opportunities.

Bridges are needed between theory and practice, research and on-road realities, and analytical tools and real-life applications. One without the other falls short of the needs.

System components and their integration into total systems must be well understood if traffic management is to be successful.

Traffic management requires the expert contributions of individuals from many disciplines: from marketing and planning to computer software and hardware, from sociology and economics to

operations research and engineering. All are needed, and many more, if traffic management is to be successful.

There must be a balance between capacity improvement and demand management. Only by working together will these approaches lead to better traffic management.

Traffic management of the major urban road system is not just a city or county or state or federal responsibility. A partnership is required if traffic management is going to work. It is not just *my* system—it is *our* system working together.

The final bridge is from the past to the future. The past has given us the major urban road system that we have. In the recent past, and today, attention has turned toward managing it. The operational problems that were easier to solve have been addressed and many lessons have been learned. The more difficult problems have been left to the future, with increased constraints being placed on the solutions. Traffic management in the future will not be easy, and there will be many unforeseen difficulties. The challenge is there; working together, we can meet it.

PART 2

Traffic Flow and Capacity

Part 2
Foreword

The 25 papers in this volume are related by their focus on highway capacity, quality of service, traffic flow measurement, and traffic flow theory. However, the papers cover a wide range of problems reflecting the concerns of practitioners as well as theoreticians.

Issues of highway capacity are receiving considerable attention because of the research effort leading toward the next edition of the Highway Capacity Manual (circa 2000). The initial group of papers in this Record examines highway capacity on uninterrupted flow facilities as it relates to rural highways, weaving sections, and freeways. Other papers deal with traffic control parameter estimation for actuated signal control, capacity of exclusive turn lanes, saturation headway at stop-controlled intersections, and roundabout capacity.

Traffic flow theory, modeling, and control applications are discussed in papers on traffic pattern estimation and route guidance, estimation of traffic flow variables for intelligent transportation system (ITS) applications, and traffic flow theories and simulation models.

Whether the reader is a traffic engineer trying to determine the capacity and level of service of freeways or a traffic flow theoretician pondering the vagaries of traffic flow equations as they relate to IVHS applications, the papers in this Record should be both interesting and informative.

Level of Service of Two-Lane Rural Highways with Low Design Speeds

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

The parameters and approach to the evaluation of levels of service for two-lane highways were changed substantially from the 1965 *Highway Capacity Manual* (HCM) to the 1985 HCM. A principal change was the introduction of percentage time delay as a parameter used to describe service quality. Another significant change was the elimination of an explicit and fully defined methodology to analyze two-lane highways with design speeds lower than 96 km/hr (60 mph). Although the 1985 HCM can be applied to highways with low design speeds, the procedure is acknowledged to be incomplete, at least as far as speed is concerned. Alternative methods are proposed to analyze the level of service for two-lane highways with design speeds of 80 km/hr (50 mph); the methods are based on relationships among speed, volume, density, and percentage time delay. The relationships were developed with the aid of the TWOPAS computer model, which is the same model used for the development of the basic relationships used in the HCM. In conclusion, a strategy for future development is proposed.

The parameters and approach to the evaluation of levels of service (LOS) for two-lane highways were changed substantially from the 1965 to the 1985 *Highway Capacity Manual* (HCM) (*1,2*). A principal change was the introduction of percentage time delay as a parameter used to describe service quality. Another significant change was the elimination of an explicit and fully defined methodology to analyze two-lane highways with design speeds lower than 96 kph (60 mph). According to AASHTO (*3*), the design speed "is the maximum safe speed that can be maintained over a specified section of highway when conditions are so favorable that the design features of the highway govern." Although the 1985 HCM can be applied to highways with low design speeds, the procedure is acknowledged to be incomplete, at least as far as speed is concerned.

It was also found that many highways with low design speeds, which had been evaluated using the 1965 HCM, needed to be reclassified in some cases to much higher LOS categories when the 1985 HCM procedure was applied. This discovery led to concern over the lack of consistency between the 1965 and 1985 methods when applied to low-design-speed highways and raised doubts about whether the new procedure is adequate for such facilities.

Another question was whether, for low-design-speed highways, the 1985 HCM procedure is true to the LOS concept presented in the 1985 HCM, which defines LOS as a measure describing operational conditions within a traffic stream "in terms of such factors as speed and travel time, freedom to maneuver, traffic interruptions, comfort and convenience, and safety."

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

As a result of these changes and concerns, the California Department of Transportation (Caltrans) initiated a research project to investigate LOS for two-lane highways with design speeds lower than 96 km/hr (60 mph). The focus was on roads with design speeds of 80 km/hr (50 mph).

The first goal of the study was to review different ways in which the LOS for two-lane highways can be defined, explore the implications of these alternatives, and use these findings to scope appropriate later study. The second goal was to conduct an empirical investigation of traffic on selected two-lane highways with low design speeds in order to extend the 1985 HCM methodology.

The study included the following tasks:

- Field data characterizing traffic operations on selected sections of low-design-speed state highways in Northern California were collected.
- Two microscopic simulation models, TWOPAS and TRARR, were calibrated and compared in terms of their abilities to reproduce the traffic conditions observed in the field. Both models generally performed well, but TWOPAS matched the field data more closely and was selected as the analysis tool for this study. The results of this model comparison are documented in a separate paper (*4*).
- The existing 1985 HCM methodology and its usage were critiqued, with a discussion of alternative methodologies and desirable properties for such methodologies.
- Several methodological alternatives were evaluated in detail, including the current HCM general terrain methodology, which uses percentage time delay to define LOS for two-lane, two-way highways.

The complete study is documented in a final report (*5*). The focus of this paper is the evaluation of methodological alternatives for defining the LOS for two-lane highways with 80-km/hr (50-mph) design speeds. These alternatives are

- Percentage time delay as basic parameter,
- Density as basic parameter,
- Functional classification of road as basis,
- Limitation on LOS at low design speeds, and
- Combined percentage time delay–density as basis.

These options will be discussed in terms of possible parameters (where not already specified), possible boundary values between LOS, and their implications regarding high LOS. To discuss the advantages and shortcomings of the various options, fundamental relationships among the different variables, developed with the aid of the TWOPAS computer model, are first presented. A discussion of a possible strategy for future development follows.

FUNDAMENTAL RELATIONSHIPS FOR ROADS WITH 80-KM/HR DESIGN SPEEDS

Fundamental traffic flow relationships for highways with 80-km/hr (50-mph) design speeds were produced with the TWOPAS model, which is the same model used to produce the values used in the 1985 HCM.

The 1985 HCM values were produced using a tangent section of highway. Values for traffic variables were obtained by varying the grade. For roads with low design speeds, the horizontal alignment is often the factor determining those lower design speeds. It was therefore not considered representative of field conditions to use a tangent section of highway to produce the required values.

Instead, two actual sections of highway with design speeds of 80 km/hr (50 mph)—or, more specifically, an average highway speed (AHS) of 80 km/hr (50 mph)—were used for this purpose. The AHS is the weighted average of the design speeds within a highway section. The results should be generally applicable to other highways with 80-km/hr (50-mph) design speeds and similar geometric properties.

The two sections have geometric characteristics that correspond to roadways in rolling and level terrain, respectively. No passing was allowed on the rolling terrain, whereas passing was allowed over 6 percent of the level terrain.

Although it is realistic to impose no-passing zones on roads that are designed for passing, it is not realistic to do the opposite. Since the model produced good results with the actual road sections, where very little passing is allowed, it was considered appropriate only to impose 100 percent no-passing on the level terrain. The types of road for which relationships were produced were, therefore,

- Rolling terrain, 100 percent no-passing;
- Level terrain, 95 percent no-passing; and
- Level terrain, 100 percent no-passing.

The directional split was 50/50. Both road sections had lane widths of 3.4 m (11 ft) and shoulder widths of 0.6 m (2 ft).

The TWOPAS simulation model was used to establish the following relationships for these road sections for a vehicle population of all passenger cars: speed-volume, density-volume, density-speed, percentage time delay–volume, percentage time delay–speed, and percentage time delay–density.

The fundamental relationships are presented in Figures 1 through 6 (NPZ = no-passing zone). In each case the simulation was run for 1 hr. Each density value was calculated as flow rate divided by speed.

The following items related to the fundamental relationships are noteworthy:

- The speed-volume relationship (Figure 1) has the same overall shape as the relationship presented in the HCM, but, as expected, speeds are lower, especially in the case of rolling terrain. The model could not produce the maximum flow rates of 2,800 passenger cars per hour (pcph) used in the HCM. A high value of approximately 2,360 pcph was attained. However, this should not necessarily be viewed as the capacity of these roads, since the values obtained in the high ranges of flow were not verified in the field. The model was validated for field flow rates of 500 to 800 vehicles per hour in both directions.
- The percentage time delay–volume relationship (Figure 2) also has the same overall shape presented in the 1985 HCM. However, the fact that the percentage time delay values for rolling terrain are about 8 percent lower than for level terrain, for the same flow rate, is unexpected. This may indicate that the model does not replicate passing behavior adequately to produce accurate results for percentage time delay. Another possible explanation is that the performance of vehicles is more likely to be constrained on the rolling terrain and may therefore not catch up to the leading vehicles. Experimentation with the model showed that rolling terrain continued to yield lower percentage time delay values than level terrain, even when increased passing opportunities were provided. However, in a separate experiment on a tangent section, the percentage time delay was higher for rolling terrain. This phenomenon warrants further investigation. Because of the relative inaccuracy of the

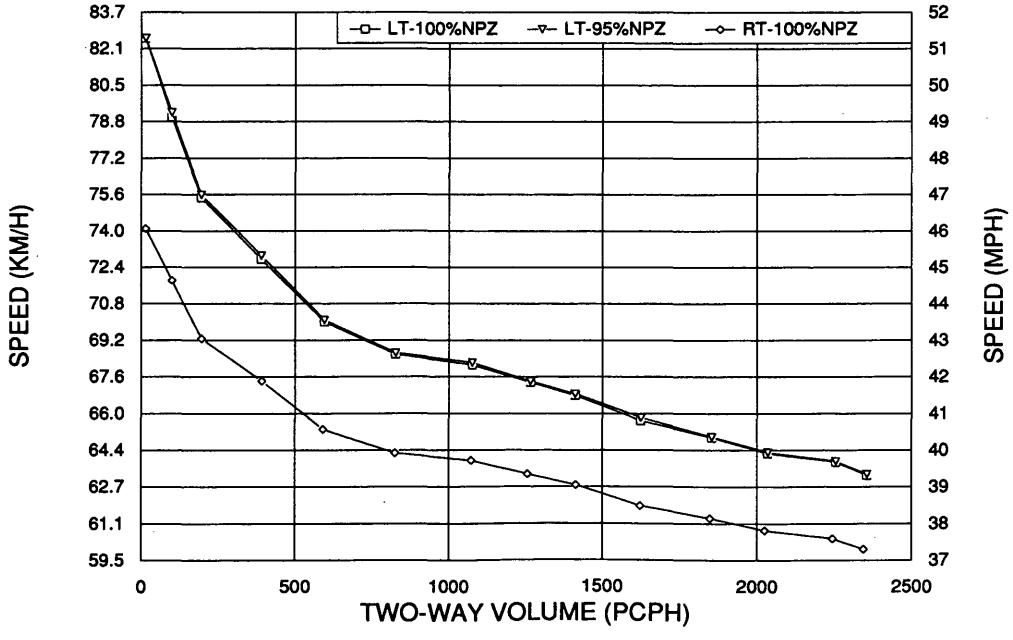


FIGURE 1 Speed-volume relationship, 80-km/hr design speed.

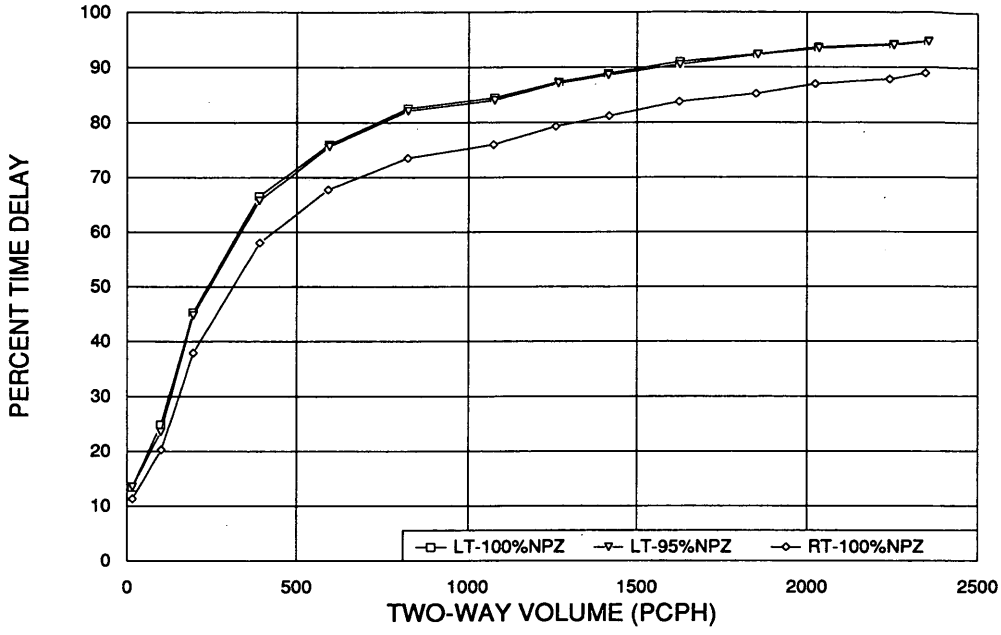


FIGURE 2 Percentage time delay–volume relationship, 80-km/hr design speed.

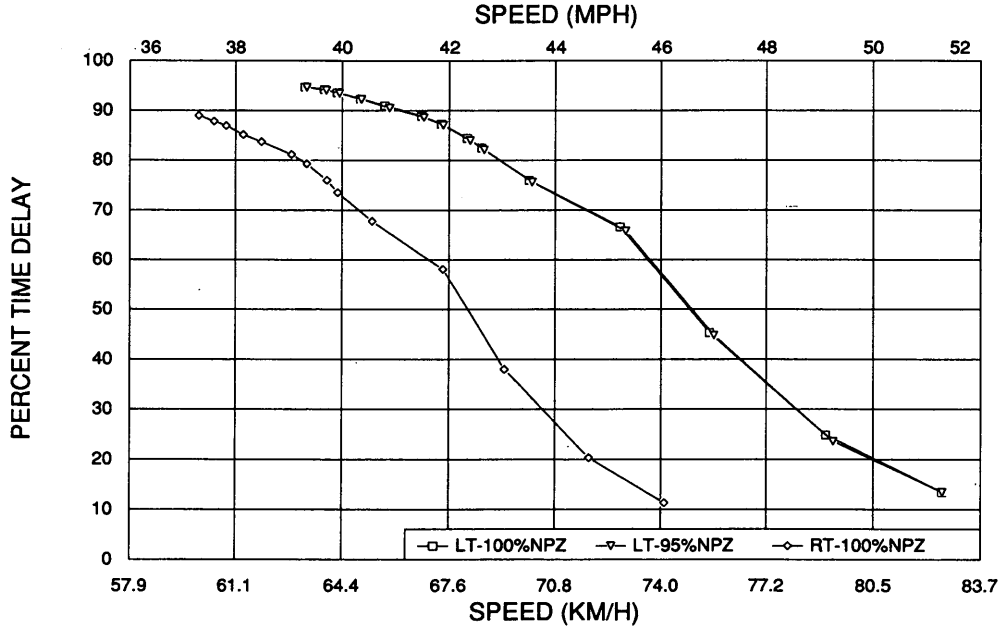


FIGURE 3 Percentage time delay–speed relationship, 80-km/hr design speed.

model predictions of percentage time delay (as found during the model calibration and validation stage) and the unexpected results discussed here, the accuracy of percentage time delay values should be questioned, even though the general shape of the percentage time delay–volume relationship appears to be reasonable.

- The percentage time delay–speed relationship (Figure 3) appears to be reasonable. It is also evident that increasing freedom to pass does not lead to substantial increases in average speed. It is noteworthy that there is a greater difference in speed due to changes in terrain than is exhibited for highways with 96-km/hr (60-mph) design speeds (*2*).

METHODOLOGICAL ALTERNATIVES FOR LOS ANALYSIS

In this section, a number of methodological alternatives for LOS analysis on roads with 80-km/hr (50-mph) design speeds are presented and discussed.

Percentage Time Delay as Basic Parameter

To be as consistent as possible with currently accepted practice, it could be argued that percentage time delay should remain the

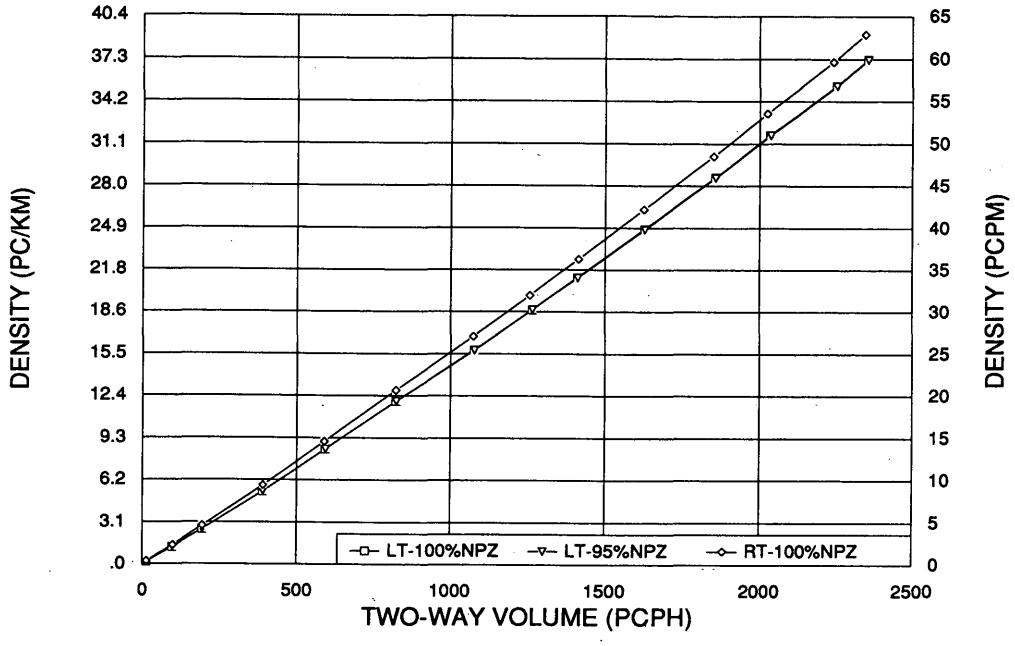


FIGURE 4 Density-volume relationship, 80-km/hr design speed.

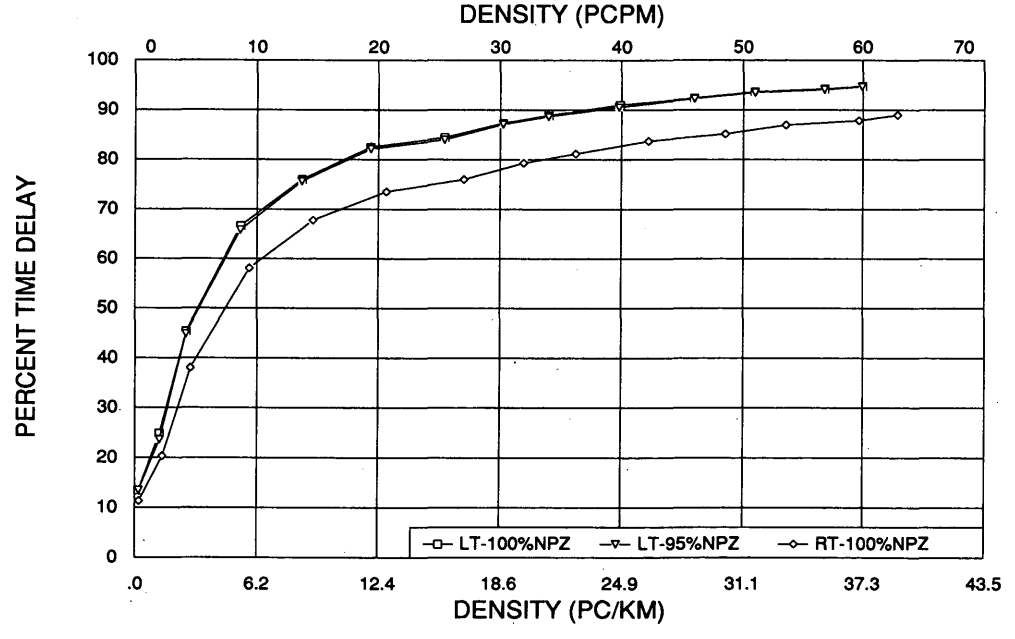


FIGURE 5 Percentage time delay–density relationship, 80-km/hr design speed.

primary parameter for general terrain applications for roads with low design speeds. Using the fundamental relationships depicted in Figures 2 and 3, the values in Table 1 were derived. The percentage time delay boundaries for LOS are identical to those in the 1985 HCM. The density values were calculated from the volume and speed.

It should be recalled that the accuracy of these percentage time delay values are questionable. A comparison of these results with the corresponding values in the 1985 HCM for 95 and 100 percent no-passing zones indicates that attainable flow rates for the 80-km/hr (50-mph) design speed are higher than those for the 96-km/hr (60-mph) speed. However, it is not necessarily only the 96-km/hr flow rates that should be questioned. The HCM gives LOS A service flow rates of 112 and 84 pcph, respectively, for rolling terrain with 80 and 100 percent no-passing zones. When adjusted for lane and shoulder width, the flow rates become 84 and 63 pcph, which yield densities of 0.9 and 0.7 passenger cars per kilometer (pc/km) [1.5 and 1.1 passenger cars per mile (pcpm)] at a speed of 91 km/hr (57 mph). The corresponding average headways are 42 and 52 sec. Perhaps it should be questioned whether such low

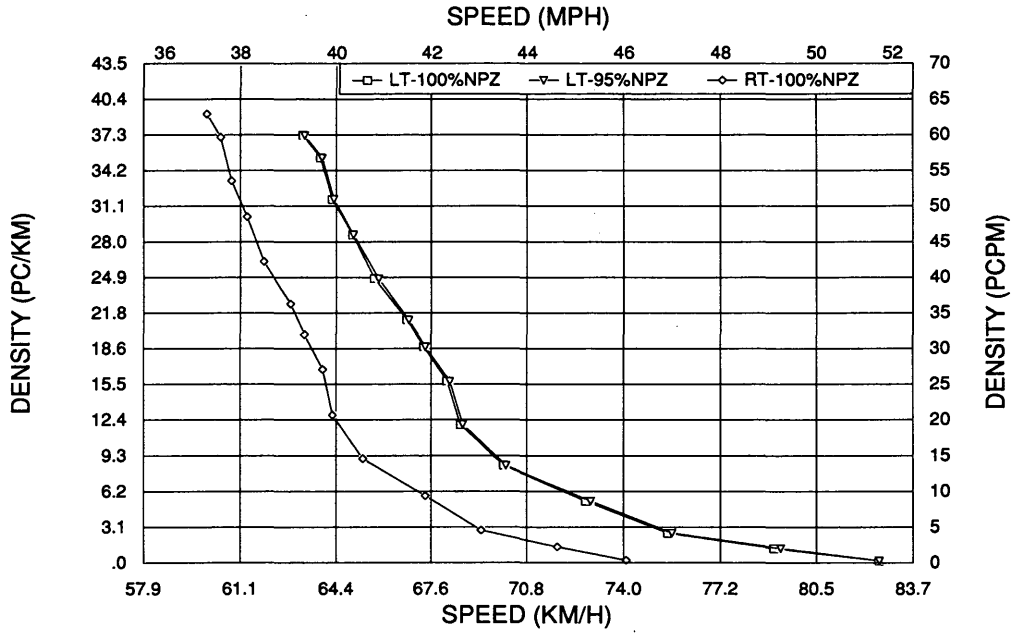


FIGURE 6 Density-speed relationship, 80-km/hr design speed.

densities and large average headways are indeed necessary to satisfy the operational standard of LOS A.

To make a completely correct comparison, an adjustment should be made to the 80-km/hr (50-mph) values to account for the narrower shoulders and lane widths. The model does not, however, explicitly take into account the lane and shoulder widths; therefore, no insights could be gained from applying the model in this respect.

Density as Basic Parameter

One of the advantages of using density as a parameter is that comparisons of LOS with other types of facilities become easier. This can be an important consideration in planning or when conducting a congestion management program. To accomplish this end, boundaries between the LOS should be established in such a way that the same quality of service is experienced at a given LOS, regardless of facility type.

The boundary values for a density-based LOS definition for two-lane highways should probably not be the same as those for multilane highways, since the operational characteristics of the facilities are very different. All else being equal, for comparable service quality, the densities on two-lane highways should be much lower than on freeways.

As a starting point, the density boundary values can be derived from the percentage time delay boundaries in the HCM. The percentage time delay boundaries and corresponding values for speeds and volumes are given in Table 2. Corresponding density values

TABLE 1 LOS Criteria[a] for 80-km/hr AHS Two-Lane Highway Sections, Percentage Time Delay

LOS	% Time Delay	Level Terrain						Rolling Terrain		
		95% No Passing Zones			100% No Passing Zones			100% No Passing Zones		
		Density (PC/KM)[d]	Speed (KM/H)[e]	Volume (PCPH)	Density (PC/KM)[d]	Speed (KM/H)[e]	Volume (PCPH)	Density (PC/KM)[d]	Speed (KM/H)[e]	Volume (PCPH)
A	≤30	≤1.7	≥77.8	130	≤1.6	≥77.8	120	≤2.1	≥70.1	150
B	≤45	≤2.7	≥75.2	200	≤2.6	≥75.0	190	≤3.9	≥68.2	265
C	≤60	≤4.6	≥73.3	340	≤4.5	≥73.1	330	≤6.4	≥66.6	430
D	≤75	≤8.3	≥69.9	580	≤8.3	≥69.9	575	≤15.3	≥63.7	975
E	>75	≤37.5	≥62.9[c]	2360[b]	≤37.5	≥62.9[c]	2360[b]	≤39.3	≥59.7[c]	2345[b]
F	100									

[a] For 3.4 m lane width and 0.6 m shoulder width
[b] Rough estimate of maximum volume
[c] Speed at maximum volume
[d] PCPM = PC/KM * 1.6
[e] MPH = KM/H * 0.625
Note: Expression q=ku may not hold exactly due to round-off in converting units.

TABLE 2 LOS Boundaries for Ideal Conditions, 1985 HCM Values

LOS	Percent Time Delay	Level Terrain			
		0% No Passing Zones			
		Speed (KM/H)[a]	V/C	Volume (PCPH)	Density[b] (PC/KM)
A	≤30	≥92.8	0.15	420	≤4.5
B	≤45	≥88.0	0.27	756	≤8.6
C	≤60	≥83.2	0.43	1204	≤14.5
D	≤75	≥80.0	0.64	1792	≤22.44
E	>75	≥72.0	1.00	2800	≤38.9
F	100	<72.0	-	-	>38.9

[a] MPH = KM/H * 0.625
[b] PCPM = PC/KM * 1.6

were calculated from speed and volume. Subsequently, these boundary values for density were used to find values for volume (from Figure 4) and percentage time delay (from Figure 5) for the 80-km/hr (50-mph) facilities. Speed values were calculated from volume and density. The results are presented in Table 3.

The volumes produced by this method for the various LOS are probably too high. This may suggest the need for an adjustment of the boundary values. It should be noted that the percentage time delay boundary values used in the 1985 HCM can be considered somewhat arbitrary since the categories coincide with 15 percent increments. Perhaps these boundaries could be selected so that the same operational quality of service is rendered as would be rendered by multilane facilities at the same LOS but at different values of density. A suggested revision in these boundaries is discussed later with the combined percentage time delay–density option.

Functional Classification of Road as Basis

Another option is to first define the function of the road (i.e., whether the road is to serve as an arterial, collector, local access, etc.). Arterials usually have higher design speeds than local access roads. The design speed reflects what is considered to be a safe comfortable speed consistent with the use and objectives of the facility. If vehicles operate at or near the design speed of the facility, then LOS A can be attained, whatever that design speed may be. Reductions in LOS can then be measured in terms of decreases in speed. If this notion is carried further, the reduction in speed could be equated with delay, which can be converted directly into economic impacts.

Boundary values could be established in terms of the percentage delay values. It appears that 5 percent increments in delay correspond approximately to 2 percent decreases in operating speed. Using 64 km/hr (40 mph) as the speed at capacity for facilities with 80-km/hr (50-mph) design speeds (which is close to the speed at the maximum flow rate produced by the model for level terrain), then LOS A could be defined as more than 77 km/hr (48 mph) and other LOS in equal descending increments. The corresponding values for volumes and percentage time delay were derived with the aid of Figures 1 and 3. The corresponding densities were calculated from the volumes and speeds. The results are given in Table 4.

Table 4 indicates that LOS A cannot be attained for rolling terrain. It should be noted that these boundaries are somewhat arbitrary (in the same way as the percentage time delay boundaries in the HCM) and that it would have been better to determine the speed at LOS E for a road with 0 percent no-passing zones. However, the 64-km/hr (40-mph) value is probably very close to the value that would be obtained for a road with 0 percent no-passing, since at flow rates approaching capacity, there are probably few passing opportunities.

Limitation on LOS for Low Design Speeds

Several options can be considered if high LOS are going to be limited for low design speed highways. One proposal is to limit LOS in the same way that it is limited for ramps in the 1985 HCM (Table 5-5 in the HCM). According to this proposal, the attainable LOS would be as follows:

TABLE 3 LOS Criteria[a] for 80-km/hr AHS Two-Lane Highway Sections, Densities Corresponding to HCM Percentage Time Delay

LOS	Density (PC/KM)[e]	Level Terrain						Rolling Terrain		
		95% No Passing Zones			100% No Passing Zones			100% No Passing Zones		
		% Time Delay	Speed (KM/H)[e]	Volume (PCPH)	% Time Delay	Speed (KM/H)[e]	Volume (PCPH)	% Time Delay	Speed (KM/H)[e]	Volume (PCPH)
A	≤4.4	≤58	≥73.1	320	≤59	≥73.1	320	≤48	≥68.6	300
B	≤8.8	≤76	≥69.8	610	≤76	≥69.8	610	≤67	≥65.1	570
C	≤14.4	≤83	≥68.2	980	≤84	≥68.2	980	≤74	≥63.7	915
D	≤22.5	≤89	≥66.1	1485	≤90	≥66.1	1485	≤81	≥62.4	1405
E	≤38.8	≤95[d]	≥62.9[c]	2360[b]	≤95[d]	≥62.9[c]	2360[b]	≤90	≥59.7[c]	2345[b]
F	>38.8									

[a] For 3.4 m lane width and 0.6 m shoulder width
[b] Rough estimate of maximum volume
[c] Speed at maximum volume
[d] Approximate
[e] PCPM = PC/KM * 1.6
[f] MPH = KM/H * 0.625

TABLE 4 LOS Criteria[a] for 80-km/hr AHS Two-Lane Highway Sections, Speed

LOS	Speed (KM/H)[d]	Level Terrain						Rolling Terrain		
		95% No Passing Zones			100% No Passing Zones			100% No Passing Zones		
		Density (PC/KM)[e]	% Time Delay	Volume (PCPH)	Density (PC/KM)[e]	% Time Delay	Volume (PCPH)	Density (PC/KM)[e]	% Time Delay	Volume (PCPH)
A	≥76.8	≤1.9	≤35	150	≤1.9	≤35	150	-	-	-
B	≥73.6	≤4.2	≤57	310	≤4.1	≤57	300	≤0.3	≤12	20
C	≥70.4	≤7.8	≤73	545	≤7.6	≤73	535	≤2.0	≤27	140
D	≥67.2	≤18.1	≤86	1220	≤17.9	≤86	1205	≤5.5	≤56	370
E	≥64.0	≤37.5[c]	≤93	2360[b]	≤37.5[c]	≤93	2360[b]	≤39.3[c]	≤73	2345[b]
F	<64.0									

[a] For 3.4 m lane width and 0.6 m shoulder width
[b] Rough estimate of maximum volume
[c] Density based on maximum volume and maximum speed
[d] MPH = KM/H * 0.625
[e] PCPM = PC/KM * 1.6

Design Speed [km/hr (mph)]	*Attainable LOS*
81 (51) or greater	A through F
65 to 80 (41 to 50)	B through F
49 to 64 (31 to 40)	C through F
33 to 48 (21 to 30)	D through F
32 (20) or less	E and F

Using these criteria, any of the values presented in Tables 1, 3, or 4 can be used for highways with 80-km/hr (50-mph) design speeds, but LOS A cannot be achieved.

It has also been proposed that both a percentage time delay and a speed criterion should be met together to attain a given LOS. For this purpose, it has been proposed to use the percentage time delay boundary values for general terrain segments and the upgrade speed criteria used in the 1985 HCM. The results of this approach are presented in Table 5. Volumes and densities corresponding to the percentage time delay values were obtained from Figures 2 and 5, respectively. Figure 3 was used to determine whether the speed requirement was met. Although the speed requirement was not met for rolling terrain at LOS D, it was considered sufficiently close to warrant inclusion in the table.

A similar exercise was carried out using the density boundaries presented in Table 3, in conjunction with speed; the volumes were obtained from Figure 4. Figure 6 was used to check whether the speed requirement was met; the results are presented in Table 6. As noted previously, the volumes in this table are probably too high and the LOS E density boundary value is unattainable, suggesting that adjustments are needed if density is to be used as a LOS criterion.

Combined Percentage Time Delay–Density as Basis

Because of apparent problems with the accuracy of the percentage time delay values in the model and the problem of determining appropriate density boundary values, an approach was devised whereby the advantages of using percentage time delay were retained without having to deal with the problem of inaccurate values.

When the HCM calculation procedures are applied, the focus is usually on the flow rate or volume-capacity (*V/C*) ratio. The percentage time delay is not usually relevant at this time, except insofar as the percentage time delay boundary values establish the *V/C*

TABLE 5 LOS Criteria[a] for 80-km/hr AHS Two-Lane Highway Sections, Time Delay and Speed

LOS	% Time Delay	Speed (KM/H)[c]	Level Terrain				Rolling Terrain	
			95% No Passing Zones		100% No Passing Zones		100% No Passing Zones	
			Density (PC/KM)[d]	Volume (PCPH)	Density (PC/KM)[d]	Volume (PCPH)	Density (PC/KM)[d]	Volume (PCPH)
A	≤30	≥88	-	-	-	-	-	-
B	≤45	≥80	-	-	-	-	-	-
C	≤60	≥72	≤4.6	340	≤4.6	330	-	-
D	≤75	≥64	≤8.3	580	≤8.2	575	≤15.3	975
E	>75	≥40	≤37.5	2360[b]	≤37.5	2360[b]	≤39.3	2345[b]
F	100	≤40						

[a] For 3.4 m lane width and 0.6 m shoulder width
[b] Rough estimate of maximum volume
[c] MPH = KM/H * 0.625
[d] PCPM = PC/KM * 1.6

TABLE 6 LOS Criteria[a] for 80-km/hr AHS Two-Lane Highway Sections, Density and Speed

LOS	Density (PC/KM)[d]	Speed (KM/H)[e]	Level Terrain				Rolling Terrain	
			95% No Passing Zones		100% No Passing Zones		100% No Passing Zones	
			% Time Delay	Volume (PCPH)	% Time Delay	Volume (PCPH)	% Time Delay	Volume (PCPH)
A	≤4.4	≥88	-	-	-	-	-	-
B	≤8.8	≥80	-	-	-	-	-	-
C	≤14.4	≥72	-	-	-	-	-	-
D	≤22.5	≥64	≤89	1485	≤90	1485	-	-
E	≤38.8	≥40	≤95[c]	2360[b]	≤95[c]	2360[b]	≤90	2345[b]
F	>38.8	<40						

[a] For 3.4 m lane width and 0.6 m shoulder width
[b] Rough estimate of maximum volume
[c] Approximate
[d] PCPM = PC/KM * 1.6
[e] MPH = KM/H * 0.625

values at the various LOS. It stands to reason then that the understanding gained through using percentage time delay can be used to determine LOS boundary values in terms of density, which can then be used for purposes of calculation and field measurement. Density is easier to calculate or obtain from field measurements, and it is also more readily convertible into speed and economic impact measures. Since the LOS boundaries in terms of percentage time delay, as currently stated in the HCM, are rarely if ever used directly, the absence of percentage time delay in the calculation procedures should not detract from the understanding gained through the continued use of percentage time delay.

With reference to Figures 2 and 5, it can be seen that the sensitivity of percentage time delay is far less at the higher ranges of volume and density than at the lower ranges. If LOS boundaries are to be defined in terms of the deterioration of service quality as volume and density increase, then smaller increments of volume and density would cause more change in LOS at the lower levels than at the higher levels. Density values corresponding to the "bending points" on the percentage time delay–density relationship, in Figure 5, are presented in Table 7. It is recognized that the boundary values are somewhat influenced by the shape of the curve as well as by the specific conditions simulated. The density boundary for LOS E was directly obtained from Table 2.

Corresponding values for percentage time delay and volume were obtained from Figures 5 and 4, respectively. Speed was calculated from density and volume. The results are also presented in Table 7.

The volumes obtained through this procedure appear to be reasonable. It is also noteworthy that the values for percentage time delay at the LOS boundaries are very close to those used in the 1985 HCM. This approach is therefore consistent to a degree with the percentage time delay–based LOS definition used in the 1985 HCM (reflected in Table 1 of this paper).

STRATEGY FOR FUTURE DEVELOPMENT

Given the widely divergent views held by the parties involved, it will take some time for consensus to be achieved on how to improve

TABLE 7 LOS Criteria[a] for 80-km/hr AHS Two-Lane Sections, Densities and HCM Percentage Time Delay Values

LOS	Density (PC/KM)[e]	Level Terrain						Rolling Terrain		
		95% No Passing Zones			100% No Passing Zones			100% No Passing Zones		
		% Time Delay	Speed (KM/H)[f]	Volume (PCPH)	% Time Delay	Speed (KM/H)[f]	Volume (PCPH)	% Time Delay	Speed (KM/H)[f]	Volume (PCPH)
A	≤1.9	≤33	≥77.3	145	≤34	≥77.3	145	≤26	≥69.3	130
B	≤3.1	≤49	≥73.6	230	≤49	≥73.6	230	≤40	≥65.6	205
C	≤5.6	≤66	≥72.0	405	≤67	≥72.0	405	≤57	≥67.5	380
D	≤12.5	≤82	≥68.0	850	≤83	≥68.0	850	≤73	≥64.0	800
E	≤38.8	≤95[d]	≥62.9[c]	2360[b]	≤95[d]	≥62.9[c]	2360[b]	≤90	≥59.7[c]	2345[b]
F	>38.8									

[a] For 3.4 m lane width and 0.6 m shoulder width
[b] Rough estimate of maximum volume
[c] Speed at maximum volume
[d] Approximate
[e] PCPM = PCPK * 1.6
[f] MPH = KPH * 0.625

the capacity and LOS analysis methodology for two-lane highways with low design speeds. In the meantime, the need remains for a fully defined procedure for analyzing the capacities and LOS for these facilities. The obvious solution would be to proceed with percentage time delay as the primary parameter until the necessary decisions can be made in forums such as TRB's Committee on Highway Capacity and Quality of Service. However, in view of the inaccuracies and other problems experienced with the percentage time delay parameter, another course of action should be considered.

A course of action that may be pursued immediately, which does not deviate a great deal from using percentage time delay as the primary parameter, is to use the procedure presented in Table 7. This method retains the principle of percentage time delay while not relying specifically on the accuracy of the simulated percentage time delay values. The method is somewhat different from the 1985 HCM method and may therefore prompt the question of whether it is appropriate to use two different methods for different design speeds. However, there does appear to be a viewpoint that the LOS for low-design-speed roads could be analyzed differently, namely, by limiting high LOS at low design speeds for other facility types. The option of having to meet two criteria to attain a given LOS, such as percentage time delay and speed or percentage time delay and density, was regarded favorably by several of the consultants who reviewed the issue paper developed during this study (*6*).

Consensus should also be reached in the medium term on whether to limit the high LOS at low design speeds. In the long term, it is essential to consider system effects since, in California for instance, the 1985 HCM is used by law for congestion management purposes. The parameters and analysis procedures should therefore take cognizance of systemwide decisions. This will be a departure from the existing framework of the HCM, in which the different types of facilities are treated independently without reference to system considerations. The questions of whether to use density as the primary parameter, the effect of the functional classification of the facility, consistency of analysis (i.e., general terrain segment versus specific grade analysis), and single-direction-based analysis should all be addressed further in future research.

It should be noted that the analysis options described in this paper are by no means the only feasible options. The fundamental relationships shown in Figures 1 through 6 can be used to test other feasible options, boundary values, etc. These results are presented in the spirit of encouraging additional exploration into methodological alternatives and constructive debate over the best direction for further evolution of the HCM.

ACKNOWLEDGMENTS

The research was funded by Caltrans and the Department of Civil Engineering and Applied Mechanics, San Jose State University. The authors would like to thank Fred Rooney, Rick Knapp, Pat Secoy, Guy Luther, Ken DeCrescenzo and Paul Vonada, all of Caltrans, for their help throughout the project. A special word of thanks is also due to Doug Harwood of the Midwest Research Institute for his interest and help with the implementation of the TWOPAS model. In addition, the authors gratefully acknowledge the assistance of Robert Layton of Oregon State University, John McLean of the Australian Road Research Board, and John Morrall of the University of Calgary for their valuable service as consultants to this project.

REFERENCES

1. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
3. *A Policy on Geometric Design of Highways and Streets.* AASHTO, Washington, D.C., 1990.
4. Botha, J. L., X. Zeng, and E. C. Sullivan. Comparison of Performance of TWOPAS and TRARR Models When Simulating Traffic on Two-Lane Highways with Low Design Speeds. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993.
5. Botha, J. L., E. C. Sullivan, and X. Zeng. *Level of Service of Two-Lane Highways With Design Speeds Less than 60 mph.* Final Report, Vols. 1 and 2. California Department of Transportation, Sacramento, June 1993.
6. Botha, J. L., and E. C. Sullivan. *Level of Service Concept for Two-Lane Roads Revisited.* Working Paper. San Jose State University; California Polytechnic University, San Luis Obispo, May 1992.

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

Economic Feasibility Assessment Procedure for Climbing Lanes on Two-Lane Roads in Mexico

ALBERTO MENDOZA AND EMILIO MAYORAL

The development of procedures for analyzing the economic feasibility of constructing climbing lanes on two-lane roads in Mexico and for assessing the levels of service before and after implementing these facilities is presented. Initially, the weight-horsepower ratio that is representative of the Mexican freight vehicles is obtained. Then this ratio is compared with the value reported for this parameter in the United States. From this comparison and a data set collected in a series of grades specific to Mexico, the different operating conditions of the trucks and the vehicular flows between the two countries are shown. The foregoing also shows the need to adapt to the Mexican conditions the procedure and criteria for specific grades in the 1985 *Highway Capacity Manual* and to generate an economic feasibility assessment procedure for climbing lanes in Mexico. These tasks are carried out by using model calibration, simulation, and regression analysis techniques. From the procedures developed, the potential use of climbing lanes in Mexico is discussed briefly. Finally, a series of conclusions and recommendations is outlined.

In Mexico two-lane roads account for more than 95 percent of the 46 000 km that constitutes the federal trunk road system. The 1988 Mexican Government Program for the Modernization of the Transportation System established the convenience of carrying out comparatively inexpensive improvements on the two-lane trunk road segments with long and steep grades (longer than 800 m and steeper than 3 percent) and considerable traffic, particularly trucks (*1*). The provision of extra climbing lanes in such segments (hereafter referred as specific grades) traditionally has been considered among the most efficient of low-cost improvements.

The preceding represents the starting point of this Instituto Mexicano del Transporte (IMT) research project, whose main features and findings are described in this paper. The objective of this research was to develop procedures for analyzing the economic feasibility of constructing climbing lanes on two-lane roads in Mexico and for assessing the levels of service (LOS) before and after implementing these facilities.

REPRESENTATIVE WEIGHT-HORSEPOWER RATIO FOR MEXICAN TRUCKS

Initially, a study was conducted to determine the weight-horsepower ratio that is representative for Mexican freight vehicles. This parameter has an important effect on reducing the speeds of heavily loaded trucks moving uphill on specific grades. If this reduction is significant, trucks impede following vehicles, degrading traffic operations.

Infrastructure Division, Instituto Mexicano del Transporte, P.O. Box 1098, 76000 Querétaro, Mexico.

Truck Weight Limits Authorized in Mexico and Other Countries

Even though the weight and dimension regulation in force today (put into effect in 1980) authorizes 16 types of freight vehicle, only the following 5 compose the truck flows that travel on Mexican roads: Type 2 (35 percent), Type 3 (22 percent), Type 3-S2 (24 percent), Type 3-S3 (15 percent), and Type 3-S2-4 (2 percent). For these truck types, Table 1 presents a summary of the weight limits specified in Mexico (according to the 1980 regulation and to a new regulation project expected to be put into effect early in 1994) and in six other countries. As indicated in Table 1, American, Japanese, and Spanish regulations allow greater weight limits for single-unit trucks (Types 2 and 3) than does the 1980 Mexican regulation. However, the Mexican regulation allows much greater limits for the heavier trucks (tractor-semitrailer, doubles, and full trailer combinations). Such dissimilarities arise as a result of the different criteria used in each country for the determination of maximum vehicle weights; for instance, in the United States, truck weights are regulated through limits on axle load (to control pavement damage), a bridge formula (to control bridge damage), and a total maximum vehicle weight of 36 T; the 1980 Mexican regulation considers only axle load. The new 1994 regulation project has already incorporated a bridge formula criterion.

Representative Truck Weight-Horsepower Ratio

A truck weight and dimension survey was carried out in 1991 by the Secretaría de Comunicaciones y Transportes on 10 points of the trunk road system. This survey provided the information needed to estimate the required ratio. In each point, for each vehicle surveyed, the information gathered consisted of the weight, dimensions, horsepower rating (specified by the manufacturer), nature of the transported freight, and origin and destination. The vehicular weights were obtained using weigh-in-motion equipment. Nearly 100,000 trucks were surveyed. After the compiled information was analyzed, the following results were obtained:

- Of all trucks surveyed (about 100,000), 22 percent were overweight.
- Of only the loaded trucks surveyed (about 70,000), 34 percent were overweight.
- The overloaded vehicles were an average of 20 percent overweight.

These figures indicate the lack of appropriate mechanisms to guarantee the adequate enforcement of weight and dimension laws.

TABLE 1 Truck Weight Limits Authorized in Mexico and Other Countries[a]

COUNTRIES		Type 2	Type 3	Type 3-S2	Type 3-S3	Type 3-S2-4
MEXICO	1980	15.5	23.5	41.5	46.0	77.5
	1994	16.5	24.5	41.1	43.6	59.1
USA		18.2[b]	24.5[b]	36.4[b]	36.4[b]	36.4[b]
CANADA		19.0	27.8	44.5	57.5	61.8
JAPAN		Total weight not to exceed 20 ton[c]				
BRAZIL		15.0	22.0	39.0	45.0	73.0[d]
SPAIN		18.0[e]	25.0[ef]	44.0[ef]	44.0	40.0
AUSTRALIA		15.0[g]	22.5[g]	39.0[g]	42.5[g]	72.0[g]

[a] In metric tons.
[b] Regulations vary by Province or State. These values are based on Federal Bridge Formula.
[c] Special vehicles up to 34 ton in total weight can be operated in the National Motorway System.
[d] Can operate only with special authorization and on specific roads.
[e] Provided it does not exceed 5 tons per meter of length measured between the first and the last axle.
[f] Subject to axle spacing restrictions.
[g] Maximum gross weight assumes single tyre steer axle with all other axles having dual tyres. Axle spacing requirements also apply to all vehicles.

For this scenario (the practical nonexistence of effective enforcement), a representative weight-horsepower ratio of 210 kg/hp was obtained (as the 95th percentile on the corresponding cumulative distribution). Under a hypothetical scenario of effective enforcement mechanisms and considering the weight limits specified by the new 1994 regulation project, a representative weight-horsepower ratio of about 160 kg/hp was obtained. It is believed that the provisions in the new 1994 regulation will eliminate the problem of overloading.

The observed weight-horsepower ratio of 210 kg/hp contrasts with the value of 135 kg/hp reported for this parameter (heavy trucks) in the United States, in the 1985 *Highway Capacity Manual* (HCM) (*2*). The main reason for this difference is the much heavier freight vehicles that travel on the Mexican roads as a consequence of the higher weight limits permitted by the regulation, as well as the truck overloads. More than 25 years ago, the representative weight-horsepower ratio of heavy trucks was very similar in both countries, that is, about 180 kg/hp (*3,4*).

In Mexico, then, the weight-horsepower ratio of trucks has grown as a result of consistent payload increments that have exceeded tare weight decrements and horsepower increments provided by the development of automotive technology. Yet in the United States, the weight-horsepower ratio of trucks has decreased. In Mexico, growth in the efficiency of freight vehicles has resulted in more heavily loaded but not in faster vehicles. In the United States, though, such growth has resulted in faster and only slightly more loaded vehicles. In essence, the conditions prevailing in each country represent two strategies for moving freight in trucks: (a) few trips in heavily loaded and slow trucks, and (b) many trips in slightly loaded and fast trucks. In general, as will be shown later, the Mexican strategy results in slower trucks traveling on the roads, which contributes to slower traffic flows. This situation is magnified by the fact that the average proportion of trucks is much higher on Mexican roads [35 percent (*5*)] than it is on American roads [14 percent (*2*)].

Such operating peculiarities of vehicular flows as well as the characteristics of Mexican roads justify the IMT's efforts to develop an economic feasibility assessment procedure for climbing lanes on two-lane roads in Mexico. For this purpose, it was necessary to first adapt to the Mexican conditions the LOS criteria for specific grades in the 1985 HCM. This task was accomplished using a data set collected in a series of specific grades and the Australian simulation model for rural roads, TRARR (*6*).

FIELD DATA COLLECTED

A data set was collected in 10 specific grades with design speeds higher than 90 km/hr, climbing lanes, and good pavement surface condition. The data collection was planned so that it would show the benefits derived from installing these facilities under different conditions of traffic, visibility, and length and percentage of grade. In all grades surveyed, the relevant alignment characteristics were taken (lane and shoulder widths, length and percentage of grade, degree of curves, and percentage of no-passing zones) along with traffic counts with vehicle classification, average upgrade and downgrade speeds, and percentage of vehicles in both directions traveling in platoons behind a leader at headways of 5 sec or less (which provides a useful measure of the quality of the traffic flow). The last three aspects were measured with the climbing lane open and closed consecutively, in 15-min intervals, during 16 hr in each grade (from 6 a.m. to 10 p.m.). Analysis of the collected data obtained the following results:

- The traffic streams in all grades surveyed were composed, on average, of 49 percent passenger cars, 10 percent buses, and 41 percent trucks.
- In all grades surveyed with the extra climbing lane closed (two-lane section), 148 hourly average upgrade speeds were recorded. In addition, on the basis of the alignment and traffic characteristics registered in each grade, the corresponding average upgrade speeds were computed according to the procedure indicated in the HCM. As indicated in Figure 1, the computed speeds were, on average, 56.5 percent higher than the recorded values. This shows the different speed levels that prevail on two-lane roads in Mexico and the United States and the need to adapt to the Mexican conditions the LOS criteria for specific grades in the HCM.
- For the downhill direction, the average speeds resulted a little higher than uphill, thus indicating better LOS in that direction. How-

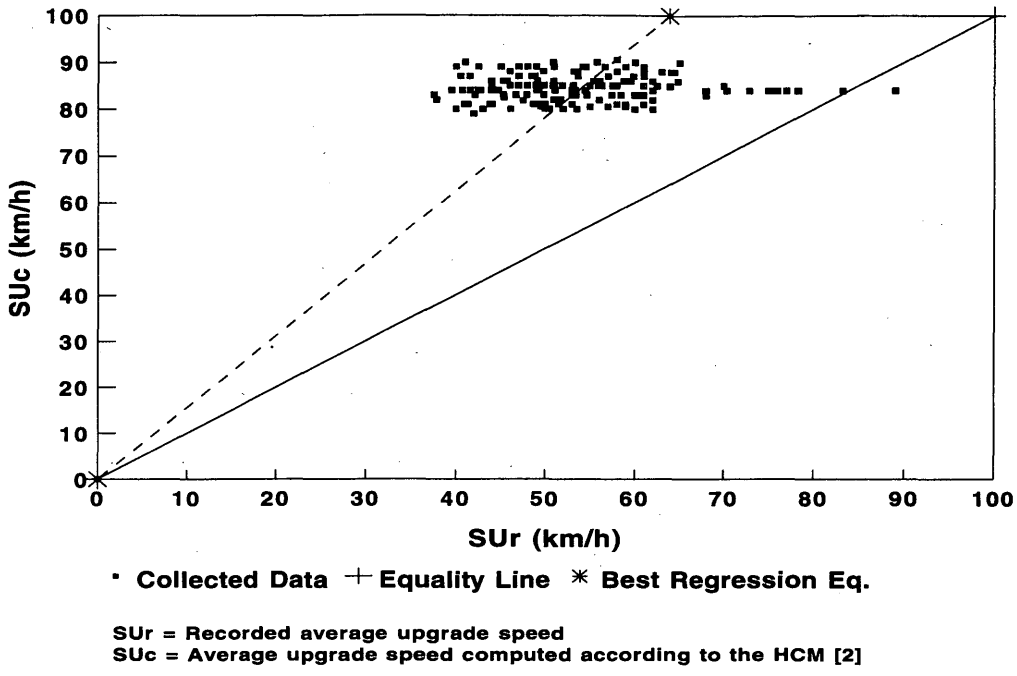


FIGURE 1 Speeds computed according to HCM versus speeds recorded in specific grades.

ever, like the uphill, the downgrade speeds showed a decreasing trend when the percentage of grade increases. This effect is attributed to the vehicles' risk of losing their brakes (which induces them to descend in low gear) and to the horizontal alignment curvatures. The following formula, valid for up to a 6 percent grade, describes the best regression relationship obtained (R^2 = .82) among the downgrade and upgrade speeds and the percentage of grade:

$$SD = SU/[1 - (M/21)] \quad (1)$$

where

SD = average downgrade speed (km/hr),
SU = average upgrade speed (km/hr), and
M = grade (%).

- The LOS criteria for specific grades in the HCM were applied to the recorded upgrade speeds and to their corresponding computed values according to the HCM procedure. For the recorded speeds, which were low with regard to the speed scale from which the LOS were defined (Table 8.2 in the HCM), the most common LOS obtained were E and F (for 83.7 percent of the hourly data recorded). This situation suggests that the LOS criteria in the HCM are too rigorous for Mexico. For the computed speeds, on the other hand, which were rather high, LOS B was the most common (for 89 percent of the computed speeds). This situation indicates that to be applicable in Mexico, the HCM speed calculation procedure should be modified.
- For a subset of 134 data, the ratio of average speed with the climbing lane open (three-lane section) to average speed with the climbing lane closed (two-lane section) was computed (obviously, for traffic traveling in the same direction as the climbing lane). This ratio indicates the percentage of speed gained when a climbing lane is installed. In Figure 2, it is apparent that this gain is higher when the speed in the two-lane section is lower. Such a trend is suitably represented by the regression equation included in Figure 2 (R^2 = .80). For the 134 data points, the average gain was about 20 percent (an average ratio equal to 1.2).
- Figure 3 depicts an observed trend for the free-flow operating speed of passenger cars versus the percentage of grade. This trend was obtained for passenger cars operating on some of the shortest grades surveyed (between 400 and 600 m long) during conditions of very low traffic density. Likewise, Figure 3 shows the corresponding trends for cars in the United States and Colombia (*2,7*). The curve for the American cars was inferred from Table 8.7 in the HCM for a flow rate–to–capacity ratio of 0 and 0 percent no-passing zones. In Figure 3, the U.S. cars are less affected by the percentage of grade than the Mexican and Colombian cars. This apparent difference in behavior is due mainly to the fact that HCM Table 8.7 was obtained from a simulation based on an ideal tangent section of highway, whereas the Mexican and Colombian curves come from field data taken in real road conditions, particularly horizontal alignment.

TRARR MODEL

TRARR is a microscopic computer model that simulates the traffic operations on rural two-lane, two-way roads without intersections. The model takes into account the effect of overtaking prohibitions, auxiliary lanes (such as passing or climbing lanes), horizontal and vertical curves, variable sight distance, and driver/vehicle characteristics. TRARR generates the traffic entering the simulated road segment and reviews the progress of the position, speed, and acceleration of each vehicle along the road at frequent intervals (typically 1 sec). The program requires four input data files containing the following information:

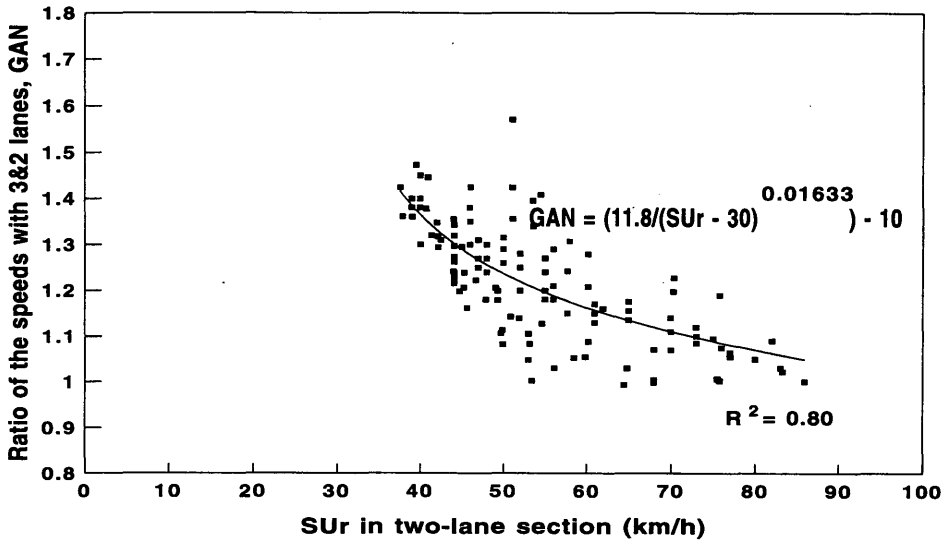


FIGURE 2 Ratio of speeds with three and two lanes versus speed in two-lane section.

1. Traffic (volume, composition, directional split, and mean and standard deviation of desired speeds; also the settling down time and duration of the simulation run);
2. Road geometry (grades, curves, barrier lines, sight distances, and auxiliary lanes);
3. Driver/vehicle characteristics (length, acceleration, and following and overtaking behavior for up to 18 vehicle types); and
4. Type of output information required and location of observing points and intervals along the simulated segment.

For each direction of travel, the model gives output values at the specified observing points (spot mean speed and percentage of vehicles in platoons) and intervals (mean and standard deviation of travel time and speed, overtaking rate, percentage of travel time spent following, and fuel consumption).

TRARR was calibrated using the geometric characteristics of the specific grades surveyed as well as the operation information collected. The program input data were taken, specifically, from the field information gathered, typical characteristics of Mexican vehicles, and default values recommended for the model (*6*). Mean desired speeds of 96, 92, and 72 km/hr and standard deviations of 13, 12, and 9 km/hr were used for passenger cars, buses, and trucks, respectively, as obtained from the field data.

The model was calibrated to minimize the difference between simulated and field values for average upgrade speeds and percentage of travel time spent following (or percentage of vehicles in pla-

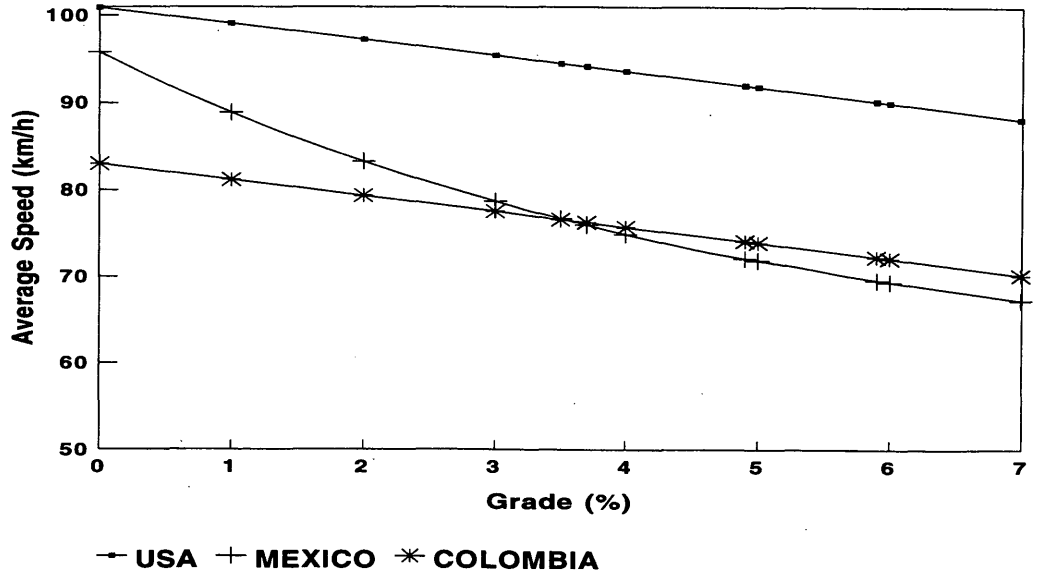


FIGURE 3 Free-flow operating speed of passenger cars versus percentage grade, in three countries.

toons). This was achieved not only by using characteristics of vehicles in Mexico, but also by redefining a series of multipliers for the desired speeds that the program uses to account for speed reductions due to certain road characteristics (mainly horizontal curves). After calibration, the model predicted average upgrade speeds for the different vehicle types and percentage of vehicles traveling in platoons with maximum errors of 15 and 17 percent, respectively, for a 90 percent confidence interval.

MODIFICATION OF LOS CRITERIA FOR SPECIFIC GRADES

From the calibrated model and the field data collected, the procedure and criteria in the HCM for LOS computation in specific grades were modified to better suit the Mexican conditions. The modifications lead to a procedure for evaluating the LOS of specific grades on two-lane roads in Mexico.

In different countries, the capacity under ideal conditions is typically assumed to be between 2,800 and 3,200 passenger cars per hour (pcph) (*2,7*). In each country, the particular value of this parameter depends on the characteristics of vehicles and drivers. For the United States, the HCM recommends a value of 2,800 pcph (*2*). Studies or field observations have not been carried out yet to determine the value of this parameter for Mexico. Some aspects suggest that it could be higher than 2,800 pcph (because Mexican drivers are more aggressive in their overtaking behavior), whereas others indicate that it could be lower (because American vehicles have more power). However, independent of the exact value of this parameter, the effect of its typical range of variation on the upgrade speeds is less important than the effect of other factors such as the presence of heavy vehicles in the traffic streams or the presence of curves in the horizontal alignment. For this reason, a value of 2,800 pcph was adopted for this study. However, it is recognized that studies are needed to refine this value.

Ratio of Flow Rate to Capacity Versus Average Upgrade Speed

For American passenger cars, ratios of flow rate to capacity versus average upgrade speed are contained in HCM Table 8.7, which was produced from simulation using a tangent section of highway.

For conditions in Mexico, these relationships were redefined using the TRARR model. However, in this case, for the different values of percentage of grade, instead of using a tangent section of highway, horizontal alignments of existing Mexican grades were used (about 400 m long). Table 2 gives the values obtained for Mexico. In the generation of Table 2, with exception of the horizontal curves, ideal conditions were used for the other factors (even directional distribution, passenger cars unaffected by grades, etc).

Table 2 is in complete agreement with the behavior of Mexican passenger cars depicted in Figure 3. Thus, in Table 2, the percentage of grade affects the operating speed of passenger cars more than

TABLE 2 Values of *F/c* Ratio[a] Versus Speed, Percentage Grade, and Percentage No-Passing Zones for Two-Lane Specific Grades in Mexico

PERCENT GRADE	AVERAGE UPGRADE SPEED	PERCENT NO PASSING ZONES (%)					
(%)	(km/h)	0	20	40	60	80	100
	76	0.17	0.00	0.00	0.00	0.00	0.00
	68	0.58	0.43	0.31	0.14	0.00	0.00
3	60	0.94	0.83	0.71	0.58	0.44	0.30
	52	1.00	1.00	1.00	0.94	0.83	0.71
	44	1.00	1.00	1.00	1.00	1.00	1.00
	72	0.19	0.02	0.00	0.00	0.00	0.00
	64	0.59	0.46	0.31	0.17	0.00	0.00
4	56	0.95	0.84	0.72	0.59	0.46	0.31
	48	1.00	1.00	1.00	0.95	0.84	0.72
	40	1.00	1.00	1.00	1.00	1.00	1.00
	68	0.23	0.06	0.00	0.00	0.00	0.00
	60	0.64	0.51	0.37	0.22	0.07	0.00
5	52	0.98	0.88	0.77	0.64	0.51	0.37
	44	1.00	1.00	1.00	0.98	0.88	0.77
	36	1.00	1.00	1.00	1.00	1.00	1.00
	68	0.10	0.00	0.00	0.00	0.00	0.00
	60	0.52	0.38	0.23	0.08	0.00	0.00
6	52	0.89	0.78	0.65	0.52	0.38	0.23
	44	1.00	1.00	0.98	0.89	0.78	0.65
	36	1.00	1.00	1.00	1.00	1.00	0.98
	64	0.20	0.03	0.00	0.00	0.00	0.00
	56	0.62	0.48	0.34	0.20	0.04	0.00
7	48	0.96	0.86	0.75	0.62	0.48	0.34
	40	1.00	1.00	1.00	0.96	0.86	0.75
	32	1.00	1.00	1.00	1.00	1.00	1.00

[a] Ratio of flow rate to ideal capacity of 2800 pcph, assuming passenger-car operation is unaffected by grades.

in HCM Table 8.7. Likewise, for the lower speed levels in Table 2 and for the horizontal curves typically associated in Mexico with the different values of grade percentage, a similar percentage of no-passing zones generates more interferences among the vehicles and therefore results in greater upgrade speed reductions than in HCM Table 8.7.

Passenger Car Equivalents

For U.S. conditions, passenger car equivalents are presented in HCM Table 8.9. These values were adapted to the Mexican conditions according to the following procedure:

- Average proportions of trucks and buses in Mexican two-lane roads equal to 35 and 10 percent, respectively, were assumed.
- On the basis of simulation of traffic operations on existing grades, the mixed flow rate corresponding to different combinations of average upgrade speed and length and percentage of grade was obtained. Then the passenger cars and heavy vehicles, components of that flow rate, were simulated separately and their separate speed distributions were assessed.
- From the speed distributions obtained, the corresponding passenger car equivalents were computed using the following expression:

$$E = r/R \tag{2}$$

where

E = passenger car equivalent for a given length and percentage of grade and average upgrade speed;
R = number of passings between passenger cars per kilometer of road, assuming that each vehicle when passing or being passed continues at its corresponding speed in speed distribution; and
r = number of passings of heavy vehicles by passenger cars per kilometer of road.

The values of R and r were computed on the basis of the speeds and relative frequencies in the corresponding speed distributions of passenger cars and heavy vehicles (*3*).

- The equivalents thus obtained were refined slightly to improve the fit between the upgrade speeds recorded in the field and their corresponding computed values after considering the earlier modifications. This final fitting is justifiable because the procedure used to determine the equivalents provides only approximate values for these parameters.

Table 3 gives the set of equivalents finally obtained. After the modifications, the prediction of upgrade speeds using the calibrated procedure improved noticeably, as may be seen from comparing Figures 1 and 4. It may also be observed that the Mexican equivalents in Table 3 are much higher than the equivalents in HCM Table 8.9.

LOS Criteria

For the U.S. conditions, HCM Table 8.2 presents the boundary speeds for the different LOS. These speeds were redefined for the Mexican conditions as follows:

1. For each flow rate/capacity ratio in HCM Table 8.7 corresponding to a specific combination of upgrade speed, percentage of grade, and percentage of no-passing zones, the respective upgrade speed for the Mexican conditions was obtained from Table 2. Through this process, a data set of speeds from both tables, equivalents in terms of flow rate/capacity, was generated (186 data pairs).
2. A regression model was fitted through the speed data. In the model developed, $R^2 = .84$. From this model, the boundary speeds for Mexico that are equivalent in flow rate/capacity to the speeds in HCM Table 8.2 were defined. Finally, these values were refined by rounding them to multiple values of 5 and leaving a uniform speed interval between LOS (similarly to HCM Table 8.2).

Table 4 presents the boundary values finally obtained. In essence, the lower speeds in Table 4 as compared with the speeds in HCM Table 8.2 indicate that Mexican drivers are more tolerant than Americans regarding their travel speeds on specific grades (or their travel times).

In addition, Equation 3, which relates the exact speed at which capacity occurs with the flow rate at that speed, is more adequate for the Mexican conditions than Equation 8.8 of the HCM:

$$S_c = 35 + 3.75\,(F_c/1{,}000)^2 \tag{3}$$

where S_c is the speed at which capacity occurs, in kilometers per hour, and F_c is the flow rate in capacity, in mixed vehicles per hour (mixed vph).

Equation 3 assumes a maximum flow rate of 2,000 vph, similar to HCM Equation 8.8. Likewise, Equation 3 considers that for specific grades between 3 and 7 percent and up to 6.4 km long, the average speed at which capacity occurs varies between the speed range specified in Table 4 for LOS E.

Other Considerations

Apart from the modifications suggested earlier, the remaining aspects of the LOS calculation procedure for specific grades in the HCM are considered applicable for Mexico. Other studies have shown that for the typical traffic conditions and cross sections of two-lane trunk roads in Mexico (lane widths between 3.30 and 3.50 m and shoulder widths between 0.6 and 1 m), the capacity adjustment factors in the HCM for directional distribution and for narrow lanes and restricted shoulder width can be used with accuracy (*5*).

In this study, Tables 2 and 3 include not only the operating limitations of Mexican vehicles in specific grades but also the effect of the horizontal alignment curves. Another way to consider this last effect besides including it in Tables 2 and 3 would consist of introducing in the procedure a capacity adjustment factor to account for the presence of curves of specific radii (a parameter that is closely related to the design speed of the road).

ECONOMIC FEASIBILITY

For specific grades shorter than 3000 m, the economic feasibility was analyzed for a series of cases corresponding to different combinations of the following factors at three levels: traffic flow rate, percentage of trucks in the traffic stream, percentage of traffic on an upgrade, percentage and length of grade, and percentage of no-passing zones. Table 5 gives the levels used for these factors.

TABLE 3 Passenger Car Equivalents for Specific Grades on Two-Lane Rural Highways in Mexico

GRADE (%)	LENGTH OF GRADE (m)	AVERAGE UPGRADE SPEEDS (km/h)								
		76	72	68	64	60	56	52	44	32
0	ALL	6.1	5.7	5.4	5.1	4.9	4.7	4.5	4.3	4.0
3	400	7.1	6.5	6.0	5.7	5.4	5.2	4.9	4.6	4.2
	800	7.4	6.7	6.2	5.8	5.5	5.2	5.0	4.6	4.2
	1600	8.9	7.9	7.2	6.6	6.2	5.8	5.5	5.0	4.4
	3200	16.9	14.4	12.5	11.1	10.1	9.2	8.5	7.4	6.2
	4800	38.3	31.0	25.9	22.2	19.4	17.3	15.5	12.9	10.3
	6400	91.2	70.4	56.7	47.0	39.8	34.4	30.1	23.9	18.0
4	400	7.7	7.1	6.6	6.2	5.8	5.6	5.3	4.9	4.5
	800	8.4	7.6	7.0	6.5	6.1	5.8	5.5	5.0	4.5
	1600	11.2	9.8	8.8	8.0	7.4	6.9	6.5	5.8	5.1
	3200	26.2	21.7	18.5	16.2	14.4	13.0	11.8	10.0	8.2
	4800	70.3	55.0	44.8	37.5	32.1	27.9	24.7	19.8	15.2
	6400	a	a	a	90.2	74.4	62.7	53.7	41.0	29.4
5	400	8.7	7.9	7.3	6.8	6.4	6.1	5.8	5.4	4.9
	800	9.9	8.8	8.1	7.5	7.0	6.6	6.2	5.7	5.1
	1600	14.6	12.6	11.2	10.1	9.3	8.6	8.0	7.1	6.1
	3200	41.4	33.6	28.1	24.2	21.1	18.8	16.9	14.0	11.2
	4800	a	98.9	78.4	64.1	53.7	45.9	39.8	31.0	22.9
	6400	a	a	a	a	a	a	96.5	70.9	48.4
6	400	9.9	9.0	8.3	7.7	7.2	6.9	6.5	6.0	5.4
	800	11.9	10.5	9.6	8.8	8.2	7.7	7.2	6.5	5.8
	1600	19.5	16.7	14.6	13.1	11.9	10.9	10.1	8.8	7.5
	3200	66.4	52.6	43.3	36.5	31.5	27.6	24.5	19.9	15.4
	4800	a	a	a	a	90.4	75.8	64.6	49.0	34.7
	6400	a	a	a	a	a	a	a	a	80.1
7	400	11.6	10.4	9.6	8.9	8.3	7.8	7.4	6.8	6.1
	800	14.6	12.9	11.6	10.6	9.8	9.1	8.6	7.7	6.7
	1600	26.6	22.5	19.5	17.3	15.5	14.1	13.0	11.2	9.3
	3200	a	83.1	67.0	55.7	47.3	40.9	35.9	28.5	21.5
	4800	a	a	a	a	a	a	a	77.6	53.0
	6400	a	a	a	a	a	a	a	a	a

[a] Speed not attainable on grade specified.

Specifically, the feasibility analysis was carried out for 243 combinations of the 729 total possible combinations (one-third of the complete factorial experiment). For these analyses, the widths of lanes and shoulders were fixed at 3.5 and 0.8 m. A uniform proportion of buses equal to 10 percent was also assumed. These values are typical of two-lane roads in Mexico.

For the different cases analyzed, the TRARR model was used to evaluate the upgrade speed gain obtained from adding a climbing lane for passenger cars, buses, and trucks. Subsequently, since these gains decrease the operating costs of the different vehicle types, the travel time value of automobile and bus passengers, and the opportunity cost of the freight transported, the corresponding economic benefits were quantified. The operating costs of passenger cars, buses, and trucks, as a function of speed, were computed on the basis of the World Bank model Vehicle Operating Costs (VOC), calibrated for Mexican conditions. The benefits were assessed using standard procedures (*8*).

In other countries, it is considered that the addition of these lanes reduces the number of serious accidents occurring on grades, particularly rear-end collisions involving trucks (*2*). However, in Mexico, since there is no reliable evidence of the preceding, benefits derived from improved safety were not included. In general, the benefits due to reductions in vehicle operating costs were much more significant than the other benefit types calculated. For this reason, while adopting a conservative approach in the evaluation of the economic feasibility, only those benefits were considered in subsequent stages of the study. The construction cost was also obtained for the cases analyzed.

On the basis of the corresponding annual benefit and cost flows, the benefit/cost ratio and the internal rate of return were computed for each case analyzed for a 10-year period, which is considered in Mexico to be reasonable for the transition toward more costly multilane solutions (*4*). The analysis resulted in the following regression equations for predicting the upgrade speed gain obtained from the addition of the extra lane (Equation 4), the first-year benefit due to the reduction of vehicle operating costs (Equation 5), the benefit/cost ratio (Equation 6), and the internal rate of return (Equation 7):

$$\text{GAN} = 0.845 + 1.61E - 4M^{2.9} + 3.71E - 8M^{2.9}\text{PU}^{2.4} + 3.55E - 3L^{0.2}F^{0.5} + 2.77E - 4\text{PNP PT}^{0.5} \quad (4)$$

$$\text{BEN}_o = \frac{1}{903.7\, e^{[22.47 - 0.032\,\text{SU}(\text{GAN} - 0.98)^{-0.653}]}} \quad (5)$$

$$\text{B/C} = \frac{0.047\,\text{BEN}_o^{0.88}\,(\text{GR} + 20)^{2.17}}{(\text{DR} + 5)^{0.444}\text{C}_o^{0.935}} \quad (6)$$

$$\text{IRR} = (2.257\text{B/C} - 1.257)\,\text{DR} \quad (7)$$

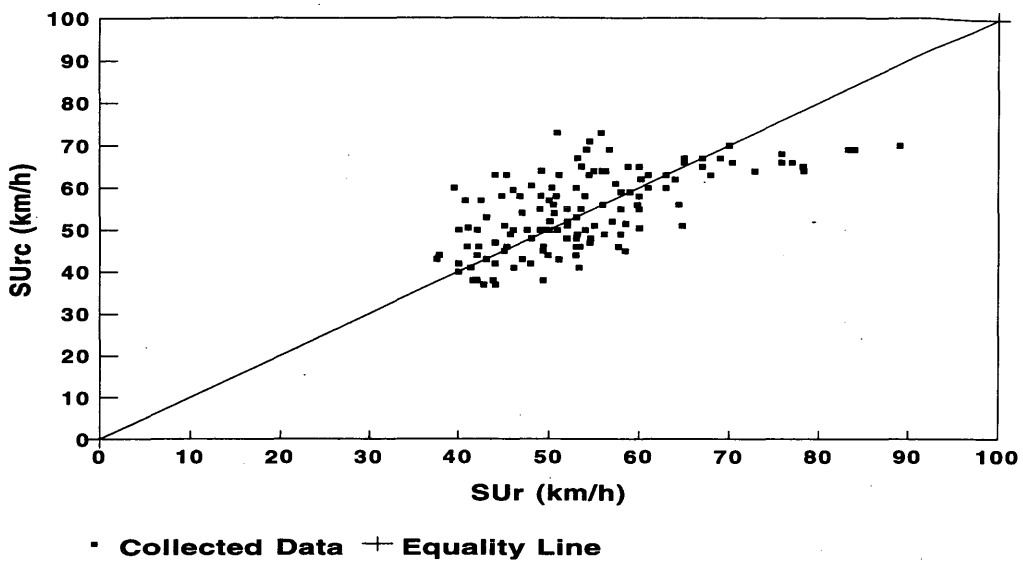


FIGURE 4 Speeds recalculated according to modified HCM procedure versus speeds recorded in specific grades.

where

GAN = upgrade speed gain obtained from adding a climbing lane (ratio of average speeds after and before installation of extra lane);
M = grade (%);
PU = traffic on upgrade (%);
L = length of grade (m);
F = hourly flow rate corresponding to rate of peak 15 min (both directions) (mixed vph);
PNP = no-passing zones (%);
PT = trucks (%);
BEN_o = first-year benefit of climbing lane due to reduction of vehicle costs (1993 pesos/km);
e = 2.71828;
SU = average upgrade speed (km/hr) (can be measured directly in the field or computed by using the HCM procedure modified as suggested earlier);
B/C = benefit/cost ratio;
GR = annual traffic growth rate (%);
DR = annual discount rate (%);
C_o = first cost of climbing lane (1993 pesos/km) (should be deflated to the 1993 value if computed for another year later than 1993); and
IRR = internal rate of return (%).

In general, an adequate goodness of fit was obtained for the preceding equations ($R^2 > .80$). They can be used, consecutively, for assessing the economic feasibility of constructing climbing lanes on specific grades shorter than 3000 m. These equations can be programmed easily in hand calculators.

As mentioned earlier, Equation 4 was developed for grades up to 3000 m long. For grades longer than 3000 m, the speed gain obtained from adding the extra lane should be computed from the regression equation in Figure 2. This equation is valid for average upgrade speeds before installing the climbing lane between 35 and 100 km/hr. Obviously, to use this equation, the upgrade speed should be measured directly in the field. The rest of the feasibility analysis can be completed by using Equations 5, 6, and 7.

TABLE 4 LOS Criteria for Specific Grades in Mexico

LEVEL OF SERVICE	AVERAGE UPGRADE SPEED (km/h)
A	>= 65
B	>= 60
C	>= 55
D	>= 50
E	>= 35 - 50[a]
F	< 35 - 50

[a] The exact speed at which capacity occurs varies with the percentage and length of grade, traffic compositions, and volume; computational procedures are provided to find this value.

TABLE 5 Factors at Three Levels Used in Economic Feasibility Analyses

FACTOR	LEVELS		
	LOW	MEDIUM	HIGH
Flow rate (mixed vehicles/hour)	0	1000	2000
Percent of trucks	0	25	50
Percent of traffic on upgrade	25	50	75
Percent of grade	0	4	8
Length of grade (m)	400	1700	3000
Percent of no-passing Zones	0	50	100

Occasionally, in grades shorter than 3000 m, the length of grade cannot be defined in some cases since the beginning or end of the grade cannot be identified adequately. In these cases, as well as in all others in which one or more of the input variables of Equation 4 cannot be defined, the speed gain obtained from adding the climbing lane should also be evaluated by using the equation in Figure 2. Evidently, this equation is less precise than Equation 4, as it does not consider separately the effect of the different variables that affect the speed gain.

NOTE ON AMERICAN WARRANTS FOR CLIMBING LANES

The criteria used in the United States to justify economically the construction of a climbing lane (*2*) are based on a more limited number of variables than those presented in this paper. For this reason, it is considered that such criteria are simpler though less precise than the ones described herein; for example, the U.S. criteria do not consider that the construction cost of the extra lane may vary within a very wide range (depending on the existing pavement width and earthwork quantities) and that its exact value may significantly affect the economic feasibility.

POTENTIAL USE OF CLIMBING LANES IN MEXICO

The procedures described earlier allow the assessment of the potential use for climbing lanes in Mexico. When additional construction of the existing pavement width is not needed, these facilities are feasible in all specific grades with average annual daily traffic (AADT) greater than 3,000; in Mexico, 30 percent of the federal trunk roads fall under this condition. As more widening of the existing pavement is required, the accompanying construction costs restrict the feasibility of such facilities to road segments with more severe traffic and geometric conditions. In Mexico, it is considered that these facilities improve operation for AADTs of up to 6,000. For vehicular flows higher than this, the operation has been observed to become unsafe (*4*).

SUMMARY AND CONCLUSION

This work is one of the tools that the IMT has developed for analyzing the economic feasibility of constructing climbing lanes on Mexican two-lane roads and for assessing the LOS before and after implementing these facilities, partly or totally, in specific grades.

This paper represents an effort to generate highway engineering procedures for Mexico and to adapt to conditions in Mexico the most important criteria developed in other countries. In Mexico, the implementation of climbing lanes is considered as a first step in the process of providing four-lane roads on the most important national freight corridors.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to the IMT for providing the financial support to carry out this research project.

REFERENCES

1. *Mexican Government Program for the Modernization of the Transportation System* (in Spanish). Secretaría de Comunicaciones y Transportes, México, 1988.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
3. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
4. *A Manual for the Geometric Design of Roads* (in Spanish). Secretaría de Comunicaciones y Transportes, México, 1971.
5. Chavelas, P. Two-Lane Roads, the Mexican Experience (in Spanish). *Proc., 1st Seminar on Capacity of Highways and Streets*, Asociación Mexicana de Ingeniería de Transporte, A.C., México, 1989.
6. Hoban, C. J., et al. *A Model for Simulating Traffic on Two-Lane Roads, User Guide and Manual for TRARR Version 3.2.* ATM 10B. Australian Road Research Board, Victoria, Feb. 1991.
7. López, M. C., and F. A. Cerquera. Capacity and Level of Service Analysis for Colombian Two-Lane Roads (in Spanish). *Proc., 6th Pan-American Congress of Transportation Engineering*, Popayán, 1990.
8. Adler, H. *Economic Appraisal of Transport Projects: A Manual with Case Studies.* World Bank, Washington, D.C., 1987.

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

Comparison of Uncongested Speed-Flow Relationships Using Data from German Autobahns and North American Freeways

FRED L. HALL AND WERNER BRILON

The use of speed-flow data from German Autobahns provides a greater range of free-flow speeds than that available from North American freeways. On the basis of these additional data, three conclusions address current questions about the speed-flow relationship. First, the speed at capacity is not independent of free-flow speed but is higher on facilities on which the free-flow speed is higher; speed approaching capacity is as high as 90 to 100 km/hr on one Autobahn. Second, the range of flows over which average speeds remain constant (at the free-flow speed) decreases with increasing free-flow speed. For example, with a free-flow speed of 135 km/hr, speeds remain constant out to flows of 1,000 passenger car units per hour per lane. Third, for the range of flows in which speeds decrease with increasing flow, the relationship between speed and flow appears to be linear. This means that there is no need to postulate a curved (e.g., quadratic) relationship near capacity, since there is not an increasing rate of decline in speeds at those flows.

The purpose of this paper is to investigate the nature of the speed-flow relationship using data from German Autobahns. Because the range of free-flow speeds on North American freeways is fairly small, free-flow speed data from the Autobahns, many sections of which have no speed limit, can be valuable in resolving three issues that North American data have not been able to answer. The first is the speed at capacity while operations remain uncongested: does it vary with free-flow speed? The second is the flow rate at which speeds begin to decrease from free-flow speeds: does this vary with free-flow speed? The third is the nature of the decrease in speeds for flows exceeding the flow-rate breakpoint: is it linear or quadratic?

The structure of the paper is as follows. The first section provides the context for the issues by providing a brief review of the current status of North American thinking. The second section summarizes the relevant German literature about speed-flow relationships on Autobahns. Because that literature is unable to resolve the issues, the third section introduces analyses of Autobahn data from two different locations. The analyses do not answer all of the questions, but they are able to shed some light on the issues. The final section draws together the results of the analyses in the form of recommendations for resolving some of the open questions about the speed-flow relationships for freeways in North America.

CONTEXT

The 1985 *Highway Capacity Manual* (HCM) (*1*) is heir to the parabolic shape of the speed-flow curve for freeways that appeared in the 1965 HCM (*2*), in that speed at capacity is shown to be roughly half of the free-flow speed for the facility. The shape of the parabola in the 1985 HCM is broader than that in the 1965 HCM, with speeds remaining at or near the free-flow value until quite high volumes, but then speeds fall off precipitously to the low value at capacity. Preparation of a revised Chapter 7 of the HCM (*3*) brought this representation into question, in that the new data collected for multilane rural roads showed that speeds at capacity were only 8 to 10 km/hr lower than the free-flow speeds. In addition, capacity was found to be higher than the 2,000 passenger cars per hour per lane (pcphpl) given in the 1985 HCM for both freeways and multilane rural roads. Both of these results implied that in some respects a multilane rural roadway operated more effectively than a freeway, which clearly did not make sense. Consequently, efforts began to revise the freeway chapter of the HCM to better reflect current conditions in the light of new data.

The new data available were summarized by Hall et al. (*4*), who also proposed a new generalized speed-flow curve for freeways based on those data. However, their proposal was only a representation of the general shape and did not provide the level of detail needed to replace the 1985 HCM curves. In particular, the HCM figure provides curves for three design speeds (70, 60, and 50 mph) and, within the 70-mph design speed, for four-, six-, and eight-lane facilities. The studies reviewed by Hall et al. do not provide the necessary details, either. The magnitudes of speed drops (i.e., the difference between the free-flow speed and speed at capacity) vary from almost nothing in some studies (*5*) to as much as 25 percent of the free-flow speed in others (*6*). An alternative hypothesis was that 80 km/hr is the speed at capacity, as found in several studies.

With the exception of a few older freeways in the northeastern United States, there is not much variation in the design speeds of North American expressways. Given the 55-mph (90-km/hr) speed limit in the United States on urban freeways (where there is the best chance of finding capacity operations) and the 100-km/hr limit in Ontario, it is not likely that a wide range of free-flow speeds will be found in North American data, except perhaps as might reflect local speed enforcement patterns. To some extent this is visible in the studies summarized by Hall et al. (*4*). The apparent free-flow speeds in those studies ranged from slightly above 100 km/hr (60 mph) (*7,8*) through 100 km/hr (*9*) and 95 km/hr (*6*) to perhaps 85 km/hr ["more than 50 mph" (*10*)]. However, the scatter in the results appears to be greater than the range of free-flow speeds, making it difficult to recommend different curves for different speeds on the basis of these studies.

GERMAN STUDIES

Because of the absence of any speed limit on many sections of the Autobahns, German data provide an opportunity to investigate

F. L. Hall, Department of Civil Engineering, McMaster University, Hamilton, Ontario L8S 4LZ Canada. W. Brilon, Lehrstuhl für Verkehrswesen, Ruhr Universität, Bochum D-44780 Germany.

speed-flow behavior under potentially different free-flow conditions. This section reviews German publications on this topic.

The earliest depiction to consider is the polygon curve from the Green Paper (*11*). This polygon contains three segments for uncongested conditions, which is the part of the speed-flow curve of interest. The resulting polygon can be envisioned as a piecewise linear approximation of the parabolic curve that appeared in the 1985 HCM. However, the polygon was for the most part a generalized curve, without specific numbers assigned to it. The first (leftmost) segment implied a decrease in speeds with each additional vehicle, although only a modest decrease. After the first breakpoint, speeds dropped more sharply. In the third segment (after the second breakpoint) speeds declined even more steeply, to capacity operation.

In a 1990 publication, Heidemann and Hotop (*12*) undertook to specify current parameters for this polygon by fitting it to recent data. They provide a variety of curves, depending on the number of lanes per direction (two or three) and the truck percentage. Their conclusion for a three-lane directional road (i.e., a six-lane freeway) with trucks making up less than 15 percent of traffic is represented as the top line in Figure 1. They found that there was a range of flow over which speeds did not decrease [extending to about 1,200 vehicles per hour (vph) over three lanes, or 400 vph per lane (vphpl)], so they added another segment to the lefthand end of the polygon. However, their data did not extend beyond about 4,000 vph. The part of the curve beyond 4,000 vph was extrapolated by analogy from two-lane data, in order to identify the last breakpoint. The righthand segment of their curve is therefore shown as a dotted line, just as in their paper.

The second representation to be discussed in Figure 1 comes from the publication *Richtlinien für die Anlage von Strassen, Teil: Querschnitte* (RAS-Q), which translates roughly as *Guidelines for the Design of Highways: Profiles* (*13*). In an appendix on level of service (Nachweis der Verkehrsqualität), graphs are included showing a portion of the speed-flow diagram for different roadway cross sections, grades, and percentage of trucks. The small segment in Figure 1 labeled as RAS-Q replicates in its entirety the curve for a six-lane divided freeway, for a continuous grade less than or equal to 1 on which heavy trucks are able to maintain speeds in excess of 70 km/hr (which could include 0 grade), for traffic with 0 percent trucks (roadway type a6ms, Steigungsklasse 1, Lkw = 0 percent).

The third curve in Figure 1 comes from a different part of the *Richtlinien für die Anlage von Strassen*, namely, RAS-W for *Wirtschaftlichkeitsundersuchungen*, or "Economic Assessment" (*14*). The curve shown in the figure is derived from the following equation:

$$V = \{136.5 - 8[\exp(0.235s)]\}[\exp(-10^{-3}\mathrm{KU}) - 0.5 \exp 10^{-3}(Q_p + 2Q_{gv})] \quad (1)$$

where

V = speed of passenger cars,
s = gradient (%),
KU = degree of curvature of roadway section,
Q_p = passenger car rate of flow, and
Q_{gv} = goods vehicles (trucks) rate of flow.

In Figure 1, the curve has been drawn for $Q_{gv} = 0$, $s = 0$, and KU = 0 (i.e., for a straight, level section with no trucks).

The fourth and final curve in Figure 1 is the most interesting. Based on recent data from a number of locations, it is an effort by Stappert and Theis (*15*) to update the curve that lies behind the short segment from RAS-Q. The underlying functional form that they used is the same as that used there, namely, a monotonically decreasing exponential function of the form

$$V = [A - \exp(\mathrm{BQ})]\,[\exp(-c) - k \exp(dQ)] \quad (2)$$

where

V = velocity,
Q = flow, and
c, d = constant Krummungs factors taking values between 0.2 and 0.003.

On the basis of fitting curves to data for nine sites for Autobahns with three lanes per direction, Stappert and Theis came up with the general curve for this road category, shown in Figure 1. There are

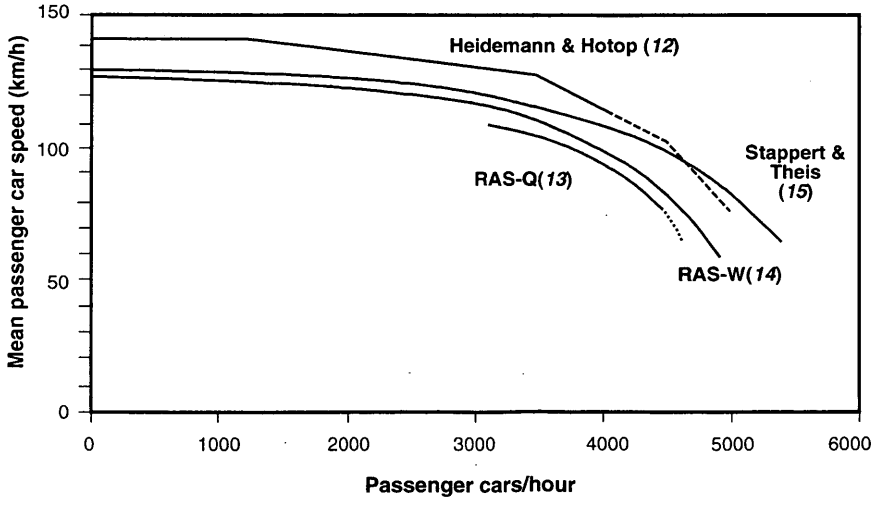


FIGURE 1 Summary of speed-flow relationships in German publications for six-lane freeways (Autobahns with three lanes per direction).

three important comments about their curve, however. The first is that it appears they used only the functional form specified previously and did not compare that function with other possibilities.

The second point is that they too had very little data above a flow of 4,000 vph for the three-lane roads. Their report contains diagrams for four of the nine three-lane sites. There are eight data points above 4,000 vph in those diagrams. Although it is possible that the sites not graphed in their report had higher flows, it appears more likely that they would have included in the report the figures showing highest flows. For the eight "high" flow data points, the mean speed is above 110 km/hr. There is no basis in the data for the 55-km/hr speed shown at a capacity of 5,500 vehicles over the three lanes, nor for the value of 5,500 itself. Those values represent an extrapolation on the basis of the assumed functional shape.

The third point is one that pertains to both the analyses by Stappert and Theis and by Heidemann and Hotop. The data that they used were all hourly counts. For lower flows, the hourly counts can give a good representation of the curve, since operations within a particular hour are likely to be on the same segment of the graph. However, for higher flows there is a good chance that a full-hour count will include data from several segments of the curve (including potentially congested operations) with the result that it will not accurately represent any one part of the curve but instead will average several types of behavior. Figure 2 provides an example of how this can occur. Successive 5-min speed-flow observations are shown, along with the moving average hourly values based on them. There is a brief period of congested 5-min data, which leads the hourly values to appear in parts of the graph where no 5-min operations occurred. Depending on the exact time selected to begin hourly observations, the hourly data may or may not reflect real operations. This point is important for interpreting the German studies, because it is likely that the underlying functional forms that both empirical studies were trying to fit were themselves originally derived on the basis of hourly data.

One of the four-lane freeways (i.e., two lanes each way) for which a graph of data is included in the Stappert and Theis study provides some useful data for the question of speeds at capacity, as well as for per-lane capacity values. On the A43 at Herne, 29 observations exceed a flow of 4,000 vph, with one point above 4,700 vph. All but three of these 29 observations have speeds above 80 km/hr, and 13 are above 90 km/hr. The lowest is 65 km/hr, but the mean would appear to be near 85 km/hr. Hence, there is evidence from the Autobahns that hourly flows in excess of 2,000 vphpl are observed repeatedly and that speeds at these flows are well above the 50 or 60 km/hr suggested by the design guide curves.

In addition to the studies summarized in Figure 1, there is one other recent empirical German study to note, conducted on Highway B10 in Karlsruhe (*16*). The section from which the data come is an urban freeway, roughly 1 km downstream of a cloverleaf interchange and 1 km upstream of a diamond interchange. Figure 3 shows the results obtained in that study, which are not in accordance with any of the curves in Figure 1. In particular, these data show a steady, albeit small, linear decrease in speeds as flow increases, beginning at the lowest observed flows. However, given the nature of the data, all that can be said about capacity is that it is greater than or equal to 3,500 vph (over two lanes), and speed at capacity is likely to be less than or equal to 80 km/hr.

In summary, then, although the conventionally used speed-flow curves for Germany show capacities lower than 2,000 vphpl, and speeds near 50 km/hr at those flows, there are very few data near capacity in the published reports. What data there are at high flows show speeds considerably higher than those portrayed by the German curves, consistent with the recently proposed curves based on Canadian and U.S. data. The data at medium flows are also consistent with the proposed American curves—that is, they show little if any decrease in speed as flow increases (with the exception of the Karlsruhe data). However, the published results in Germany are not adequate to answer the issues raised at the start of this paper.

NEW DATA FROM GERMAN AUTOBAHNS

In an effort to resolve those issues, data from two Autobahn measurement locations have been analyzed. Because data on operations upstream of these points are not available, one cannot be certain that capacity has been reached in the data. Nonetheless, flow values have been observed that are sufficiently high to warrant some tentative conclusions about the behavior of speeds in the vicinity of capacity. In each of the two subsections that follow, the data collection location is described, followed by a description of the analyses that were

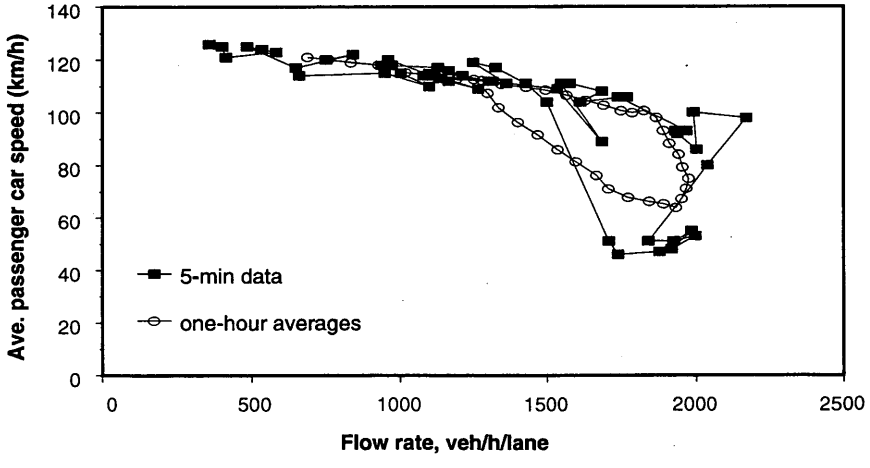


FIGURE 2 Effect of using hourly average data when there is congestion: Moving average hourly data versus 5-min observations, data from A60.

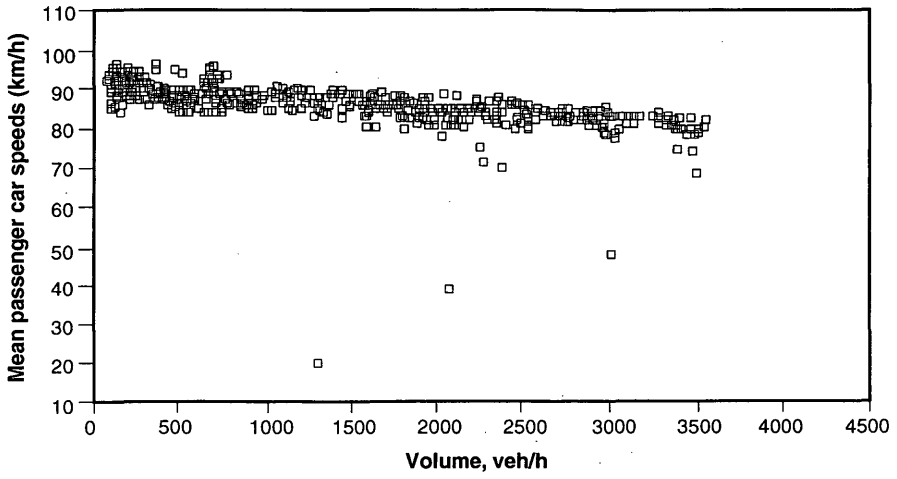


FIGURE 3 Speed-flow data from four-lane urban freeway in Karlsruhe (*16*).

performed. Conclusions from the two analyses appear in the final section of the paper.

A3 at Heusenstamm

The A3 near Heusenstamm, south of Frankfurt, has three lanes in each direction and no speed limit. The measurement location, at km 84, is more than 2 km from any entrance or exit ramps. Because it is sometimes asserted that German Autobahns have a much higher percentage of trucks than do North American freeways, Figure 4 is included to show the daily pattern of traffic volumes together with the truck percentages (for the westbound traffic, i.e., toward Frankfurt). The data in Figure 4 are the actual 5-min counts. There is a fairly steep morning peak in the traffic flow, which is the main focus of the investigation that follows. During this peak, truck percentages fall to their lowest level during the day and are comparable to North American peak-period values. At other times of day, especially at night, the truck percentages are much higher.

Three days of data were used, for May 29–31, 1990. There are two ways to investigate the data, one based on the German procedure of stratifying the analysis by truck percentage, the other based on the U.S. procedure of converting to passenger car units by means of a passenger-car-equivalent truck factor. Figure 5 is based on the German procedure and shows the mean speed of all vehicles versus volume per lane for trucks less than or equal to 15 percent of the traffic volume. This figure suggests a flat segment out to about 600 vphpl, followed by a linearly decreasing function.

The numerical analysis, however, has been based on the use of truck equivalence factors. Figure 6 shows the data for the morning peak period (5:00 to 10:00 a.m.) for these three days together. Peak-period data were used because most North American capacity analyses have relied on peak period data. In converting from vehicles per hour to passenger cars per hour, a truck equivalence factor of 2.0 was used, since that is the accepted German value, as indicated in Equation 1. Three functional forms were attempted (Table 1): linear, quadratic, and piecewise linear. All performed well, in the sense that the R^2 was quite high (above .75 in all cases) and the root-mean-square (*RMS*) error reasonably low. The quadratic function stands out as being better than the linear, but the coefficient on the first-order term is positive, which is counter-intuitive. None of the piecewise linear models had any slope on flow for the first segment

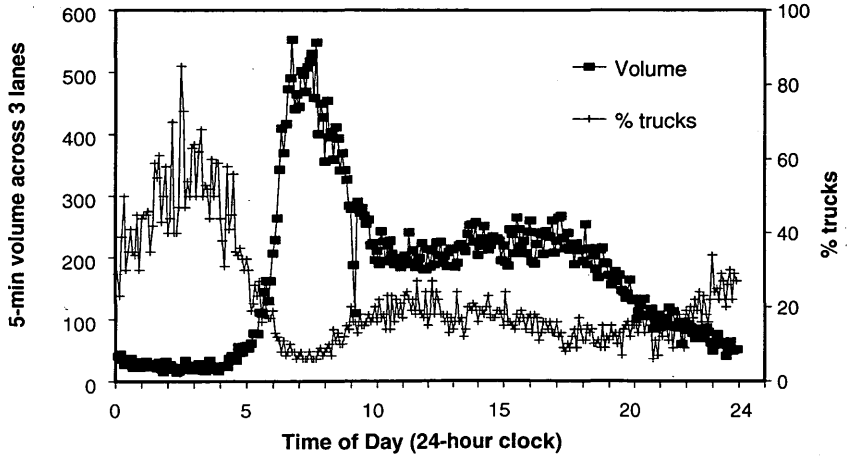


FIGURE 4 Total vehicles and percent trucks versus time of day; from A3 near Heusenstamm, traffic toward Würzburg, May 30, 1990.

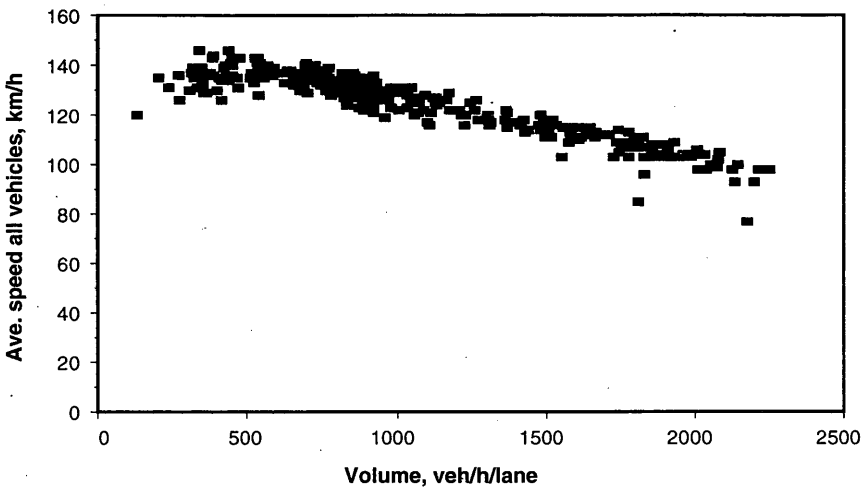


FIGURE 5 Speed-flow data from A3 near Heusenstamm for trucks less than or equal to 15 percent of volume.

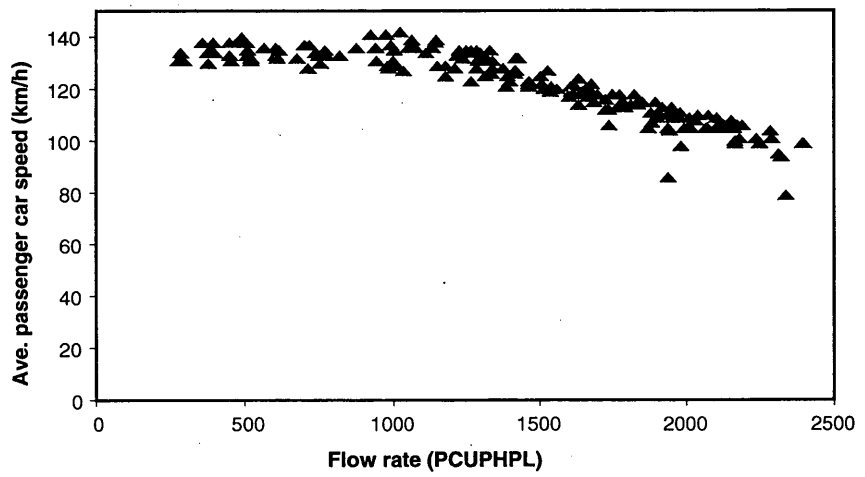


FIGURE 6 Speed-flow data from A3 near Heusenstamm, for three days of morning peak-period data, with flow converted to passenger car units.

of the line, and some performed slightly better than the quadratic. Hence it seems fair to prefer one of the piecewise linear models. The one using a breakpoint of 1,100 vehicles has the highest R^2, but the differences in the R^2 and *RMS* values among several breakpoints were so small that it would be more appropriate to say there is really no difference among several possible breakpoints. The selection of 1,000 vph should probably be made because it is a number that implies approximation, which would be appropriate. The free-flow speed for the equation is 134 km/hr.

A60 near Ginsheim

The A60 near Ginsheim, between Frankfurt and Mainz, has two lanes in each direction and no speed limit. Data were available for the same three days as for the A3. The traffic pattern over the day is similar to that shown in Figure 4 for the A3, so it is not shown here.

Again, presenting the data in the German fashion (Figure 7) suggests a segment with 0 slope out to a flow of perhaps 600 vphpl, followed by a linearly decreasing segment out to capacity. (The nine data points with speeds below 75 km/hr should not be included in the estimation of the function for uncongested data. They are either observations within a queue or queue discharge data.)

Quantitative analysis of the speed-flow relationship was again concentrated on the morning peak period (5:00 to 10:00 a.m.), converting trucks to passenger car units (Figure 8). The first part of Table 2 is based on all of the data in Figure 8. (The nine congested points have been omitted from the figure.) In general, the functions do not fit quite so well as they did for the A3 data: maximum R^2 values are down by about 0.05. In addition, the nature of the piecewise linear function appears to be changing between breakpoints of 400 and 900. The fact that the best R^2 values are associated with quite high flows (1,300 to 1,800 pcuphpl) implies that the data near capacity are affecting the function. Hence, all data above 1,500 pcuphpl (the optimum breakpoint) were deleted, and the

TABLE 1 Results of Analysis of Functions, A3 Data Westbound

Function type	R^2	RMS Error	Equation (all coeff. sig at .00001)	
Simple linear	0.7777	6.09	150 -0.0203 Q	
Quadratic	0.8840	4.41	130 +0.0157 Q - 0.000014 Q^2	
Piecewise linear Breakpoint	(for Q< breakpoint, D=0; for Q =,> breakpoint, D=1)			
400	0.8096	5.65	133 + 20 D	- 0.0222 QD
500	0.8249	5.42	135 + 21 D	- 0.0235 QD
600	0.8466	5.07	134 + 24 D	- 0.0252 QD
700	0.8644	4.77	134 + 27 D	- 0.0267 QD
800	0.8892	4.31	134 + 32 D	- 0.0293 QD
900	0.8928	4.24	134 + 33 D	- 0.0298 QD
1000	0.8963	4.17	134 + 35 D	- 0.0307 QD
1100	0.9000	4.09	134 + 37 D	- 0.0318 QD
1200	0.8993	4.11	134 + 38 D	- 0.0323 QD
1300	0.8986	4.12	134 + 37 D	- 0.0316 QD

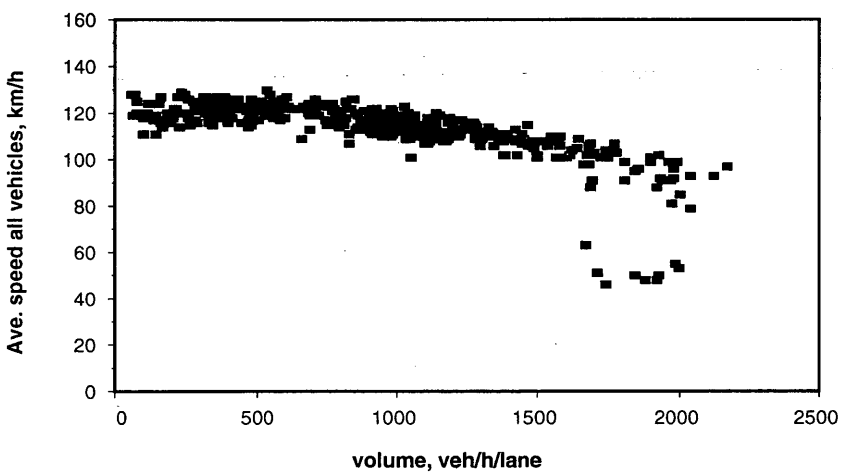


FIGURE 7 Speed-flow data from A60 near Ginsheim, for trucks less than or equal to 15 percent of volume.

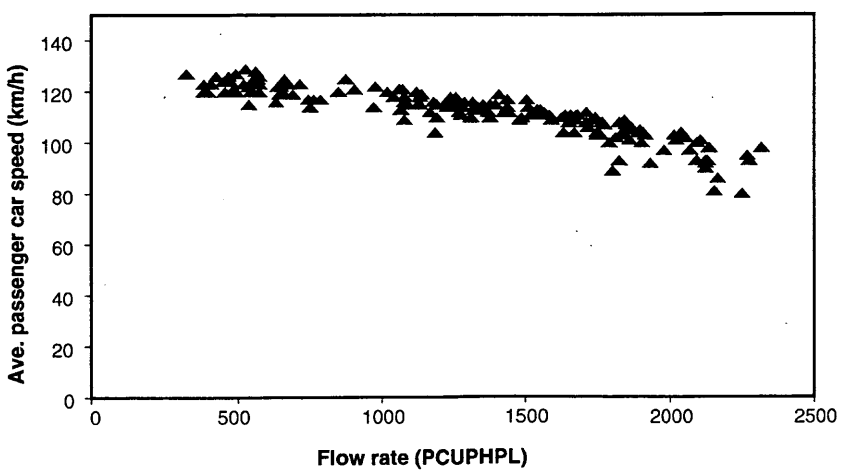


FIGURE 8 Speed-flow data from A60 near Ginsheim, for three days of morning peak-period data, with flow converted to passenger car units.

TABLE 2 Results of Analysis of Functions, A60 Data Eastbound

Function type	R^2	RMS Error	Equation (coeff. sig at .0001 unless noted)
Simple linear	0.7817	4.48	133 -0.0160 Q
Quadratic	0.8329	3.91	124 - 6.328 $(Q/1000)^2$
Piecewise linear Breakpoint	(for Q< breakpoint, D=0; for Q =,> breakpoint, D=1)		
400	0.7879	4.43	134 - 0.0304Q + 0.0140 QD (sig at 0.026)
500	0.7937	4.37	123 + 12 D - 0.0171 QD
600	0.7988	4.31	123 + 13 D - 0.0180 QD
700	0.8064	4.23	123 + 16 D - 0.0193 QD
800	0.8191	4.09	133 + 7 D - 0.0205 Q
900	0.8190	4.10	128 + 13 D - 0.0098 Q - 0.0110 QD (sig at .043) (sig at .027)
1000	0.8201	4.09	127 + 14 D - 0.0093 Q - 0.0118 QD (sig at .017) (sig at .004)
1100	0.8238	4.05	128 + 15 D - 0.0103 Q - 0.0117 QD
1200	0.8318	3.96	129 + 18 D - 0.0115 Q - 0.0124 QD
1300	0.8359	3.91	128 + 22 D - 0.0109 Q - 0.0149 QD
1400	0.8428	3.82	128 + 27 D - 0.0110 Q - 0.0176 QD
1500	0.8436	3.81	128 + 30 D - 0.0109 Q - 0.0192 QD
1600	0.8426	3.83	128 + 30 D - 0.0107 Q - 0.0191 QD
1700	0.8417	3.84	129 + 30 D - 0.0112 Q - 0.0189 QD
1800	0.8367	3.90	129 + 28 D - 0.0119 Q - 0.0174 QD (sig at .001)
Excluding data above 1500 pcuphpl			
linear	0.6043	3.17	128 - 0.0109 Q
breakpoint			
400	0.6079	3.17	122 + 6 D - 0.0113 DQ (sig at 0.001)
500	0.6065	3.17	123 + 6 D - 0.0115 DQ
600	0.6109	3.15	123 + 4 D - 0.0106 DQ (sig at 0.008)
700	0.6010	3.19	123 + 5 D - 0.0104 DQ (sig at 0.044)
800	0.5728	3.31	122 + 10 D - 0.0129 DQ (sig at 0.004)

analysis run again. In this case, R^2 values dropped even further, and the quadratic function dropped out entirely. (Neither of its coefficients was significant.) The piecewise linear function with a breakpoint at 600 is a viable candidate, but so is the simple linear function. Free-flow speed is either 128 km/hr (linear function) or 123 km/hr (piecewise linear, 700).

CONCLUSIONS

These results provide a positive indication that capacity varies with the free-flow speed of the facility. The A3 data suggest a value between 90 and 100 km/hr for a free-flow speed of 134 km/hr (on a six-lane roadway). The A60 data (for a four-lane road) suggest values above 90 km/hr (before queue discharge effects arise) for a free-flow speed of 123 km/hr. The data discussed by Stappert and Theis for the four-lane A43 at Herne suggested a speed of 85 km/hr for a free-flow speed of 110 km/hr. The higher values are associated with higher free-flow speeds. Thus speed at capacity may not be independent of free-flow speed. The implication for speed-flow curves in the HCM is to support the new curves in Chapter 3 (*17*) which show different speeds at capacity for different free-flow speeds. Although the effect appears to be present in these German data, some studies referred to earlier provide contradictory results. In particular, the works by Persaud and Hurdle (*6*) and Hurdle and Datta (*9*) contain high speeds at capacity even though they do not have particularly high free-flow speeds.

With regard to the flow rate at which speeds begin to decrease from free-flow speeds, the German data suggest that the range of flows over which speed is constant can vary, depending in part on how the data are analyzed. Figures 5 and 7 suggest only a short range of constant speeds, out to about 600 vphpl when speed is averaged over all vehicles. The analyses in Tables 1 and 2 suggest a larger range when flow is converted to passenger cars and speed is

averaged over only passenger vehicles. Nevertheless, there are some clear indications even in these analyses that higher free-flow speeds are associated with a lower value of flow for the breakpoint than in North American studies. Heidemann and Hotop's curve in Figure 1 has a breakpoint at about 400 pcphpl for a free-flow speed of 143 km/hr. Table 1 suggests a breakpoint of about 1,000 pcphpl for a free-flow speed of 134 km/hr. The lowest breakpoint in the new Chapter 3 for the HCM is 1,300 pcphpl, for a free-flow speed of 70 mph (115 km/hr). One interpretation is that the constant speeds seen across a wide range of flows on North American freeways are probably an artifact of the presence of a speed limit that is considerably below the speed at which drivers could travel comfortably. This interpretation receives some support from Heidemann and Hotop's analysis (*12*) of sections of the Autobahn with speed limits, in that they show the constant speed segment continuing to higher volumes when there is a posted speed limit. For example, on a six-lane road with a speed limit of 80 km/hr, the constant speed segment extends to 1,300 pcphpl.

These data have not been adequate to resolve the final issue raised at the start of this paper, namely, the nature of the decrease in speeds for higher flows. The analyses with the Autobahn data (especially for the A60) suggest that a linear function in this range is entirely adequate and that there is no steeper decrease in speeds at the highest observed flows. However, if the highest observed flows are not at capacity, then it remains possible that there is a steeper decline in the last few hundred vehicles per hour of flow, which the data in Figure 8 suggest. Nevertheless, given the large range of flows with decreasing speeds (from 500 to 2,000 vphpl), a linear function appears reasonable.

In summary, then, the analysis of German data supports the general picture proposed by Hall et al. for the uncongested portion of the speed-flow curve and adds some detail to the general picture in a way that is consistent with the depiction in the recently approved version of Chapter 3 of the HCM (*17*). The one difference between these data and the depiction of speed-flow relationships on freeways in the new Chapter 3 of the HCM is that only a linear trend was observed in these German data, although it is possible that capacity flows did not occur in the data. Certainly these few sites are not enough to settle the matter decisively, but they do provide useful confirmation for the new Chapter 3 curves, which in several key aspects appear to have been developed with minimal empirical support.

ACKNOWLEDGMENTS

The authors are pleased to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada, and of the Deutsche Forschungsgemeinschaft, through the bilateral exchange agreement between those two agencies. In addition, the authors would like to thank Michael Grossman for his assistance in summarizing the relevant German literature.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
3. Committee A3A10, Subcommittee on Multilane Highways, Chapter 7: Capacity and Level of Service Procedures for Multilane Rural and Suburban Highways. TRB, National Research Council, Washington, D.C., 1990.
4. Hall, F. L., V. F. Hurdle, and J. H. Banks. Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992, pp. 12–18.
5. Wemple, E. A., A. M. Morris, and A. D. May. Freeway Capacity and Level of Service Concepts. In *Highway Capacity and Level of Service* (U. Brannolte, ed.), Balkema, Rotterdam, the Netherlands, 1991, pp. 439–455.
6. Persaud, B. N., and V. F. Hurdle. Some New Data That Challenge Some Old Ideas About Speed-Flow Relationships. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988, pp. 191–198.
7. Chin, H. C., and A. D. May. Examination of the Speed-Flow Relationship at the Caldecott Tunnel. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 75–82.
8. Hall, F. L., and K. Agyemang-Duah. Freeway Capacity Drop and the Definition of Capacity. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 91–98.
9. Hurdle, V. F., and P. K. Datta. Speeds and Flows on an Urban Freeway: Some Measurements and a Hypothesis. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 127–137.
10. Banks, J. H. Flow Processes at a Freeway Bottleneck. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 20–28.
11. Bundesminister für Verkehr (BMV). Abhangigkeit zwischen Streckenkennwert, Verkehrsstärke und Verkehrsgeschwindigkeit. (sog. "Grüne Blätter") Anlage zum *Gesetz über den Ausbau der Bundesfernstrassen in den Jahren 1971–1985 (FStrAbG), 2. Überprüfung des Bedarfsplanes.* Aachen, Germany, 1978.
12. Heidemann, D., and R. Hotop. Verteilung der Pkw-Geschwindigkeiten im Netz der Bundesautobahnen—Modellmodifikation und Aktualisierung. *Strasse und Autobahn*, Heft 3, 1990, pp. 106–113.
13. *Richtlinien Für die Anlage von Strassen, Teil: Querschnitte (RAS-Q).* Forschungsgesellschaft für Strassen und Verkehrswesen, Cologne, Germany, 1982.
14. *Richtlinien Für die Anlage von Strassen, Teil: Wirtschaftlichkeitsundersuchungen (RAS-W).* Forschungsgesellschaft für Strassen und Verkehrswesen, Cologne, Germany, 1986.
15. Stappert, K. H., and T. J. Theis. *Aktualisierung der Verkehrsstärken-Verkehrsgeschwindigkeitsbeziehungen des BVWP-Netzmodells. (Actualisation of the v-q Relations for the Federal Transportation Masterplan Network).* Research Report VU-18-0009-V89. Heusch/Boesefeldt, Aachen, Germany, 1990.
16. Brilon, W., and F. Weiser. *Verkehrsplanerische Untersuchung zur Frage eines sechsstreifigen Ausbaus der Rheinbrücke der B10 bei Karlsruhe.* Ruhr University, Bochum, Germany, 1992.
17. *Special Report 209: Highway Capacity Manual,* 3rd edition, revised. TRB, National Research Council, Washington, D.C., Ch. 3 (in publication).

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

Revisions to Level D Methodology of Analyzing Freeway Ramp Weaving Sections

JOHN R. WINDOVER AND ADOLF D. MAY

For ramp weaves on a eight-lane freeway, the total point flow method has been demonstrated to predict point flows in each of the rightmost two lanes more accurately than the Level D methodology. The Level D methodology is one of the current methods used by the California Department of Transportation (Caltrans). Research undertaken at the Institute of Transportation Studies at the University of California, Berkeley, improved the Level D estimates of point flow for weaving sections operating under various flow ranges. The Level D estimate of freeway-to-freeway (FF) percentage in the rightmost through lane was modified to improve the point flow predictions. An equation was developed that predicts the FF percentage as a function of weaving section length, upstream section demand, on-ramp demand, and off-ramp demand. The process involved a calibration effort that compared total point flow and Level D estimates of volumes in the rightmost through freeway lane. From this comparison an FF percentage estimating equation was developed that, when incorporated in Level D, would result in Level D producing volume estimates comparable in accuracy to the total point flow method. It was validated with the empirical data that were used to develop the total point flow method. The FF estimating equation used with Level D produced significant improvements in the point flow prediction. The equation is recommended for inclusion in the Level D methodology incorporated in FRELANE.

The Institute of Transportation Studies at the Berkeley campus of the University of California (ITS-UCB) is developing a computer model, FRELANE, to analyze isolated freeway sections. A ramp weave, a type of simple weaving section, is one of the eight types of sections that FRELANE is capable of analyzing. FRELANE currently applies two methodologies to analyze ramp weaves: the total point flow method and the Level D method. Both methods estimate the total point flow in each of the rightmost two lanes at 152-m (500-ft) intervals. The Level D method was developed for analyzing weaving sections that operate at near-capacity conditions. The total point flow method has been found more accurate in estimating total point flows in a ramp weaving section for a wide range of conditions, including those for which Level D was designed. The Level D methodology is, however, one of the current methods used by the California Department of Transportation (Caltrans) for analysis of ramp weaves. Modifications in the Level D factors are being investigated to revise the Level D method in order to make it an accurate methodology for weaving sections under a wide range of flows. Ultimately, the revisions in the Level D factors will be incorporated into the FRELANE program.

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

DEFINITION OF SIMPLE WEAVING SECTIONS

A simple weave has only one on-ramp and one off-ramp connected by one or more auxiliary lanes. This research concentrates on a simple weave with one on-ramp and one off-ramp connected by a single auxiliary lane, a ramp weave. Figure 1 illustrates the ramp weaving section and terminology used in this report.

The ramp-to-ramp (RR) flow enters the weaving section from the on-ramp and exits by the off-ramp. The freeway-to-freeway (FF) flow enters and exits the weaving section on the mainline freeway. The ramp-to-freeway (RF) flow enters the weaving section from the on-ramp and exits by the mainline freeway. The freeway-to-ramp (FR) flow enters the weaving section from the mainline freeway and exits by the off-ramp.

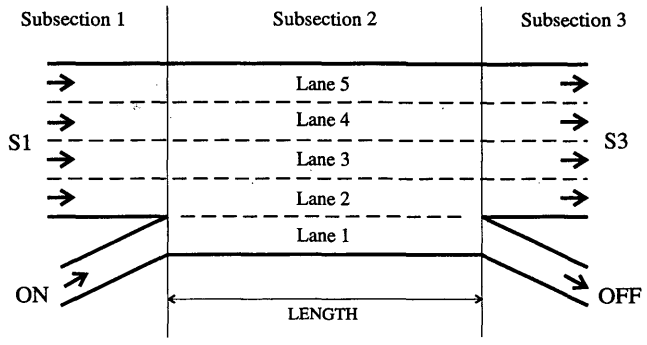
HISTORICAL REVIEW

The 1950 HCM presented the first freeway weaving analysis method, which predicted the capacity and operating speeds of freeway weaving sections (*1*). The 1965 HCM contained a revised version of this 1950 HCM method with added emphasis on quality of flow (*2*). The revised version was based on publications by Norman (*3*), Hess (*4*), and Leisch (unpublished studies, Bureau of Public Roads, U.S. Department of Commerce, 1958–1964). In 1981, TRB published the PINY method and the Leisch method (*5*). The PINY method, developed at the Polytechnic Institute of New York, was a new method for estimating weaving and nonweaving speeds for simple weaving sections (*6*). The Leisch method (*7*) was an enhancement of the nomograph method of the 1965 HCM. The 1985 HCM chapter on weaving analysis predicted the average speeds of weaving and nonweaving vehicles using regression-based equations (*8*).

In 1987 a 6-year research program was initiated at ITS-UCB that was based on the need for additional research on freeway weaving in California, which has produced a number of publications (*9–17*).

CURRENT RESEARCH AND METHODOLOGIES

The first phase of research at ITS-UCB was to evaluate the existing methods of analyzing major weaves. The second phase involved evaluating existing methods for analyzing other types of freeway sections including ramp weaves. The methods evaluated included the speed estimating methods previously listed, the JHK method (*18*), the Fazio method (*19*), and the following flow estimating methods: Caltrans *Traffic Bulletin 4* method (Level D) (*20*) and total point flow method (*21*). The total point flow method was deter-



Ramp Weaving Variables and their Valid Ranges

Weaving Length (LENGTH):	305 m (1000 ft) - 610 m (2000 ft)
Subsection 1 Demand (S1):	4816 - 7200 pcph
On-Ramp Demand (ON):	222 - 1800 pcph
Off-Ramp Demand (OFF):	338 - 1004 pcph
Subsection 3 Demand (S3):	5378 - 7200 pcph
Subsection 2 Demand (S2):	5914 - 9000 pcph
Ramp to Ramp Demand (RR):	4 - 148 pcph
Freeway to Freeway Demand (FF):	4242 - 6500 pcph
Ramp to Freeway Demand (RF):	214 - 1480 pcph
Freeway to Ramp Demand (FR):	330 - 856 pcph
Weave Volume (RF + FR):	950 - 2780 pcph

FIGURE 1 Schematic and variables for ramp weaving section.

mined to produce the best predictions. The next best candidate was the Level D method (*22*).

Ongoing research at ITS-UCB explored the use of the total point flow approach in analyzing freeway weaving sections. The total flow at a point may be estimated directly or found as a sum of the individual movements. A computer model, FRELANE, for predicting traffic performance in weaving sections based on total point flow has been developed from this research. FRELANE uses the predicted point flow at specific locations to calculate the traffic density at these locations. The calculated density is then used to select the appropriate level of service (LOS) for each location. The locations at which the analysis is done in FRELANE include the merge, the diverge, 76 m (250 ft) downstream of the merge, and 152-m (500-ft) increments from the merge to the end of the section. For ramp weaving sections, FRELANE has two methodologies to predict the point flows at the preceding locations in Lanes 1 and 2 along a weaving section: total point flow and Level D.

Total Point Flow Method

The total point flow method, proposed by Holmes, is a regression-based methodology that directly predicts the total flow at an analysis point within a weaving section (*11*). The flow is calculated for each point as a function of length of weaving section, lane being considered, location being considered, mainline freeway flow, RF flow, FR flow, and RR flow. These equations were determined to predict total point flow within 10 percent of the empirical values for 90 percent of the analysis data.

Level D Method

The Level D method was developed by Caltrans in the early 1960s (*20*). The Level D method is appropriate for ramp weaving and non-ramp-weaving sections operating under conditions of high or near-capacity traffic flow. Given the section length and volumes in the weaving section, Level D predicts the point flow as a sum of the individual movements. The point flows are predicted for each of the two rightmost lanes of the freeway weaving section at the same locations as the total point flow method. The point flow 76 m (250 ft) downstream of the merge was not estimated by the Level D method initially but has since been added. The RF and FR percentages in each lane at each location are solely a function of section length. The amount of through traffic in the rightmost through freeway lane (Lane 2) is a function of FF traffic flow and is assumed to be constant throughout the weaving section. The estimates of total flows in Lane 2 are highly sensitive to the estimate of through traffic in the rightmost through freeway lane. The current errors in the estimation of total volumes at points in this lane can be attributed principally to incorrect predictions of FF volumes. The current Level D method predicts the total point flows within 10 percent of the empirical values for 40 percent of the analysis data.

ASSUMPTIONS AND LIMITATIONS OF ANALYSIS

The current Level D method assumes the FF percentage in the rightmost through lane of the weaving section to remain constant along the weaving section. This is also assumed true for this analysis. The

RF and FR percentages in Lane 2 are assumed to be predicted correctly by the Level D methodology.

The entire analysis is limited to a basic four lane one-directional mainline freeway segment. All traffic flows input and calculated are in passenger cars per hour. The estimating equations are considered acceptable for analysis only when all of the input variables are within the range of the empirical data that was used to develop the equations. These ranges are different for the total point flow method and the Level D method. The overlap of these two regions, shown in Figure 1, is the region that was used to develop equations in this analysis.

LEVEL D REVISION METHODOLOGY

Level D required modification in order to improve its accuracy over a wide range of flows. The Level D estimation of the FF percentage in the rightmost through lane was identified as the main factor contributing to the inaccuracies in the Level D estimates in Lane 2. Several approaches were available in attempting to improve the FF percentage estimation in Lane 2. It was decided that a two-step process would be followed, consisting of a calibration effort and a validation effort. In the calibration process, a formulation for estimating FF percentage by the Level D method was based on forcing the Level D method to agree with the total point flow method in terms of total flow at selected points along the rightmost through lane for various flow ranges. The calibration stage assumed that the total point flow estimates were accurate for the various flow ranges in order to derive an equation to estimate an FF percentage that, when incorporated in Level D, would allow the Level D method to accurately predict point flows under various flow ranges. In the validation process, the performance of the derived formulation for estimating FF percentage in Lane 2 was checked. The validation process used the empirical data from four freeway ramp weaving sections used to develop the total point flow method. Therefore, the calibration process derived an equation for estimating FF percentage using the values calculated by the total point flow method over a wide range of flows, and the validation process checked the derived equation using the empirical data, which were the same data that were used to develop the total point flow method.

Calibration Methodology

To improve the FF percentage in Lane 2, three options were considered:

1. Modify the existing FF percentage tables,
2. Use the available data to derive a method for calculating FF percentage, or
3. Simulate new data to derive a method for calculating FF percentage.

The current method uses a table of averages to calculate the FF percentage in Lane 2 on the basis of the through freeway volume (FF) only. A consistent trend between the through freeway volume and the average FF percentages based on the existing empirical data could not be determined, thus a modification of the existing FF percentage table was rejected. Simulation of a data set to estimate FF percentages was deferred because there was no indication that the available data were inadequate to produce accurate results. Therefore, the use of available information was selected to derive new FF percentages for the Level D methodology.

A mathematical formula was derived to calculate the FF percentages needed to improve the current Level D estimates using the existing data. This formula required correct FF percentages for calibration. The FF percentages for calibration could be obtained by using either empirical values or values derived from the total point flow method. The total point flow method was determined from past research to closely replicate reality. The values derived by this method can be considered as valid as the empirical values. The FF percentages derived by the total point flow method also allowed a wider range of combinations of flow conditions than was available using the empirical data. The mathematical equation would then be validated with the empirical data.

Methodology to Regression Equation Derivation

The first phase in deriving an expression for the percentage FF was the development of a set of inputs to estimate volumes by the Level D method and total point flow method. The 23 input data points developed, given in Table 1, covered the widest range of values possible for all input variables and their combinations within the overlapping valid ranges of each estimating method. A wide range of values was desired in order to derive an FF percentage estimating equation that would improve the Level D estimates over a wide range of conditions. Ramp weaving section lengths of 305, 457, and 610 m (1,000, 1,500, and 2,000 ft) were then tested separately using these data sets. The total point flow was calculated at 0, 76, 152, and 305 m (0, 250, 500, and 1,000 ft) along the weaving length for Lane 2. The calculation went to the 305-m (1,000-ft) location to include as much of the information as possible to calculate the FF percentage and to account for as much weaving as possible. Most weaving occurs within the first 152 m (500 ft) of the weaving section, so FF percentage calculations to 305 m (1,000 ft) should cover the majority of the weaving action.

The next phase was to calculate FF percentages for calibrating an FF percentage equation. The assumed FF demand in Lane 2 was calculated by taking the difference between the total point flow calculated by the total point flow method and the Level D method and adding it to the Lane 2 FF demand calculated by the Level D method. A new percentage of FF traffic in Lane 2 was calculated for each location in Lane 2 along the weaving length. The FF percentage required in the Level D method to produce the same total movements as the total point flow method could be one of the following: the average of the FF percentages from 0 to 305 m (1,000 ft), the FF percentage at the critical point (the location with the highest total point flow) along the weaving section, or the FF percentage at 0 m (the merging point) where the FF percentages are calculated in the empirical data.

The total movements calculated by Level D with updated FF percentages were plotted along with total movements calculated by the total point flow method and the current Level D method for each analysis location. These graphs were produced to verify that the updated Level D estimates were a significant improvement in total movement estimation.

Since significant improvements were found in the updated Level D estimates, the next phase was to derive a method to incorporate these revised FF percentages into Level D. A regression equation as a function of the input values entered was chosen over averaging the FF percentages because earlier findings showed that averaging the FF percentages over certain FF volumes would not satisfactorily update the volume-dependent tables already in the Level D method. Regression equations were derived for the following independent

TABLE 1 Input Data Created for Ramp Weaving Section

Data Set	Input Volumes				Calculated Volumes					
	S1	ON	OFF	RR	S2	S3	RF+FR	FF	RF	FR
1	5000	1000	500	125	6000	5500	1250	4625	875	375
2	5000	1600	500	125	6600	6100	1850	4625	1475	375
3	5000	1600	700	125	6600	5900	2050	4425	1475	575
4	5500	1000	500	125	6500	6000	1250	5125	875	375
5	5500	1000	700	125	6500	5800	1450	4925	875	575
6	5500	1000	900	125	6500	5600	1650	4725	875	775
7	5500	1600	500	125	7100	6600	1850	5125	1475	375
8	5500	1600	700	125	7100	6400	2050	4925	1475	575
9	5500	1600	900	125	7100	6200	2250	4725	1475	775
10	6000	400	900	125	6400	5500	1050	5225	275	775
11	6000	1000	500	125	7000	6500	1250	5625	875	375
12	6000	1000	700	125	7000	6300	1450	5425	875	575
13	6000	1000	900	125	7000	6100	1650	5225	875	775
14	6000	1600	500	125	7600	7100	1850	5625	1475	375
15	6000	1600	700	125	7600	6900	2050	5425	1475	575
16	6000	1600	900	125	7600	6700	2250	5225	1475	775
17	6500	400	900	125	6900	6000	1050	5725	275	775
18	6500	1000	500	125	7500	7000	1250	6125	875	375
19	6500	1000	700	125	7500	6800	1450	5925	875	575
20	6500	1000	900	125	7500	6600	1650	5725	875	775
21	6500	1600	900	125	8100	7200	2250	5725	1475	775
22	7000	400	900	125	7400	6500	1050	6225	275	775
23	7000	1000	900	125	8000	7100	1650	6225	875	775

Note: All demands are in passenger cars per hour (pcph).
See Figure 1 for the empirical limits of the above variables.

variables used in various combinations: Section 1 demand (S1), on-ramp demand (ON), off-ramp demand (OFF), FF demand (FF), RF demand (RF), and FR demand (FR). The standard error of the estimate was used to select the best equation containing significant variables.

Ramp Weaving Section Analysis at 457 m

The 457-m (1,500-ft) weaving section was the first section analyzed. The FF percentages were calculated for all the data. The average difference in FF percentage along the section from 0 to 305 m (1,000 ft) was 3.7 percent. The average difference in FF percentage along the weave length was larger than expected. Thus, using the FF percentage at a single point, either at distance 0 or at the critical point, to derive an equation was not considered the way to reflect properly the general trend in FF percentage along the section. Instead, the average FF percentage between 0 and 305 m (1,000 ft) was used to calibrate a regression equation. Figure 2 illustrates that the calculated average FF percentages are higher than the FF percentages used in the current Level D method for the 23 data points. The critical FF percentages are also plotted on this graph. The critical FF percentage and the average FF percentage are effectively interchangeable.

Figure 3 illustrates that the Level D method using the average FF percentages predicted total movements that were very close to the total point flow predictions for the 457-m (1,500-ft) section. The predictions were close for all distances and for a wide range of total point flows—900 to 2,200 passenger cars per hour (pcph). These FF percentages were considered satisfactory to develop a regression equation. An attempt was made to develop a regression equation to estimate FF percentage as a function of FF demand only, which is contained in the 1965 and 1985 HCMs, but there was a very low correlation between FF percentage and FF demand. The following two regression equations were determined to best replicate the average FF percentages for rightmost through lanes calculated for the 23 data points:

$$FF\% = 25.4 - 0.00209(S1) - 0.00512(ON) + 0.0152(OFF) \quad (1)$$

$$FF\% = 26.6 - 0.00208(FF) - 0.00512(RF) + 0.0132(FR) \quad (2)$$

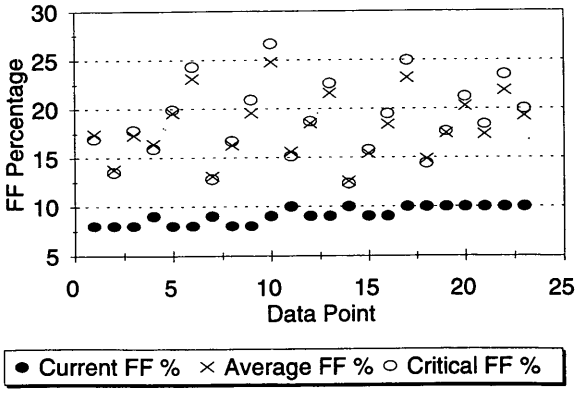


FIGURE 2 FF percentages for 457-m (1,500-ft) ramp weaving section.

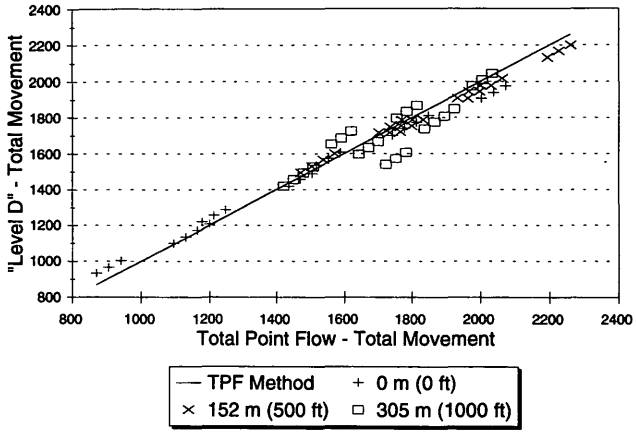


FIGURE 3 Level D method using average FF percentages for 457-m (1,000-ft) section.

The standard error of the estimate was approximately 0.30 for both equations. All the input variables in the Equations 1 and 2 were determined to be statistically significant in the equation by using the *t*-distribution at a 95 percent confidence level. Regression Equation 1 was chosen for further analysis because Section 1 demand, on-ramp demand, and off-ramp demand are variables that are more easily measured directly in the field.

Ramp Weaving Section Analysis at 305 and 610 m

The FF percentages calculated for the 305-m (1,000-ft) weaving section had an average difference in FF percentage along the weaving length of 5.0 percent. Again, the average FF percentage was the most suitable FF percentage to use for further analysis. The graph of the total movements predicted by Level D with these average FF percentages and by the total point flow method was similar to the same graph produced for the 457-m (1,500-ft) section (Figure 3). The following regression equation, with a standard error of 0.35, was determined to best replicate these FF percentages:

$$FF\% = 15.6 - 0.00103(S1) - 0.00619(ON) + 0.0140(OFF) \quad (3)$$

The FF percentages calculated for the 610-m (2,000-ft) weaving section had an average difference in FF percentage along the weaving section of 6.1 percent. Again, an average FF percentage was the most suitable value to use for calibration. The graph of the total movements predicted by Level D with these average FF percentages and the total point flow predicted total moments was similar to the 457-m (1,500-ft) graph. The following equation, with a standard error of 0.28, was determined to best replicate these average FF percentages:

$$FF\% = 35.5 - 0.00322(S1) - 0.00402(ON) + 0.0172(OFF) \quad (4)$$

Equation Extension to All Simple Weaving Section Lengths

The FF percentage equations developed for the three weaving section lengths were determined to be statistically different using a *t*-test at a 95 percent confidence level. Figure 4 shows a pattern of FF percentages increasing consistently as the weaving section length increased. Thus, to produce an FF percentage estimation equation that can be applied to all weaving lengths from 305 to 610 m (1,000 to 2,000 ft), an equation that includes length as a variable was required. To develop this equation, the data points for each of the three weaving lengths already analyzed and the corresponding average FF percentages were combined. A regression analysis was performed on these 69 data points, and the following regression equation, with a standard error of 0.77, was derived:

$$FF\% = 7.92 + 0.0117(LENGTH) - 0.00211(S1) - 0.00511(ON) + 0.0155(OFF) \quad (5)$$

The Level D method using Equation 5 to estimate FF percentage will be referred to herein as the modified Level D method.

VALIDATION

After the general equation (the regression equation for all lengths) was developed on the basis of values calculated by the total point flow method, the next step was to attempt to validate the equation. The data used for validation were the 22 empirical data points used to develop the total point flow method. These data points were obtained at four freeway ramp weaving sections. The first site (eight data points) was eastbound Interstate 580 from Oakland Avenue to Grand Avenue in Alameda County, California, which had a weaving length of 372 m (1,220 ft). The second site (four data points) was southbound I-5 from Palomar Street to Main Street in San Diego County, California, which had a weaving length of 381 m (1,250 ft). The third site (three data points) was eastbound CA-60 from Paramount Boulevard to San Gabriel Boulevard in Los Angeles County, California, which had a weaving length of 418 m (1,370 ft). The fourth site (seven data points) was westbound CA-91 from 183rd Street to Artesia Boulevard in Los Angeles County, California, which had a weaving length of 578 m (1,895 ft) (*22*).

The average weaving volumes at the four sites were 2,388 pcph at the first site, 1,145 pcph at the second, 615 pcph at the third, and 1,043 pcph at the fourth. The weaving volume is the combined RF and FR traffic flow. Thus the first site, which is operating at near-capacity conditions, has an average weaving volume that is more than twice as high as the average weaving volumes observed at the other sites.

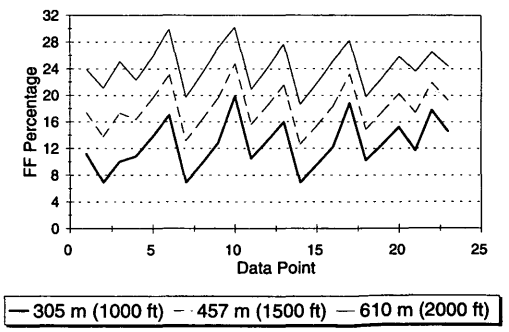


FIGURE 4 Average FF percentages for three ramp weaving sections.

Testing of Developed Equation

The overall performance of the modified Level D method was first determined by comparing the accuracy of the modified Level D total point flow estimations to the current Level D and total point flow method estimates. The FF percentages used in the current Level D method can be found in Table 5-3 of the 1985 HCM. The accuracy of these methods was determined by calculating the average residual of each method's estimates of total point flow for the 22 empirical data points. The merge point in Lane 2 was the location used for this validation effort. The current Level D method had an average residual of 339 pcph, the total point flow method had an average residual of 62 pcph, and the modified Level D method had an average residual of 89 pcph. Therefore, the modified Level D estimates of total point flow were on average 250 pcph closer to the empirical value than the current Level D estimates, for the 22 empirical data points. The modified Level D was comparable in accuracy to the total point flow method, but the total point flow method was slightly more accurate than the modified Level D method.

The validation process also compared each estimate of total point flow by the current Level D method and by the modified Level D method with the empirical value for the 22 empirical data points on a site-by-site basis. The results of this comparison are illustrated in Figure 5. Figure 5 showed that the modified Level D method is predicting total point flows closer to empirical values for Sites 2, 3 and 4, which are not operating at near-capacity conditions. For Site 1, which is operating under near capacity conditions, the current Level D estimates were closer. The current Level D method was designed for sections near capacity, thus reasonable estimates by the current Level D method were expected for this first site. The current Level D method total point flow estimates were generally too low for the other sites, Sites 2, 3 and 4, which were not operating close to capacity. The modified Level D method over estimated the point flows for near-capacity Site 1. However, the modified Level D method reasonably estimated the total point flow for Sites 2, 3 and 4, which were not operating at near-capacity conditions. Therefore, the modified Level D method produced reasonable estimates for all operating conditions with a tendency to overestimate flows for weaving sections operating near capacity.

Validation of Ordinary Least Squares Assumption

The ordinary least squares assumption was also checked using residual plots. The residual plots showed that the ordinary least squares assumption was reasonable. However, the variances exhibited some site dependency, which implied that a factor was probably missing from the general equation.

CONCLUSION

For ramp weaves on an eight-lane freeway, the total point flow method had been demonstrated to predict point flows more accurately than the current Level D methodology, which is one of the methods used by Caltrans. This analysis determined that the overall accuracy of Level D can be improved by modifying the Level D estimation of FF percentage in the rightmost through lane. The FF percentages currently used in the Level D methodology were determined to be consistently low during both the calibration and validation stages of this analysis. The following regression equation

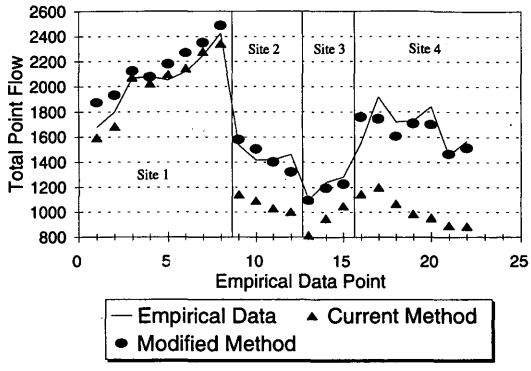


FIGURE 5 Comparison of current and modified Level D methods.

was generated to improve the current Level D estimation of FF percentage, in order to improve the current Level D total point flow estimation:

$$FF\% = 7.92 + 0.0117(LENGTH) - 0.00211(S1) - 0.00511(ON) + 0.0155(OFF)$$

The modified Level D method, using the generated regression equation, increased the current Level D methods accuracy on average by 250 pcph. Thus, the modified Level D method showed a significant improvement in accuracy of estimating total point flow. The modified Level D method did show a tendency to overestimate the total point flow for near-capacity weaving sections. A conservative estimation of total point flow for near-capacity weaving sections was acceptable. Thus, the modified Level D method, which uses the generated Regression Equation 5, is recommended for adoption in the FRELANE model to improve the Level D predictions of point flow.

FUTURE RESEARCH

The RF and FR percentages in the rightmost through lane, which currently depend on only the distance along the weaving section length, appeared to be volume-dependent also. The next phase of research is to determine if RF and FR percentages are dependent on the traffic movements in the weaving section. The calibration of the RF and FR curves was beyond the scope of this project.

ACKNOWLEDGMENTS

This research was sponsored by Caltrans and FHWA. The authors wish to thank Barbara Ostrom for her advice and help throughout this research as well as James Holmes and Lannon Leiman.

REFERENCES

1. *Highway Capacity Manual.* Bureau of Public Roads, U.S. Department of Commerce, 1950.
2. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965.
3. Norman, O. K. Operation of Weaving Areas. *Bulletin 167.* HRB, National Research Council, Washington, D.C., 1957, pp. 38–41.
4. Hess, J. W. *Traffic Operation in Urban Weaving Areas.* Bureau of Public Roads, U.S. Department of Commerce, 1963.

5. *Circular 212: Interim Materials on Highway Capacity*. TRB, National Research Council, Washington, D.C., Jan. 1980.
6. Roess, R., et al. *Freeway Capacity Analysis Procedures*. Final Report. Polytechnic Institute of New York, May 1978.
7. Leisch, J. E. *Capacity Analysis Techniques for Design and Operation of Freeway Facilities*. Report FHWA-RD-74-24. FHWA, U.S. Department of Transportation, 1974.
8. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
9. Cassidy, M. J., A. Skabardonis, and A. D. May. *Operation of Major Freeway Weaving Areas: Recent Empirical Evidence*. Working Paper UCB-ITS-WP-88-11. Institute of Transportation Studies, University of California, Berkeley, Dec. 1988.
10. Chan, P., M. J. Cassidy, and A. D. May. *Operation of Freeway Ramps and Multiple Weaving Sections; Research Proposal and Candidate Sites*. Working Paper UCB-ITS-WP-90-4. Institute of Transportation Studies, University of California, Berkeley, Sept. 1990.
11. Holmes, J. R., D. Preslar, D. Thompson, L. Leiman, B. K. Ostrom, and A. D. May. *Simulation and Empirical Relationships in Developing a Freeway Lane Model*. Technical Document UCB-ITS-TD-92-2. Institute of Transportation Studies, University of California, Berkeley, 1992.
12. Leiman, L., B. K. Ostrom, and A. D. May. *FREWEV: A Design and Analysis Model for Major Freeway Weaving Sections: User's Guide*. Research Report UCB-ITS-RR-92-5. Institute of Transportation Studies, University of California, Berkeley, March 1992.
13. Ostrom, B. K., L. Leiman, and A. D. May. *FREWEV: A Design and Analysis Model for Major Weaving Sections: Data and Assumptions*. Technical Document UCB-ITS-TD-92-1. Institute of Transportation Studies, University of California, Berkeley, 1992.
14. Leiman, L., B. K. Ostrom, and A. D. May. *An Analysis Model for Major Freeway Weaving Sections and Empirically Limited Analysis of Selected Freeway Segments: User's Guide*. Working Paper UCB-ITS-WP-93-3. Institute of Transportation Studies, University of California, Berkeley, June 1993.
15. Ostrom, B. K., L. Leiman, and A. D. May. *Suggested Procedures for Analyzing Freeway Weaving Sections*. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993.
16. Cassidy, M. J., P. Chan, B. Robinson, and A. D. May. *A Proposed Analytical Technique for the Design and Analysis of Major Freeway Weaving Sections*. Research Report UCB-ITS-RR-90-16. Institute of Transportation Studies, University of California, Berkeley, 1990.
17. Robinson, B. W., M. A. Vandehey, G. D. Mazur, and A. D. May. *Improved Freeway Analysis Techniques: Ramp and Weaving Operations for Freeway Lane Model*. Research Report UCB-ITS-RR-92-4. Institute of Transportation Studies, University of California, Berkeley, March 1992.
18. Reilly, W. R., J. H. Kell, and P. J. Johnson. *Weaving Analysis Procedures for the New Highway Capacity Manual*. Offices of Research and Development, Traffic Systems Division, FHWA, U.S. Department of Transportation, 1984.
19. Fazio, J. *Development and Testing of a Weaving Operational Analysis and Design Procedure*. M.S. thesis. University of Illinois, Chicago, 1985.
20. Moskowitz, K., and L. Newman. *Traffic Bulletin 4: Notes on Freeway Capacity*. Division of Highways, California Department of Public Works, Sacramento, July 1962.
21. Holmes, J. R., D. Preslar, D. Thompson, L. Leiman, B. K. Ostrom, and A. D. May. *Simulation and Empirical Relationships in Developing a Freeway Lane Model*. Technical Document UCB-ITS-TD-92-2. Institute of Transportation Studies, University of California, Berkeley, 1992.
22. Fong, H. K., and F. D. Rooney. *Weaving Areas Near One-Lane Ramps*. Report FHWA-CA-TO-TOS-90-1. California Business, Transportation, and Housing Agency; FHWA, U.S. Department of Transportation, Sept. 1990.

The contents of this paper reflect the views of the authors and do not necessarily reflect the official view or policies of Caltrans or FHWA.

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

Proposed Analytical Technique for Analyzing Type A Weaving Sections on Frontage Roads

VICTOR E. FREDERICKSEN AND MICHAEL A. OGDEN

The analysis of nonfreeway or slow-speed weaving sections is documented. Previous research in this area has been limited almost exclusively to freeway weaving sections. Specifically, Type A weaving areas on frontage road facilities with ramps on the left side were evaluated. Special consideration should be given to both the length of the section and the number of lanes when designing the geometrics of a weaving section. Access points such as driveways can also have a significant effect on traffic operations within these sections. It was determined by previous weaving studies as well as this research that speed was not an adequate measure of effectiveness because of its insensitivity to volume. Two additional measures of effectiveness were studied: density and lane changing intensity (LCI). Density was also eliminated because of its relationship at constant speeds. Models were developed to predict LCI using three levels of service. These models require only the identification of geometric conditions and traffic volumes to predict lane change operations. The evaluation of performance measures for the LCI model found it to be an effective means of nonfreeway weaving analysis. This methodology is also consistent with the approach used in the 1985 *Highway Capacity Manual.*

A weaving section is formed when a merge area is followed closely by a diverge area. Weaving is defined by the 1985 *Highway Capacity Manual* (HCM) as "the crossing of two or more traffic streams traveling in the same direction along a significant length of highway, without the aid of traffic control devices" (*1*).

Weaving sections have unique operational characteristics and require special design consideration. In the past, weaving section research has concentrated almost exclusively on freeway weaving sections. Consequently, methodologies for analyzing weaving sections do not provide adequate means for analyzing nonfreeway or slow-speed weaving sections. A procedure is needed for analyzing frontage road and arterial weaving sections.

A typical Type A frontage road weaving section is shown in Figure 1. A Type A weaving section is defined in the 1985 HCM as requiring "that each weaving vehicle make one lane change in order to execute the desired movement" (*1*). Weaving occurs between the merge and diverge points of the section and can be affected by several factors. These factors include lane balance through the section, lane widths, lane configuration, section length, speed limits on the frontage road and ramps, and shoulder widths.

Research at the Institute of Transportation Studies at the University of California, Berkeley, has shown that current weaving section analysis methods are not reliable (*2,3*). This is primarily due to the use of speed as a performance measure. Speed has been found to be insensitive to other traffic factors and is therefore difficult to predict. Measures of effectiveness considered for this project were density and lane changing intensity (LCI).

One of the first methods for analyzing the operations and design of freeway weaving sections was published in the 1950 edition of the HCM (*4*). This procedure was based on an empirical analysis of data collected before 1948. The U.S. Bureau of Public Roads initiated an effort in 1953 that resulted in a new method for the analysis and design of freeway weaving sections and was published in the 1965 HCM (*5*).

The Polytechnic Institute of New York (PINY) developed a methodology that was published in *NCHRP Report 159* (*6*) in 1976. The PINY procedure was found to be difficult to apply because of its complexity and therefore was not widely accepted as a useful methodology. A modified PINY procedure was presented in TRB's *Circular 212* (*7*) in 1980 to simplify the structure of the procedure.

Circular 212 also contained a procedure previously published in the *ITE Journal* (*8*). This method, developed by Leisch, was similar in structure to the 1965 HCM procedure and used two nomographs: one for two-sided configurations, and one for one-sided configurations.

FHWA sponsored a project from 1983 through 1984 to compare the PINY and Leisch procedures and make recommendations for a procedure to be included in the 1985 HCM. This study, conducted by JHK & Associates (*9*), concluded that neither method was adequate for analyzing operations of freeway weaving areas. The study proposed a method consisting of two equations: one for the prediction of the average speed of weaving vehicles, and one for the prediction of the average speed of nonweaving vehicles.

NCHRP Project 3-28B in 1984 recalibrated equations similar to those in the JHK method for the prediction of weaving and nonweaving speeds in weaving sections for the three basic types of configurations and for constrained and unconstrained operations. The result was a procedure consisting of 12 calibrated equations that was subsequently approved by TRB's Committee on Highway Capacity and Quality of Service and included in the 1985 HCM (*1*).

Fazio and Raiphail (*10*) revised the JHK method by using an increased amount of calibration data and introducing a new "lane shift" variable into the speed equations. This variable represents the minimum number of lane shifts that must be executed by the driver of a weaving vehicle from his lane of origin to the closest destination lane.

Researchers at the Institute of Transportation Studies began a study in 1987 that examined six existing methods for the design and analysis of freeway weaving sections. The study found that the existing models did not accurately predict weaving and nonweaving speeds and that speed was insensitive to changes in geometric and traffic factors over the range of values in the data set used. The

V. E. Fredericksen, Wilbur Smith Associates, 908 Town & Country Boulevard, Houston, Tex. 77024. M. A. Ogden, Texas Transportation Institute, Texas A&M University, College Station, Tex. 77843.

FIGURE 1 Typical frontage road weaving section.

study suggested that average travel speed is not an ideal measure of effectiveness (*2*).

Cassidy and May developed a new analytical procedure for the capacity and level of service (LOS) for freeway weaving sections that uses prevailing traffic flow and geometric conditions to predict vehicle flow rates in critical regions within the weaving section. Predicted flows are then used to assess the capacity sufficiency or LOS of a weaving area (*3*).

The Center for Transportation Studies and Research at the New Jersey Institute of Technology published a report in 1991 in which a model for analyzing weaving areas under nonfreeway conditions was proposed. The proposed model consisted of equations for predicting weaving and nonweaving speeds similar to those used in the 1985 HCM (*11*).

STUDY DESIGN AND METHODOLOGY

The objective of this project was to develop a method of analyzing Type A weaving areas on collector-distributor and frontage road facilities that is both reasonably accurate and simple to use. This procedure should define a measure of effectiveness and take into consideration weaving and nonweaving volumes, weaving section length and width, and any intermediate disturbances within the weaving section. To accomplish this, it was necessary to establish a data base to provide operational and physical information needed to formulate a method for analyzing the weaving sections. Another objective was to provide some general guidelines for the design of weaving sections on collector-distributor and frontage road facilities.

Data Collection and Analysis

Data for this study were collected in two phases. Phase 1 consisted of the data used to formulate the proposed models, and Phase 2 consisted of the data used to test the proposed models.

Data Requirements

Data collection activities for this study included traffic volume, vehicle classification, lane changing activity, speed, density, and weaving section geometry. All operational data were collected by personnel at the Texas Transportation Institute using video recording equipment. The weaving section geometry was obtained from roadway plans and field measurements.

Study Site Selection

Data were collected at eight sites in Texas (Table 1). The sites were chosen using the following criteria:

- Weaving sections should be less than 457 m (1,500 ft) in length from gore point to gore point, preferably less than 305 m (1,000 ft), and
- Intermediate disturbances such as intersections and driveways should be minimal.

Originally, only study sites in the Houston area were to be considered for this project. However, not enough sites were found in Houston, and it was necessary to use sites in other Texas cities. Several sites were chosen in Austin and the Dallas–Fort Worth area. The two Houston sites were not used for the Phase 1 analysis of the study because of reconstruction activities in the area. Data from these two sites were collected after the completion of the reconstruction activities and used in the Phase 2 analysis for model testing purposes.

RESULTS

Once the required traffic data were collected, the appropriate operational data were extracted directly from the videotape documen-

TABLE 1 Nonfreeway Weaving Study Sites

LOCATION	PHASE	CITY	LANES	WIDTH m (ft)	LENGTH. m (ft)
IH35 SB-FR @ Felix	I	Ft. Worth	3	11 (36)	136 (447)
SH360 SB-FR @ Green Oaks	I	Arlington	4	13 (44)	142 (467)
IH820 WB-FR @ Wichita	I	Ft. Worth	4	15 (48)	184 (604)
US75 SB-FR @ Midpark	I	Dallas	4	13 (44)	230 (755)
US75 NB-FR @ Spring Valley	I	Dallas	4	13 (44)	256 (841)
IH35 NB-FR @ Riverside	I	Austin	4	15 (48)	335 (1100)
US59 SB-FR @ Beechnut	II	Houston	3	11 (36)	293 (960)
US59 NB-FR @ Fondren	II	Houston	3	11 (36)	342 (1120)

tary. These data were summarized in 5-min intervals. This time interval was used to increase the sample size. All large vehicles traveling on the weaving sections were converted to passenger car equivalents according to procedures for freeways given in the 1985 HCM (*1*). Data sets with average flow rates of fewer than 200 vehicles per hour were excluded, as the focus of this project was on operations at higher volumes. From the eight sites, 335 data points were obtained.

Volumes of traffic entering the weaving sections and volumes of weaving vehicles were measured from the videotaped data. Densities were also obtained directly from the videotapes by counting the number of vehicles in a weaving section at a given time, as opposed to calculating densities on the basis of speeds and volumes. This was done by pausing the videotape every 10 sec, recording the densities for each lane, and averaging the readings to obtain a density value for each 5-min period. Average speeds were calculated by two methods, the first by using the stopwatch feature on the video camera to determine the time it takes a vehicle to travel a known distance, and the second by dividing the average volumes by the average densities.

It was not possible at most locations to obtain speeds via the first method described. The direct measurement method was instead used to verify the average speed calculations for the volume/density method. Lane changes were counted directly from the videotaped data. All lane changes within the entire weaving section were counted and summed for each 5-min period; these values were then converted to lane changes per hour per mile per lane. Weaving section lengths were measured between the painted gore points.

Data Verification

The accuracy of the data used to develop and calibrate the weaving models was a vital aspect of this project. Approximately 10 percent of the data were extracted from the videotape a second time to serve as an accuracy check. Any data sets with discrepancies of more than 5 percent were extracted a third time. Only one data set was found to have any discrepancies in the density values and was therefore extracted a third time. In many instances, total movements (i.e., ramp to frontage road, frontage road to frontage road, frontage road to ramp, and ramp to ramp flows) could be compared with lane changing activity data.

Fundamental Relationships

Before a model was developed for analyzing weaving section performance, the relationships between speed, flow, density, and LCI were examined to gain a better understanding of the operational characteristics of weaving sections. Frontage road flow rates in the weaving sections are generally limited by the intersection capacities upstream of the weaving sections. Each of the weaving sections in this study was preceded by an upstream traffic signal: consequently, the flow rates are lower than those on a freeway weaving section. Vehicle platooning significantly affects the operational characteristics of frontage road weaving sections. However, this study did not attempt to quantify this effect.

Speed-Flow Relationships

Relationships between speed and volume were studied initially. Average flow rates per lane were used to normalize the weaving section volumes, and speeds were obtained from the videotaped data by calculating speeds from the volume and density data. A scatter plot of average speed versus average flow is illustrated in Figure 2 (pcphpl = passenger cars per hour per lane). Aggregated 5-min observation data from the six Phase 1 study sites were used to construct the scatter plot.

Figure 2 reveals a high degree of scatter among the data. Speed appears to be insensitive to flow for the flow rates measured (e.g., fewer than 600 vehicles per hour per lane). There is less scatter at higher volumes, however, indicating that speed may be somewhat sensitive to flow as it nears capacity. From the data collected, no

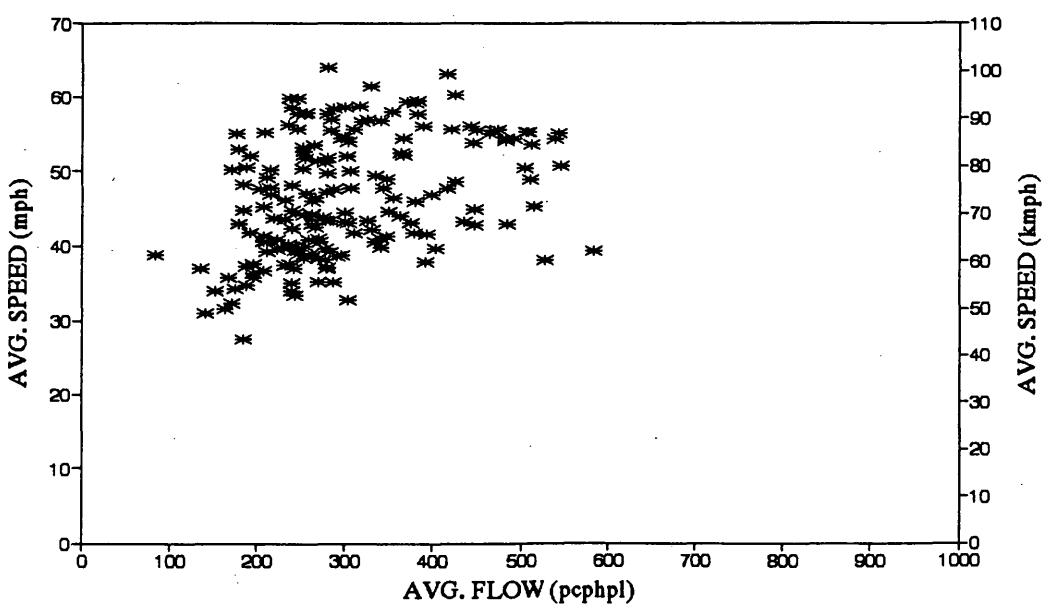


FIGURE 2 Speed versus flow.

obvious relationship between speed and flow was found, supporting the conclusions of other weaving studies that speed is not an adequate performance measure.

Density-Flow Relationships

Relationships between density and volume were also examined. Densities were measured directly from the videotaped data over the length of each weaving section. Figure 3 illustrates the density-flow relationship using average densities and average flows for 5-min periods (vpmpl = vehicles per mile per lane, vpkmpl = vehicles per kilometer per lane). There is much less scatter among the density-flow data than the speed-flow data. This is due partly to volume being contained in both axes of the plot. Density appears to be sensitive to flows, although the scatter increases at higher flows.

There is a conceptual flaw in the relationship between density and flow, however. For a given weaving section, the average speeds are nearly constant until traffic flows approach the capacity level. In this study, traffic flows for the weaving sections studied did not approach capacity. This resulted in density values consisting of volumes divided by an essentially constant speed. In this case (generally uniform speeds), the plot of density versus flow is the same as flow versus flow, which would obviously be a strong linear relationship. It was determined that a model for predicting densities on the basis of flow would not be the most effective procedure for predicting traffic operations in weaving areas on frontage roads.

LCI-Volume Relationships

In previous weaving studies (2), LCI was suggested as a possible measure of effectiveness, but none of these studies developed this concept. LCI is a more direct measure of the turbulence experienced within a weaving section than speed; it can be expressed as the number of lane changes per hour per mile per lane, as shown in the following equation:

$$\text{LCI} = \frac{\text{number of lane changes per hour}}{(\text{number of lanes})(\text{length of weaving section})} \quad (1)$$

LCI was found to be sensitive to flow. The data were stratified for different lengths of weaving sections to improve the relationship as illustrated by the degree of scatter in the data and represented by the coefficient of correlation, r^2. The data were separated by weaving section length into three groups; the first, 122.0 to 182.6 m (400 to 599 ft); the second, 182.9 to 274.1 m (600 to 899 ft); and the third, 274.4 to 365.9 m (900 to 1,200 ft). Scatter plots for each weaving section group are illustrated in Figures 4, 5, and 6 (lcphpmpl = lane changes per hour per mile per lane, lcphpkmpl = lane changes per hour per kilometer per lane).

Proposed Models for LCI Prediction

A linear model was constructed for each of the three weaving section length groups using a regression program. These models estimate the LCI in a frontage road weaving section on the basis of the average volume per lane. The three LCI models, developed from 5-min observation data, are listed here:

$$\text{122.0–182.6 m (400–599 ft): LCI} = 10.46\,(V/n) + 372 \quad (2)$$

$$\text{182.9–274.1 m (600–899 ft): LCI} = 8.52\,(V/n) + 79 \quad (3)$$

$$\text{274.4–365.9 m (900–1200 ft): LCI} = 391\,(V/n) + 590 \quad (4)$$

where

LCI = lane changes per hour per lane per mile (to convert to kilometers, divide by 0.621),
V = hourly volume entering weaving section, and
n = number of lanes in weaving section.

The coefficient of correlation (r^2) is a measure of how much of the variability of the dependent variable, LCI in this case, is

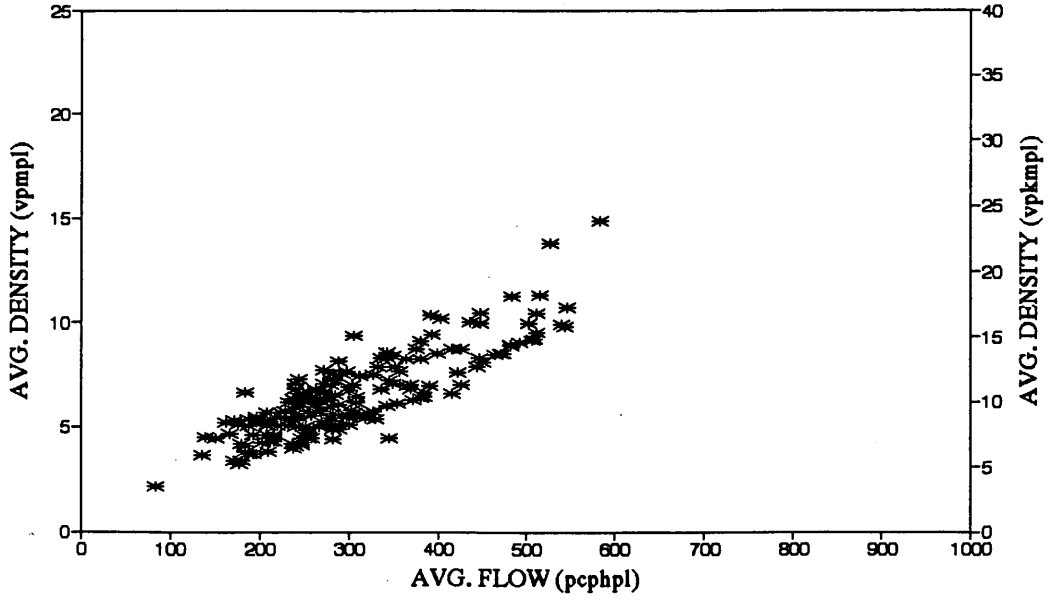


FIGURE 3 Density versus flow.

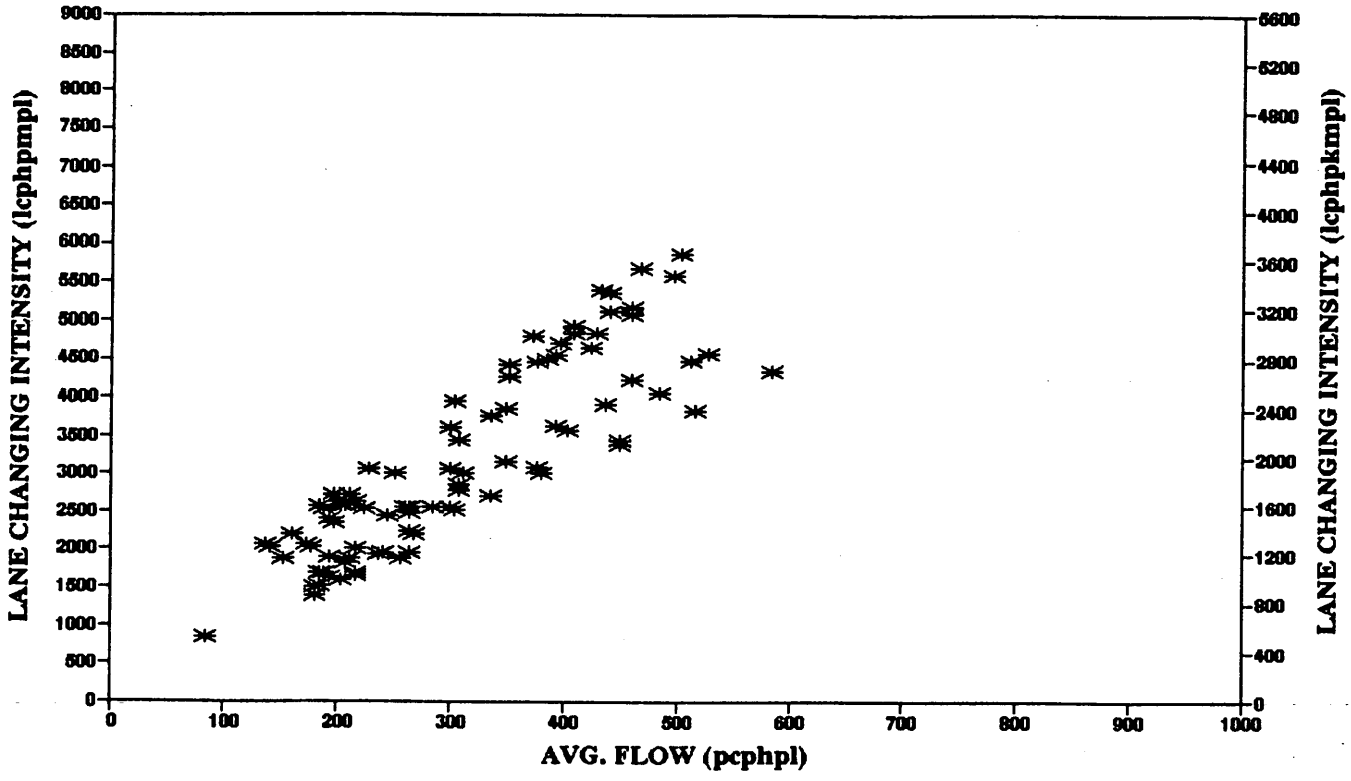


FIGURE 4 LCI versus average flow, 122.0 to 182.6 m (400 to 599 ft).

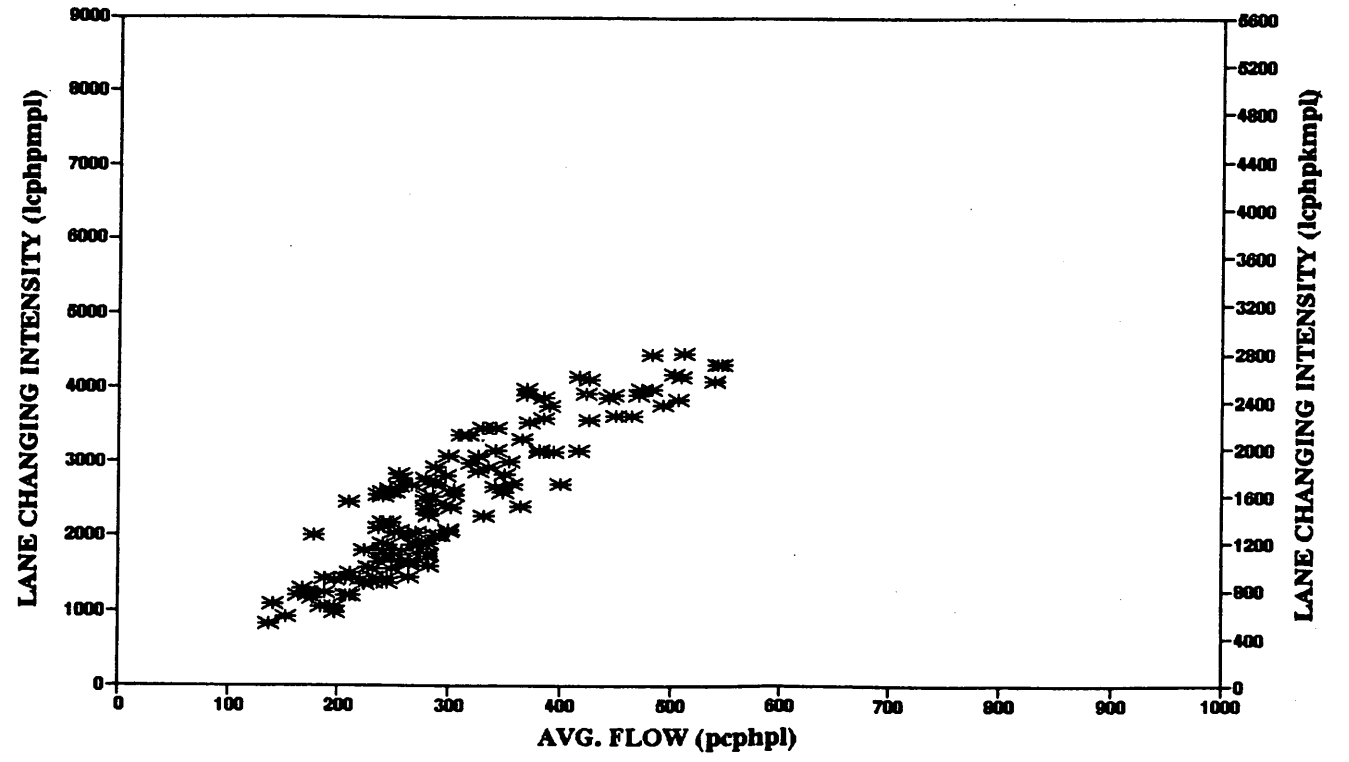


FIGURE 5 LCI versus average flow, 182.9 to 274.1 m (600 to 899 ft).

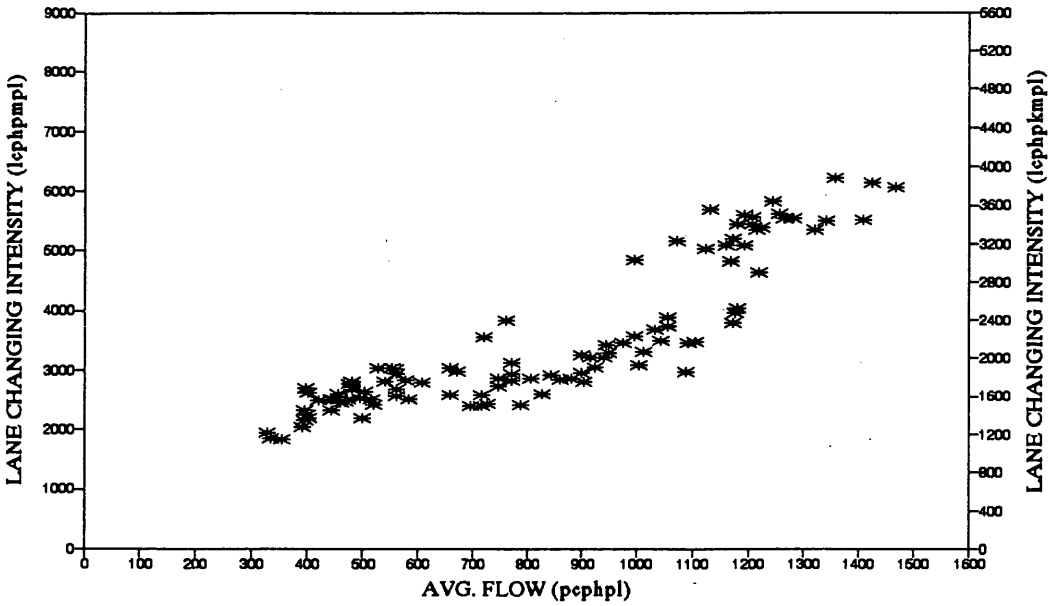


FIGURE 6 LCI versus average flow, 274.4 to 365.9 m (900 to 1,200 ft).

explained by the variability of the independent variable, average volume in this case. A value of +1.00 or −1.00 is perfect, and a value of 0.00 is the lowest possible. The adjusted r^2 value for Equation 2 is 0.94, the adjusted r^2 value for Equation 3 is 0.78, and the adjusted r^2 value for Equation 4 is 0.82. The three LCI equations are shown graphically in Figure 7.

The LCI models were developed using the Jandel Scientific Curve Table Software Package, and the analysis of variance was performed using the Statistical Analysis Software Package (SAS). A linear equation was chosen for each model for simplicity and because there were no obvious patterns in the data that suggested that the relationships might be nonlinear. It is possible, however, that as traffic operations near capacity, the relationships will become nonlinear. The equations each have a constant associated with them because the relationship between volume and LCI is not known as volume approaches 0. Although it is intuitively obvious that each model should begin at the origin, it is possible that the relationship is nonlinear at very low volumes. The models presented in this paper should be used only for the volume ranges shown in Figure 7.

Model Testing

Data were collected for Phase 2 at two sites in Houston for the purpose of testing the LCI models. The two weaving sections were in the range of 274.4 to 365.9 m (900 to 1,200 ft) and thus were applic-

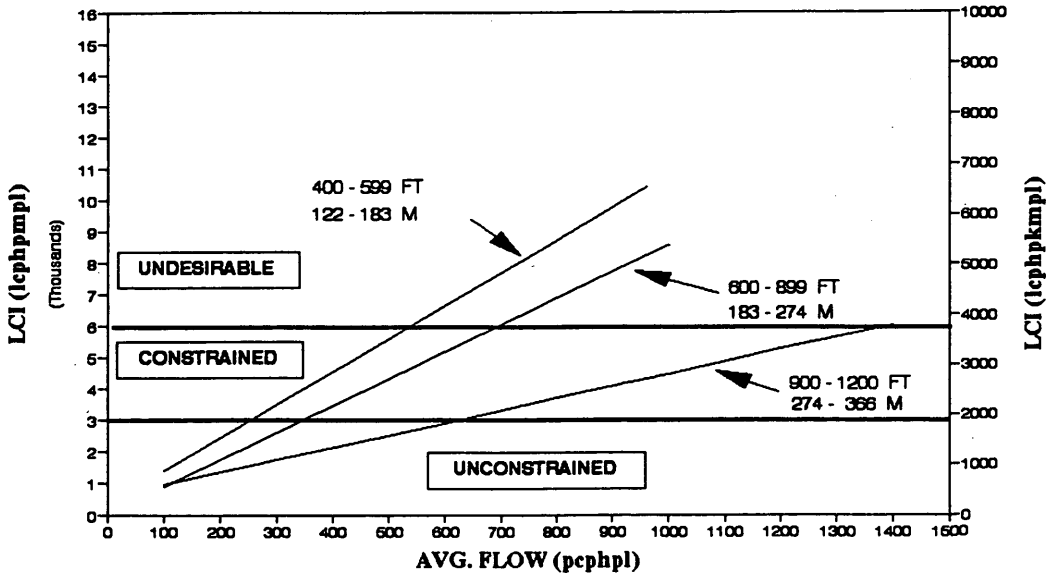


FIGURE 7 Proposed LCI models.

able to only one of the three models. Attempts to locate Phase 2 weaving sections in the Houston area to test the models for shorter weaving sections were unsuccessful.

The testing procedure consisted of a statistical analysis of the data collected at the two Phase 2 test sites by comparing the values observed for LCI with those predicted by the model. The two test sites experienced higher volumes than any of the original study sites, thereby enabling the boundaries of the model to be tested at these higher volumes. The data collected at the two test sites compared favorably with the predicted LCI values from the model and the other study sites. The adjusted r^2 value for this test was .75, indicating that the model is reasonably accurate. This result indicates a more reliable method than the previously mentioned current methods used to predict performance in weaving sections.

LOS Estimation

The criteria for determining LOS were developed to be consistent with those in the 1985 HCM, but with some differences. The 1985 HCM describes six levels (A through F). The criteria proposed in this paper have only three levels—unconstrained, constrained, and undesirable—because of the difficulty in differentiating between six levels over the range of data. It can also be argued that six separate levels do not exist. The criteria proposed in this paper can be compared to 1985 HCM criteria as follows:

- Unconstrained: A and B,
- Constrained: C and D, and
- Undesirable: E and F.

The unconstrained LOS represents free to stable flow conditions in which individual behavior is relatively unaffected by other traffic, and comfort and convenience levels are high. The constrained LOS represents a stable flow condition in which individual behavior is significantly affected by others and may become restricted. Comfort and convenience levels are noticeably lower. The undesirable LOS represents flow conditions approaching capacity in which comfort and convenience levels are poor and breakdowns in flow may occur with small changes in volume. The average speeds under these conditions would also be noticeably lower. The proposed LOS criteria are presented in Table 2 and shown graphically in Figure 7.

The values given in Table 2 were selected subjectively by viewing the videotaped data and identifying the periods in which each LOS was represented. The LCIs were determined at each LOS for all the weaving sections, and an average value was selected to represent each LOS boundary. This method of selection is subjective, and these values do not represent exact divisions in LOS. These values are intended to provide a general idea of what can be expected at a given weaving section. For example, in Figure 7, weaving sections greater than 274.4 m (900 ft) long reach the undesirable LOS at relatively high volumes. This suggests that at lengths greater than 274.4 m (900 ft), weaving is not a major concern on frontage roads. This topic is discussed later in this paper.

Design Procedures

Design procedures were established to properly analyze and develop Type A weaving sections on frontage roads. The necessary criteria are given in the following.

TABLE 2 LOS Criteria for LCI

LOS	Lane Changing Intensity (LCI)	
	Metric (lcphpmpl)	lcphpmpl
Unconstrained	0 - 1863	0 - 3000
Constrained	1863 - 376	3000 - 6000
Undesirable	> 3726	> 6000

Step 1: Establish Roadway Conditions

Existing or proposed roadway conditions must be specified before proceeding with the analysis. Roadway conditions include the length and number of lanes for the weaving section being studied (Figure 8).

Step 2: Determine Traffic Volumes

Traffic volumes should be expressed as hourly flow rates, which are obtained by identifying the peak 15-min interval within the hour of interest and multiplying this value by four. These values should be converted to passenger car equivalents. As shown in Figure 8, volumes are needed for ramp traffic and frontage road traffic entering the weaving section.

Step 3: Convert Traffic Volumes to Average Volume per Lane

Traffic volumes developed in Step 2 are converted to an average lane volume by adding the freeway exit ramp and frontage road volumes to obtain a total volume entering the weaving section and dividing this value by the number of lanes in the weaving section.

Step 4: Calculate LCI

LCI can be calculated using Equations 3 through 5 or can be obtained graphically from Figure 7.

Step 5: Determine LOS

LOS can be determined from the LCI by using the ranges of values given in Table 2 or by using Figure 7, which graphically illustrates the LOS boundaries.

FINDINGS AND RECOMMENDATIONS

Obviously, it is desirable not to have any weaving sections in a roadway design, but there are times when the alternatives are even less desirable. When a weaving section is to be part of a design,

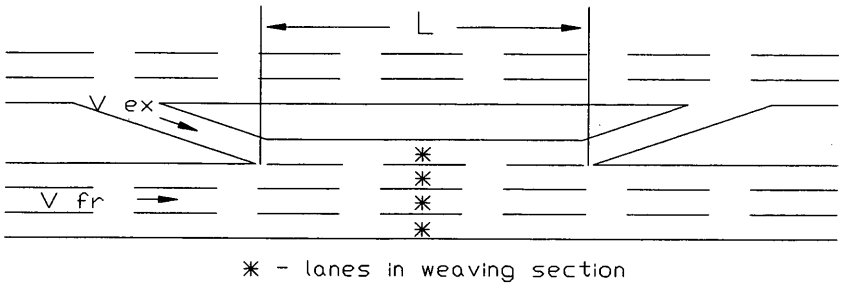


FIGURE 8 Weaving section analysis requirements.

special consideration should be given to both the length of the section and the number of lanes in the section. The projected LOS for a weaving section can be improved by adjusting the roadway conditions.

Lane Length

It is desirable to have a weaving section length in the range of 274.4 to 365.9 m (900 to 1,200 ft) as shown in Figure 8. A length in this range would help to ensure that weaving problems were minimized. It is desirable not to have a weaving section shorter than 182.9 m (600 ft). Weaving sections shorter than 182.9 m (600 ft) and significant traffic volumes will most likely experience operational problems.

Number of Lanes

A minimum of three lanes is recommended for weaving sections; this includes two through lanes and one auxiliary lane connecting the two ramps. Four lanes are recommended for weaving sections with significant volumes. The addition of a lane can help alleviate existing or projected operational problems, assuming the added lane is actually used. A lane could be added and not improve conditions if most of the traffic is weaving traffic and the additional lane is not used because there is little demand for through lanes in the weaving section.

Intermediate Disturbances

The design of weaving sections should not include intersections or driveways. The presence of driveways can have a significant effect on the operations of any facility, and this is especially true of weaving sections. The combination of the turbulence caused by weaving traffic and the effect of traffic turning into and out of cross streets or driveways could cause not only operational problems but safety problems as well.

Summary of Findings

The objective of this project was to develop a procedure for analyzing weaving section operations on nonfreeway facilities that was both reasonably accurate and simple to use. It has been determined by previous weaving studies and by this research that speed is not an adequate measure of effectiveness because of its insensitivity to traffic volumes typically experienced on frontage roads. Two possible measures of effectiveness were studied: density and LCI.

Density was eliminated as a possible measure of effectiveness because at uniform speeds, density is simply volume divided by a constant, and any model depicting this relationship would not be useful in predicting weaving operations.

Models were developed to predict LCI for three ranges of weaving section lengths. The resulting models had reasonable r^2 values and are easily used. LOS criteria were established for the LCI model, providing LCI ranges for three levels. Only three levels were defined because of the difficulty in determining the boundary values for each level.

LCI appears to be an effective performance measure for weaving sections. The relationship between LCI and average volume provided r^2 values that were higher than r^2 values for relationships currently being used (typically 0.50 to 0.60) for weaving section analysis (*11*). Application of the methodology outlined in this report is relatively simple and requires few data. Only geometric conditions and traffic volumes are required, both of which are easily attained. The methodology is also consistent in its approach to analyzing weaving sections with the 1985 HCM, other than using a different measure of effectiveness.

Future Research

Future research is required to calibrate the LCI model for different weaving configurations and to test sections of various lengths. The data used to develop the LCI model were obtained exclusively from Type A frontage road weaving sections with ramps on the left side. The LCI model is also intended to be used to analyze weaving sections on collector-distributor roads with ramps on the right side. It is possible that the LCI model will need to be recalibrated for these weaving sections.

ACKNOWLEDGMENTS

This paper is based on a study conducted by the Texas Transportation Institute and is sponsored by the Texas Department of Transportation.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. Cassidy, M. J., A. Skabardonis, and A. D. May. Operation of Major Weaving Sections: Recent Empirical Evidence. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1987.
3. Cassidy, M. J., and A. D. May. Proposed Analytical Technique for Estimating Capacity and Level of Service of Major Freeway Weaving Sections. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991.
4. *Highway Capacity Manual.* Bureau of Public Roads, U.S. Department of Commerce, 1950.
5. *Special Report 87: Highway Capacity Manual.* HRB, National Research Council, Washington, D.C., 1965, pp. 160–186.
6. Pignataro, L. J., et al. *NCHRP Report 159: Weaving Areas—Design and Analysis.* TRB, National Research Council, Washington, D.C., 1975.
7. *Circular 212: Interim Materials on Highway Capacity.* TRB, National Research Council, Washington, D.C., 1980.
8. Leisch, J. E. A New Technique for Design and Analysis of Weaving Sections on Freeways. *ITE Journal*, Vol. 49, No. 3, March 1979.
9. Leisch, J. E. *Procedure for Analysis and Design of Weaving Sections.* Final Report. Project RD-82/54. FHWA, U.S. Department of Transportation, 1984.
10. Fazio, J., and N. M. Raiphail. Freeway Weaving Sections: Comparison and Refinement of Design and Operations Analysis Procedures. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986.
11. Sadegh, A., et al. *Operation of Weaving Areas Under Non-Freeway Conditions.* Center for Transportation Studies and Research, New Jersey Institute of Technology, Newark, 1991.

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

Methodology for Determining Level of Service Categories Using Attitudinal Data

SAMER M. MADANAT, MICHAEL J. CASSIDY, AND WAN-HASHIM WAN IBRAHIM

Level of service (LOS) is standard terminology used for characterizing the operational quality of a transportation facility as perceived by the user of that facility. Given that transport systems are commonly designed and operated to maintain a specified LOS, it is a matter of some concern that the measures of effectiveness currently adopted for assessing LOS, as well as the threshold values for partitioning LOS designations, have been established subjectively. A methodology for partitioning LOS designations by using an ordered probit model calibrated with attitudinal data collected by transportation "users" is described. The application of this methodology is demonstrated by using survey data of bus riders. The basic approach, however, can be applied to all types of transportation facilities.

The 1985 *Highway Capacity Manual* (HCM) defines level of service (LOS) as "a qualitative measure describing operational conditions within a traffic stream, and their perception by motorists and/or passengers" (*1*). Thus, a designation of A through F is intended to characterize the operating quality of a subject transportation facility or system as perceived by the user. Although the HCM does state that its published analysis techniques are not intended to serve as legal standards for designing transportation systems, LOS has become a deeply embedded concept in the transportation psyche. Both the professional and the layman use it to depict existing or projected conditions. And, most important, LOS designations are used to influence decisions of tremendous economic consequence.

In a typical jurisdiction, for example, transportation systems may be designed and operated to maintain a stipulated LOS. Where changing environmental conditions (e.g., increased vehicle demand) cause LOS to erode below a stipulated designation, mitigating measures may be obligated at great cost to taxpayers, developers, and users.

Given the consequences of decisions made in response to measured or predicted LOS, it is imperative that LOS designations truly reflect that which they are intended. That is, the parameters thought to best characterize operating conditions for a particular type of transportation system (called measure of effectiveness, or MOE) must actually reflect user perceptions of operational quality. Likewise, the parameter values that separate LOS A from B, B from C, and so on must reflect boundaries that are consistent with the perceptions of the user population.

It is therefore a matter of some concern that the measures of effectiveness currently used to characterize LOS, as well as the threshold values used to separate LOS designations, reflect nothing more than the consensus of those involved in developing the HCM techniques. In short, LOS parameters and threshold values represent the judgment of a TRB committee. There appears to be no body of work conclusively relating LOS parameters to the perceptions or attitudes of the user population.

The work described in this paper has focused primarily on the identification of appropriate threshold values for partitioning one LOS designation from the next. The paper describes a technique to establish threshold values by making use of an ordered probit model (*2*) calibrated with survey data of user attitudes. Because the threshold values identified in this work actually reflect user perceptions, the proposed methodology is a considerable improvement over the somewhat arbitrary manner in which LOS designations are now partitioned.

The paper demonstrates the proposed methodology by applying it in conjunction with survey data reflecting LOS conditions perceived by bus riders. Although bus transit is only one type of transportation system, the methodology presented in this paper can be applied to any type of transportation facility currently addressed in the HCM. The decision to use attitudinal data collected from bus passengers was motivated solely by the relative ease with which such data could be collected.

RESEARCH APPROACH

Threshold values for partitioning LOS designations were identified by using an ordered probit model. Ordered probit modeling is one of several commonly used econometric techniques for the analysis of rating data. Specifically, where respondents are asked to evaluate a product or service on an ordinal scale (e.g., from 1 to 10 or from A to F), the correct methodology is to use a class of models with ordered dependent variables such as ordered probit or ordered logit. These techniques allow the analyst to correlate user responses to a host of explanatory variables (i.e., potential measures of effectiveness). Simultaneously, these techniques facilitate the identification of the thresholds between successive ratings.

Calibrating the ordered probit model required a data base relating the LOS designation perceived by users to the actual parameter values of the MOE. For example, the adopted MOE for signalized intersection LOS is delay. One could measure the intersection delay imparted to a specific motorist and then, in theory, ask the motorist to rate his or her perceived LOS at the conclusion of the delay period. Repeating this experiment for numerous motorists would provide the necessary data base for calibrating the ordered probit model.

The obvious procedural problem is that of usurping from motorists their perceptions of service quality. Conducting controlled experiments using a selected study group represents one feasible approach to collecting such motorist data. However, such an

S. M. Madanat and W.-H. W. Ibrahim, School of Civil Engineering, Purdue University, West Lafayette, Ind. 47907. M. J. Cassidy, Institute of Transportation Studies, University of California at Berkeley, 109 McLaughlin Hall, Berkeley, Calif. 94720.

endeavor was considered to be well beyond the scope of the research presented in this paper. Attitudinal data could, however, readily be collected from bus riders.

The HCM does include a chapter dedicated exclusively to transit LOS (and capacity). According to the HCM, the LOS imparted to bus passengers directly corresponds to the level of crowding on the bus. More specifically, the selected MOE is available square feet per passenger. The following table reproduces the MOE thresholds adopted by the HCM:

LOS	*Space per Passenger (ft^2)*
A	≥ 13.1
B	13.0 to 8.5
C	8.4 to 6.4
D	6.3 to 5.2
E (maximum scheduled load)	5.1 to 4.3
F (crush load)	< 4.3

The primary task in this work was to compare the MOE thresholds for bus riders arbitrarily adopted by the HCM with the thresholds rationally established using the stated perceptions of bus riders themselves.

DATA COLLECTION

Data reflecting rider perceptions were collected on numerous buses in the Chicago Transit Authority (CTA) system on Monday, March 22, 1993. To conduct the survey, a data collector individually asked riders to specify their perceived LOS. Specifically, the data collector identified himself as a CTA employee, stated that he was conducting a passenger survey, and asked each rider to rate his or her "present level of comfort on the bus on a scale of one to six; where one corresponds to a rating of very comfortable and six to a rating of unacceptable discomfort." Note that a rating of 1 to 6 corresponds to a LOS of A to F.

Coincident to each rider response, the data collector kept a running count of the number of passengers on board the bus. In this way, perceived LOS designations were correlated with the MOE currently used in the HCM: available square feet per passenger.

The data collector strived to randomly sample riders in an effort to avoid systematic bias in the data base. Moreover, the data collector spatially sampled individual passengers within the bus so that respondents would not be influenced by the responses of those around them. In total, 174 responses were collected from passengers riding standard 40-ft-long buses. The following table summarizes the total number of responses for each of the six specified ratings:

Stated Response	*Count*	*Percentage*
1	65	37.4
2	31	17.8
3	35	20.1
4	19	10.9
5	5	2.9
6	19	10.9

The survey instrument used in this work (i.e., the question posed by the data collector) was a fast and simple way to obtain the needed attitudinal data. The responses provided an adequate data base to satisfy the methodological objectives of this research (i.e., to demonstrate the application of ordered probit for partitioning LOS designations). In terms of its ability to obtain unbiased data, the survey instrument is suspect. The conclusions of this paper include a discussion of how the instrument might be improved as part of a comprehensive effort to identify LOS thresholds appropriate for generalized application.

MODEL SPECIFICATION AND ESTIMATION

For each observation (individual) i, the following variables are available:

y_i = stated level of comfort, $y_i \in \{1, 2, 3, 4, 5, 6\}$;
x_i = passenger density on bus at time individual i provided a response (passengers/ft^2); and
$i = 1, 2, \ldots, 174$.

Define as the latent comfort of individual i at the time of his or her response

$$U_i = \alpha + \beta x_i + \epsilon_i$$

where α and β are parameters to be estimated (where α can be thought of as absorbing the mean of ϵ), and ϵ_i is the random error term, accounting for all unobserved attributes contributing to individual i's perceived comfort; because the error term is the sum of a large number of random effects, it can be assumed normally distributed, that is,

$$\epsilon_i \sim N(0, \sigma^2) \quad (1)$$

Thus U_i can be divided into two components: a systematic component, $V_i = \alpha + \beta x_i$, and a random contribution, ϵ_i.

The stated level of comfort for individual i, y_i, is related to his or her latent comfort in the following manner:

$$y_i = 1 \text{ if } U_i \leq k_1 \Rightarrow \alpha + \beta x_i + \epsilon_i \leq k_1 \quad (2a)$$

$$y_i = 2 \text{ if } k_1 < U_i \leq k_2 \Rightarrow k_1 < \alpha + \beta x_i + \epsilon_i \leq k_2 \quad (2b)$$

$$y_i = 3 \text{ if } k_2 < U_i \leq k_3 \Rightarrow k_2 < \alpha + \beta x_i + \epsilon_i \leq k_3 \quad (2c)$$

$$y_i = 4 \text{ if } k_3 < U_i \leq k_4 \Rightarrow k_3 < \alpha + \beta x_i + \epsilon_i \leq k_4 \quad (2d)$$

$$y_i = 5 \text{ if } k_4 < U_i \leq k_5 \Rightarrow k_4 < \alpha + \beta x_i + \epsilon_i \leq k_5 \quad (2e)$$

$$y_i = 6 \text{ if } U_i > k_5 \Rightarrow \alpha + \beta x_i + \epsilon_i > k_5 \quad (2f)$$

where $k_1, \ldots, k_5$ are the unobserved thresholds on the latent scale separating consecutive levels of comfort.

Equations 1 and 2 fully describe the model specification. Such a specification represents an ordered probit model (2). An ordered probit structure is an extension of a simple binary probit model to a case in which the observed indicator variable is ordinal and takes a value between 1 and $m > 2$.

The objective of the estimation is to provide statistical estimates of the model parameters α, β, and $k_1, \ldots, k_5$. This objective is achieved through the use of maximum likelihood estimation (MLE).

Not all parameters of Model 2, however, are uniquely identifiable by MLE. This can be readily observed if Equation 2 is rewritten as

$$y_i = 1 \text{ if } \beta x_i + \epsilon_i < k_1 - \alpha \quad (3a)$$

$$y_i = 2 \text{ if } k_1 - \alpha < \beta x_i + \epsilon_i \leq k_2 - \alpha \quad (3b)$$

$$y_i = 3 \text{ if } k_2 - \alpha < \beta x_i + \epsilon_i \leq k_3 - \alpha \quad (3c)$$

$$y_i = 4 \text{ if } k_3 - \alpha < \beta x_i + \epsilon_i \leq k_4 - \alpha \quad (3d)$$

$$y_i = 5 \text{ if } k_4 - \alpha < \beta x_i + \epsilon_i \leq k_5 - \alpha \quad (3e)$$

$$y_i = 6 \text{ if } \beta x_i + \epsilon_i > k_5 - \alpha \quad (3f)$$

It can be seen that α is not distinguishable from the thresholds $k_1, \ldots, k_5$. Only the differences $k_i^* = k_i - \alpha$, $i = 1, \ldots, 5$ are statistically identifiable. Therefore, the MLE procedure will only provide estimates of β, $k_1^*, \ldots, k_5^*$. This is basically equivalent to the normalization $\alpha = 0$.

The standard normalization $\sigma^2 = 1$ is also required. This latter normalization is common to all probit models and determines the scale of the model parameters.

Model 3 can be estimated by using a general-purpose MLE routine available in most statistical software or by using specialized probit estimation programs. This research has used the standard probit procedure available in SST (*3*).

The estimation results are presented in Table 1. Referring to Table 1, all threshold parameters, with the exception of k_1^*, are highly significant (i.e., all t-statistics $>> 2$). This reflects a high level of confidence in their values. The density parameter, β, is significantly different from 0 (t-statistic = 3.19), indicating that passenger density does influence perceived LOS. The overall fit of the model, however, is low ($\rho^2 = 0.083$), indicating that passenger density alone does not explain variations in rider responses to the level of comfort question.

To compare threshold values estimated through the ordered probit approach with those documented in the HCM requires that all thresholds be of equal scale. To convert those thresholds generated by the ordered probit model, the values of $\hat{k}_1^*, \ldots, \hat{k}_5^*$, (the estimated k_i^* values) were first divided by $\hat{\beta}$ (the estimated value of β) to obtain $\hat{k}_i' = \hat{k}_i^*/\hat{\beta}$, $i = 1, \ldots, 5$ thresholds on the density scale. These density thresholds $\hat{k}_1', \ldots, \hat{k}_5'$ are then inverted to obtain area thresholds $\hat{t}_1, \ldots, \hat{t}_5$, compatible with the scale used for thresholds in the HCM.

$$\hat{t}_i = 1/\hat{k}_i' \quad i = 1, \ldots, 5$$

where $\hat{t}_i$ equals thresholds on the scale of available area per passenger, in square feet per passenger.

The following table presents the threshold values estimated by the ordered probit procedure:

LOS	*Space per Passenger (ft²)*
A	≤ 305.0
B	305.0 to 13.2
C	13.1 to 6.0
D	5.9 to 4.3
E	4.2 to 3.9
F	< 3.9

ANALYSIS OF FINDINGS

If one were to assume that the responses collected from CTA riders in this research at least approach or approximate the perceptions of bus riders in general, the MOE and threshold values adopted by the HCM are highly suspect. To begin, both the currently adopted thresholds presented in Table 1 and the values generated from the ordered probit model in the previous table reflect LOS thresholds relevant to a standard 40-ft bus with an interior area of about 340 ft². Significant differences exist between the threshold values presented in these two tables.

The probit-generated thresholds in the previous table suggest that LOS A conditions are difficult to obtain on an urban transit bus as the presence of more than one passenger results in an available area below the LOS A threshold. In contrast, the HCM thresholds indicate that as many as 26 passengers can be aboard before operating conditions erode to LOS B. For Levels B and C, the probit-generated thresholds differ from the currently adopted values by

TABLE 1 Model Estimates

Independent Variable	Estimated Coefficient	Standard Deviation	t-statistic
k1*	0.021	0.141	0.147
k2*	0.478	0.146	3.270
k3*	1.051	0.150	6.985
k4*	1.479	0.150	9.850
k5*	1.631	0.155	10.526
Density	6.313	1.980	3.189

$L(0) = -294.90$

$L(\hat{\beta}) = -270.46$

Rho-Squared = 0.083

approximately one step size—that is, the HCM thresholds delineating LOS A from B and LOS B from C are very close to the probit-generated thresholds delineating LOS B from C and LOS C from D, respectively. For the lower LOS conditions, the probit-generated thresholds suggest that riders are more willing to tolerate higher passenger densities than those implied by the HCM thresholds. Given that transit operators may establish service frequencies and bus sizes with reference to maximum allowable (i.e., crush) loads, improving the LOS thresholds by exploiting the proposed methodology with an expanded data base would provide worthwhile information. If indeed LOS F is defined by an area smaller than 4.9 ft^2/passenger (the value recommended by the HCM), a reduced frequency of service might be acceptable. This could lead to substantial cost savings for the transit agency.

Findings from this work, however, are not limited to the identification of large differences between the LOS thresholds currently adopted and those derived through the ordered probit approach. The models calibrated in this research effort indicate that passenger density, while being a significant predictor, does not in itself strongly characterize perceived LOS. The significance of this finding is discussed further in the next section.

CONCLUSIONS

The specific findings resulting from this research are by no means definitive: the probit-generated thresholds presented herein are not values that the authors propose for adoption by the HCM or any transit agency. The objective behind this work has been to demonstrate a more rational methodology for establishing LOS parameters. The small data set collected in this effort provided a simple means to this end. However, the size of the data base and the instrument used for acquiring these data are far from ideal. Obtaining a more reliable and representative data base would require (a) a significantly enhanced survey instrument for measuring latent LOS designations and (b) an expanded number of observations reflecting rider perceptions under a greater variety of operating conditions, bus systems, geographic regions, and so forth.

Regarding the first concern, the instrument used in this research fell short of commonly adopted standards (*4*). For our application, the exclusive use of stated preference data potentially promotes policy-response bias, as respondents may believe that responding negatively to any questions concerning passenger comfort might induce mandated improvements to "their" bus system. The potential for this bias was likely exacerbated by asking respondents a single question reflecting an obvious objective. At the very least, the survey instrument could be enhanced for future surveys by providing riders with a questionnaire incorporating a number of bipolar options characterizing passenger comfort. To further minimize the validity problems commonly associated with stated preference data, a passenger questionnaire could be developed incorporating both stated and revealed preferences (*5*).

The need also exists to identify operating items, in addition to passenger density, that influence LOS perceptions. Such items might include factors such as bus condition and aesthetics, demographic features of the riders and routes, waiting times at the bus stop (i.e., service frequencies), and required number of transfers. Such items (and their associated significance) can be identified only through an extensive data collection effort to measure the values of the potential influences. These values could then be correlated with individual survey responses as part of a comprehensive model-building process. The effort might result in an expression for estimating a performance index characterizing LOS. As noted in this paper, this type of research could also be carried out to assess motorists' perceptions of LOS relevant to other types of transportation facilities.

In the final assessment, the value of the research described in this paper does not lie in the specific parameter values identified. Instead, the contribution of this work has been to demonstrate the manner in which a commonly used modeling technique—namely, ordered probit—can be applied to address an important but overlooked transportation issue: LOS as perceived by the user.

Findings from the specific application described in this paper should prove relevant to transit agencies. Transit operators are certainly concerned with passenger comfort and the service-scheduling and fleet-sizing issues related to comfort. However, the authors hold that the relevance of this work extends well beyond application to bus riders. The proposed methodology applies to virtually all transportation facilities in which LOS is a relevant issue.

If the operating quality of a transportation facility is to be evaluated from the perspective of the user (and it seems logical that it should), adopted LOS designations must truly reflect these perceptions. The lack of existing research in this topic, and the rather subjective manner in which LOS is currently defined, are therefore matters of significant concern. A great deal of money might be spent to improve the operating conditions of a given transportation facility by one or two LOS designations, yet the extent to which these improvements actually influence user perception of LOS is practically unknown. Perhaps more important, the federal government is allocating millions of dollars to fund research projects directed at developing and improving the accuracy of analytical procedures for predicting (arbitrarily selected) MOEs. Still, there is no certainty concerning the significance of these MOEs for characterizing LOS from a user's perspective.

LOS designations must be better understood and applied in transportation engineering and planning. The research described in this paper proposes an approach for addressing this fundamental issue.

ACKNOWLEDGMENTS

The authors wish to thank Ross Patronsky, CTA Strategic Planning Department, for his assistance in acquiring survey data. The authors also thank Chris Williams, Purdue University, for collecting the data.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985, pp. 1–3.
2. Ben-Akiva, M., and S. Lerman. *Discrete Choice Analysis.* MIT Press, Cambridge, Mass., 1985.
3. Dubin, J., and D. Rivers. *Statistical Software Tools, Version 2, Reference Manual.* Dubin-Rivers Research, Pasadena, Calif., 1990.
4. Eagly, A., and S. Chaiken. *The Psychology of Attitudes.* Harcourt Brace Jovanovich Publishers, 1993.
5. Morikawa, T. *Incorporating Stated Preference Data in Travel Demand Analysis.* Ph.D. dissertation. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, 1989.

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

Estimation of Green Times and Cycle Time for Vehicle-Actuated Signals

RAHMI AKÇELIK

An analytical method for estimating average green times and cycle time at vehicle-actuated signals is presented. The examination is limited to the operation of a basic actuated controller that uses passage detectors and a fixed gap time setting. Both fully actuated and semiactuated control cases are discussed. The practical cycle and green time method for computing fixed-time signal settings is also outlined. A discussion of the arrival headway distributions is presented since the estimation of arrival headways is fundamental to the modeling of actuated signal timings. The method given provides essential information for predicting the performance characteristics (capacity, degree of saturation, delay, queue length, and stop rate) of intersections controlled by actuated signals and for investigating the optimization of actuated controller settings. Further work is needed to validate and calibrate the formulas given using real-life and simulation data.

This paper presents an analytical method for estimating average green times and cycle time at vehicle-actuated signals. This information is essential for the prediction of the performance characteristics (capacity, degree of saturation, delay, queue length, and stop rates) of intersections controlled by actuated signals. The method can be seen as an extension of the current Australian, U.S. *Highway Capacity Manual* (HCM), United Kingdom, and similar methods for the analysis of fixed-time (pretimed) signals (*1–4*). The practical cycle and green time method for computing fixed-time signal settings is also outlined (*1,2*).

This paper is limited to the operation of a basic actuated controller that uses passage detectors and a fixed gap time setting. Both fully actuated and semiactuated control cases are discussed. The author is preparing a more comprehensive report that discusses actuated signal controllers that use various gap reduction methods with passage and presence detectors (*5*).

The literature on actuated signal operations is limited compared with that on fixed-time signals. However, there are still many useful papers on actuated signals, mostly based on the use of simulation methods, and a few of them describe analytical techniques. A detailed literature review is outside the scope of this paper. The descriptions of actuated controller operations provided by Staunton (*6*) and the analytical methods provided by Lin (*7,8*) were used in the development of the work reported in this paper.

The method presented for the analysis of actuated signal operations can be implemented manually. However, implementation through computer software such as SIDRA (*2*) is useful for dealing with complex intersection geometry and phasing arrangements and for obtaining solutions that require iterations.

The arrival headway distributions are discussed first, since the estimation of arrival headways is fundamental to the modeling of actuated signal timings.

Australian Road Research Board, Ltd., P.O. Box 156, Nunawading, Victoria 3131, Australia.

ARRIVAL HEADWAY DISTRIBUTIONS

Accuracy in predicting small arrival headways (up to about 12 sec), rather than the whole range of headways, is particularly important in modeling actuated signal operations. A class of arrival headway distributions referred to as M1 (negative exponential), M2 (shifted negative exponential), and M3 (bunched exponential) is considered. The M3 model was proposed by Cowan (*9*) and used extensively by Troutbeck (*10–12*) for estimating capacity and performance of traffic circles and other unsignalized intersections. A special case of the M3 model has been used by Tanner (*13,14*) for unsignalized intersection analysis. The M1 and M2 models can be derived as special cases of the M3 model through simplifying assumptions about the bunching characteristics of the arrival stream.

The M1 and M2 models are more commonly used in the traffic analysis literature as models of random arrivals. However, the M3 model is found to be more representative of real-life arrival patterns. The more commonly used shifted negative exponential (M2) model is found to give poor predictions for the range of small headways, which is of particular interest when modeling actuated signal operations and gap acceptance at intersections.

This paper uses the bunched exponential (M3) model for deriving various formulas for the analysis of actuated signal operations. It is recommended that this model be used consistently for all urban traffic analysis (gap acceptance modeling at signalized and unsignalized traffic facilities, modeling of traffic performance, and so on). For a detailed discussion of the bunched exponential model of arrival headways, see a recent paper by Akçelik and Chung (*15*).

The cumulative distribution function, $F(t)$, for the bunched exponential distribution of arrival headways, representing the probability of a headway less than t sec, is

$$F(t) = \begin{cases} 1 - \phi \exp[-\lambda(t - \Delta)] & \text{for } t \geq \Delta \\ 0 & \text{for } t < \Delta \end{cases} \quad (1)$$

where

Δ = minimum headway in arrival stream (sec),
ϕ = proportion of free (unbunched) vehicles, and
λ = model parameter calculated from $\lambda = \phi q_t/(1 - \Delta q_t)$, where q_t is the total arrival flow (vehicles/sec).

The proportion of bunched vehicles in the arrival stream is $(1 - \phi)$. The free (unbunched) vehicles are those with headways greater than the minimum headway (Δ), and the proportion of free vehicles (ϕ) represents the unbunched vehicles with randomly distributed headways. All bunched vehicles are assumed to have the same intrabunch headway (Δ).

The M1 and M2 models can be derived from the M3 model by setting the bunching parameters as follows:

Negative exponential (M1) model:

$$\Delta = 0$$

$$\phi = 1 \text{ (therefore } \lambda = q_l) \tag{2a}$$

Shifted negative exponential (M2) model:

$$\phi = 1 \tag{2b}$$

A known value of ϕ can be specified for use in the M3 model. For general application purposes, ϕ can be estimated as a function of the arrival flow rate. The following relationship has been derived by the author by generalizing the bunching implied by the negative exponential model:

$$\phi = \exp(-b\Delta q) \tag{3}$$

where b is a bunching factor and q is the arrival flow rate in vehicles per second.

The M3 model with estimates of ϕ obtained from Equation 3 will be referred to as the M3A model and will be used in this paper with the following parameter values:

Single-lane case:

$$\Delta = 2.0$$

$$b = 1.5 \tag{4a}$$

Multilane case (number of lanes = 2):

$$\Delta = 1.0$$

$$b = 1.0 \tag{4b}$$

Multilane case (number of lanes > 2):

$$\Delta = 0.5$$

$$b = 1.0 \tag{4c}$$

The bunching factor of $b = 1.5$ for the single-lane case was derived as an approximation to the values predicted by the following linear model used by Tanner (*13,14*):

$$\phi = 1 - \Delta q \qquad \text{for } q < 1/\Delta \tag{5}$$

The M3 model with estimates of ϕ obtained from Equation 5 will be referred to as the M3T model.

The bunching factors for multilane cases are based on the treatment of arrival flows in all lanes as a single stream. The values given in Equations 4b and 4c were derived through comparison with lane-by-lane treatment of multilane situations.

Research carried out after writing this paper to calibrate the M3A model using real-life and simulation data (*15*) indicated lower levels of bunching than those predicted by Equations 4a through 4c.

Figure 1 shows cumulative distribution functions for the arrival headway models M1, M2, M3A, and M3T for a single-lane traffic stream with arrival flow rate of 900 veh/hr. There are significant differences in the predictions of arrival headways by different models, especially for small arrival headways (up to about 12 sec). Generally, the shifted negative exponential model does not appear to be a satisfactory model. The amount of bunching as represented by parameter ϕ in Model M3 has a major effect on the prediction of arrival headways.

The following formulas provide two fundamental parameters for actuated signals (used for estimating the extension time before a gap change after queue clearance; see Equations 13 and 14):

$$n_g = -1 + (1/\phi)\exp[\lambda(e_o - \Delta)] \tag{6a}$$

$$h_g = (1/n_g)\{-(e_o + 1/\lambda) + (\Delta/\phi + 1/\lambda)\exp[\lambda(e_o - \Delta)]\} \tag{6b}$$

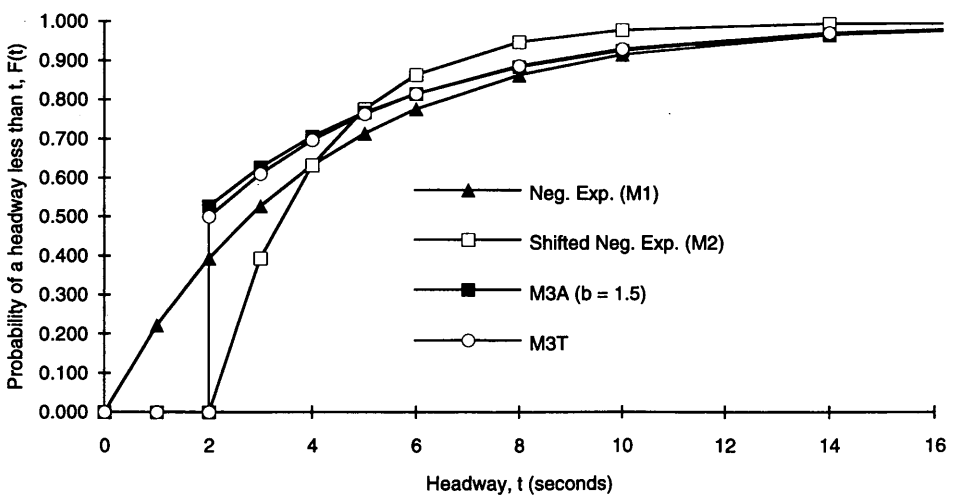


FIGURE 1 Cumulative headway probabilities predicted by Models M1 ($\Delta = 0$), M2 ($\Delta = 2$), M3A ($\Delta = 2$, $b = 1.5$), and M3T ($\Delta = 2$): single-lane case with arrival flow rate = 900 veh/hr.

where

n_g = average number of arrivals before a gap change after queue clearance (average number of consecutive arrivals with headways $t < e_o$ followed by an arrival with a headway $t \geq e_o$),
h_g = average headway before a gap change after queue clearance (average headway corresponding to n_g), and
e_o = gap time setting.

Equations 6a and 6b can be used for the M1 and M2 models by choosing appropriate parameters as given in Equations 2a and 2b.

Another useful formula is the probability of no arrivals during a given period of T sec (equivalent to the probability of a headway longer than T):

$$p_o = \phi \exp[-\lambda(T - \Delta)] \quad (7)$$

ACTUATED SIGNAL TIMINGS

An analytical method is presented for estimating the average green times and cycle time for a basic actuated controller that uses passage detectors and a fixed gap time setting. Both fully actuated and semiactuated control cases are considered. Most studies of actuated signals reported in the literature assume this type of actuated controller operation. Akçelik further discusses more complex types of actuated signal controllers that use various gap-reduction techniques with passage and presence detectors (*5*).

At vehicle-actuated signals, the green time, and hence the cycle time, is determined according to the vehicle demands registered by detectors. This may be on the basis of phase (stage) control or group (movement) control. Phase sequence may be fixed or variable. A phase can be skipped when there is no demand for it. After a minimum green period, the running phase waits in the rest position when no conflicting demand is present (an automatic call for another phase has the same effect as a conflicting demand).

For a given phasing system, efficient operation of vehicle-actuated signals depends on the values of various controller settings. The three basic controller settings that determine the length of the green period are minimum green, gap time (the terms "vehicle interval," "vehicle extension," and "unit extension" are also used), and maximum extension (or maximum green) settings. Modern controllers have additional settings that differ according to the type of controller: minimum gap time, time to reduce, waste time, headway time, and so on. The location, number and other characteristics of detectors affect the choice of vehicle-actuated settings also.

Figure 2 shows the operation of a basic actuated controller that uses passage detectors and a fixed gap time setting for terminating the green time.

To facilitate easier analytical formulation:

1. The early cut-off and late start intervals are ignored;
2. Actuated signal operation is expressed in terms of movements rather than phases (although the discussion is valid for both group controllers and phase controllers); and
3. The formulas for estimating the green and cycle times are expressed in terms of effective green times rather than displayed green times.

Refer to Akçelik (*1*,pp. 1–5) for introductory discussions related to Items 2 and 3. Appropriate conversion of some controller settings to effective values is required before the formulas given in this paper can be used. For example, the displayed maximum green time

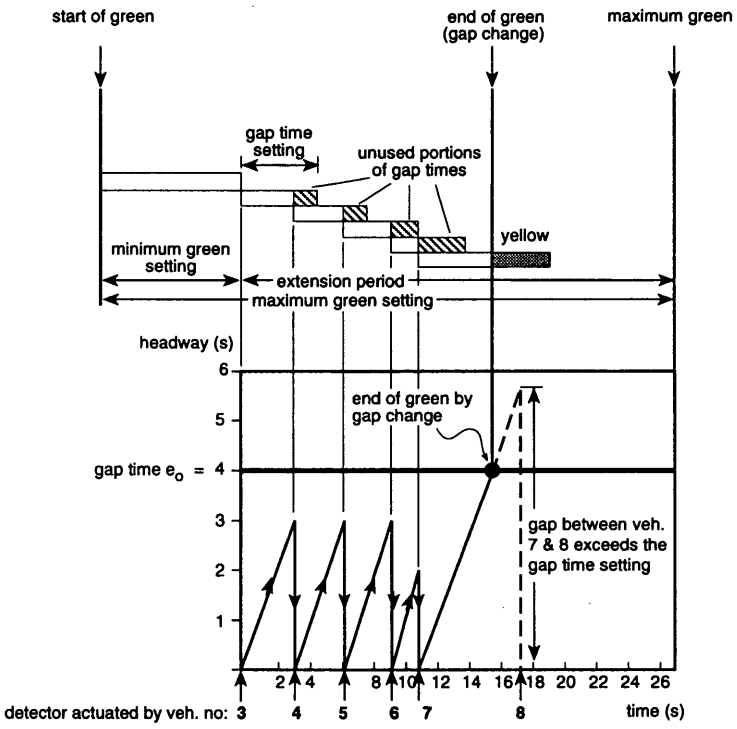


FIGURE 2 Extension sequence in basic vehicle-actuated control using passage detectors and a fixed gap time setting (*6*).

setting (G_{max}) needs to be converted to an effective maximum green time value (g_{max}) using $g_{max} = G_{max} + I - l$ where I is the intergreen time (yellow plus all-red) and l is the lost time. In most cases, $I \cong l$, and therefore $g \cong G$ can be assumed. However, all incremental settings such as the gap time, waste time, and vehicle increment settings can be used as controller settings without any adjustment for effective green time calculations.

Average Green Time for Fully Actuated Signals

The green time (g) allocated to a movement (group) under actuated control comprises a minimum green time (g_{min}) and a green extension time (g_e):

$$g = g_{min} + g_e \tag{8}$$

subject to

$$g \leq g_{max}$$

or

$$g_e < g_{emax}$$

where g_{max} is the maximum green setting and g_{emax} is the maximum extension time setting.

The minimum green time consists of a fixed minimum green time and a variable initial green time. The fixed minimum green time is determined as a safe minimum green that should be long enough for the first vehicle to start moving and enter the intersection (typically 4 to 6 sec). The variable initial green time is an additional variable period determined by the number of vehicle actuations (after the first vehicle) during the red period. Vehicle increment and maximum initial green settings are used in relation to this. The sum of the fixed minimum green time and maximum initial green time must be long enough to clear the vehicles waiting in the critical lane between the detection point and the stop line. For this purpose, the critical lane is defined as the lane with the highest flow rate (5).

The value of the maximum green time (or maximum extension time) setting to be used in practice must be chosen with due consideration to traffic flows at different times (morning peak, evening peak, day off-peak, night off-peak, weekend, and shopping periods) and to the peaking characteristics of traffic. The choice of the design period as a basis of green time calculations is important in this respect. The objective should be to obtain green times that are not too restrictive for maximum possible flow rates (e.g., during a peak 15-min period). On the other hand, long maximum settings coupled with a bad choice of the gap time and other controller settings can lead to unduly long green and cycle times, resulting in inefficient operation during a larger proportion of the time.

Traditionally, the green time calculation methods for fixed-time signals are used for determining suitable maximum green settings. However, the method given in this paper for the analysis of actuated operations could be used to determine appropriate values of the maximum green settings directly without resorting to fixed-time signal analysis.

The method for estimating the green extension time (g_e) for the basic actuated controller operation is given in the following. It assumes that a conflicting demand is registered before the termination of the minimum green period, and therefore, the extension period starts immediately after the expiration of the minimum green period (see Figure 2).

A basic actuated controller uses a fixed value of the gap time (vehicle interval or unit extension) setting (e_o) for terminating the green time (typically e_o = 2.5 to 4 sec). As seen in Figure 2, detection of each additional vehicle extends the green period by an amount equal to the gap time (e_o). The controller starts timing a new gap time at each vehicle actuation. The green period terminates when

1. The time between successive vehicle actuations exceeds the gap time setting, $h > e_o$ (gap change), or
2. The total green extension time after the expiration of minimum green time equals the maximum extension setting, $g - g_{min} = g_{emax}$ or $g = g_{max}$ (maximum change).

During a gap change (see Figure 2), the green period terminates after the gap time expires. In some controllers, a passage time setting (e_p) is used instead of the gap time for the last vehicle to be able to travel the distance between the detector and the stop line before the start of yellow signal. Thus, the terminating time (e_t) at gap change is either the gap time ($e_t = e_o$) or passage time ($e_t = e_p$). The gap timing logic operates from the start of the green period to enable a green termination at the end of the minimum green time.

During the saturated portion of the green period (i.e., during the queue clearance period), the headways are assumed to be equal $h = h_s = 1/s$, where h_s is the saturation headway and s is the combined saturation flow rate (in vehicles per second) for all lanes, allowing for any lane underuse. The standard methods for the calculation of saturation flow can be used (*1–4*). For a single lane, typically s = 1,800 veh/hr = 0.5 veh/sec, therefore h_s = 2.0 sec. For multilane cases, $s = s_c/p_c$ can be used, where p_c is the proportion of total flow in the critical lane and s_c is the critical lane saturation flow. This is a simplistic formula that assumes the same saturation flow (s_c) for all lanes but allows for unequal lane flows. For equal lane flows, $p_c = 1/n_l$, where n_l is the number of lanes, thus $s = n_l s_c$.

A gap change during the saturated portion of green period (after the expiration of the minimum green period) is theoretically possible, at least for a single-lane movement. This would occur if $h_s > e_o$ (for example, for a turning movement with s = 1,200 veh/hr, h_s = 3.0 sec, and e_o = 2.5 sec). Gap change during the saturated portion of the green period indicates an inefficient operation (insufficient green to clear the queue). Therefore, e_o should be set to ensure that a gap change does not occur during the saturated portion of the green period, particularly for single-lane movements. The analyses presented in the rest of this paper assume that the gap time is set to ensure that a gap change does not occur during queue clearance.

A gap change during the unsaturated portion of green—that is, after the queue clearance period—corresponds to conditions when the vehicles in the arrival stream pass through the intersection without queueing. The arrival headway distributions discussed in the previous section are applicable in this case.

The green time in the case of a gap change after queue clearance is

$$g = g_s + e_g \tag{9}$$

subject to

$$g_{min} \leq g \leq g_{max}$$

where g_s is the saturated portion of green period (queue clearance time) and e_g is the extension time by gap change after queue clearance.

From Equation 8, the green extension time can be calculated from

$$g_e = g - g_{\min} = g_s + e_g - g_{\min} \quad (10)$$

The saturated portion of green period can be estimated from

$$g_s = f_q(n_q/s + yr)/(1 - y) \quad (11)$$

where

f_q = calibration factor to allow for variations in queue clearance time,
n_q = residual queue from previous green period in case of two green periods per cycle ($n_q = 0$ for the more common case of a single green period per cycle),
s = saturation flow rate (veh/sec),
r = red time (sec), and
y = flow ratio ($y = q/s$ where q = arrival flow rate).

In multilane cases, the saturated portion of green should represent the time to clear the queue in the critical lane (i.e., the longest queue for any lane) considering all lanes of all approaches in the signal group (or phase). Ideally, g_s should be calculated for each lane in each approach using parameters relevant to each lane (e.g., flow, saturation flow, and effective red time for the lane). This method is used by SIDRA (2). Alternatively, g_s can be calculated for each lane group (or approach) by appropriate adjustments to flow, saturation flow, and effective red times to ensure that g_s for the lane group approximates the critical lane value.

Equation 11 allows for two green periods per cycle. For the more common case of a single green period per cycle, n_q is 0 (this should not be confused with overflow queues due to oversaturation in the cycle), and the green time can be expressed in terms of the cycle time (c) rather than the red time (r):

$$g = f_q yc + (1 - y)e_g \quad (12)$$

subject to

$$g_{\min} \leq g \leq g_{\max}$$

This is an approximate equation (exact if $f_q = 1$), and the use of Equations 9 and 11 is preferred.

The average extension time by gap change can be estimated from

$$e_g = n_g h_g + e_t \quad (13)$$

where

n_g = average number of arrivals before a gap change after queue clearance (due to a headway $h > e_o$), given by Equation 6a;
h_g = average headway before a gap change after queue clearance (due to a headway $h > e_o$), given by Equation 6b; and
e_t = terminating time at gap change (equals the gap time setting, $e_t = e_o$, or the passage time setting, $e_t = e_p$).

For the case when $e_t = e_o$, Equation 13 is equivalent to

$$e_g = \frac{\exp[\lambda(e_o - \Delta)]}{\phi q} - \frac{1}{\lambda} \quad (14)$$

See Equations 1 through 5 for parameters in this formula. For negative exponential distribution of headways, set $\Delta = 0$, $\phi = 1.0$, and $\lambda = q_t$ (total arrival flow). For shifted negative exponential distribution, set $\phi = 1.0$. The resulting formula is then similar to that given by Lin (*7,8*) except that Lin recommends $\Delta = 1$ sec for the single-lane case and $\Delta = 0$ sec for the multilane case (therefore equivalent to the negative exponential model).

When the gap timer operates from the start of the green period (including the minimum green period) and non-stop-line detectors are used, it is necessary to reduce n_g by the number of vehicles that arrive early during the green period, cross over the detector, and join the back of the queue downstream of the detection point. These vehicles are counted as part of the vehicles departing during queue clearance (sg_s vehicles) as well as part of n_g.

Figure 3 shows an example of average extension time by gap change after queue clearance (e_g) as a function of the total arrival flow (q) for a single-lane case with $e_o = 3$ sec (from Equation 14). The extension times are predicted using the arrival headway models M1 (negative exponential), M2 (shifted negative exponential), and bunched exponential models M3A and M3T. It is seen that there are sizable differences in the predictions of extension times by different headway models. The amount of bunching as represented by parameter ϕ in M3 has a significant effect on the prediction of extension times.

Figure 4 shows average extension time by gap change after queue clearance (e_g) as a function of the total arrival flow (q) using the M3A model for a single-lane case ($\Delta = 2.0$ sec, $b = 1.5$) and a four-lane case ($\Delta = 0.5$ sec, $b = 1.0$) with $e_o = 3.0$ and 4.5 sec. It is seen that the difference between extension times (e_g) for gap time settings of $e_o = 3.0$ and 4.5 sec increases with increasing flows to substantial levels at very high flows.

Semiactuated Signals

Semiactuated signal operation as a simple two-phase system controlling a major-minor road intersection is considered. For the sake of notations, the minor road will be referred to as the "side street" and the major road will be called the "main road." The side street vehicles are detected and controlled as in the case of fully actuated control. On the other hand, the main road has no detections. It receives only a minimum green time after a change of phase to the main road (e.g., by a gap change or maximum change). The main road phase is terminated after a conflicting demand is registered on the side street. Therefore, this type of operation is suitable only when the side street flows are low.

The formulas given here are close to those by Lin (*8*), but the more general M3 arrival headway distribution is used, the saturated part of the green period is dealt with differently, and the assumption about how a conflicting demand is registered in deriving the durations of main road and side street green periods is slightly different.

The average green time for the main road can be estimated from

$$g_M = g_{\min M} + (\phi_s/\lambda_s) \exp[-\lambda_s (e_{tS} - \Delta_s + l_M + g_{\min M})] \quad (15)$$

where

ϕ_s, λ_s, Δ_s = headway distribution parameters calculated from Equations 1 through 5 considering total flow in all lanes of all side street movements;

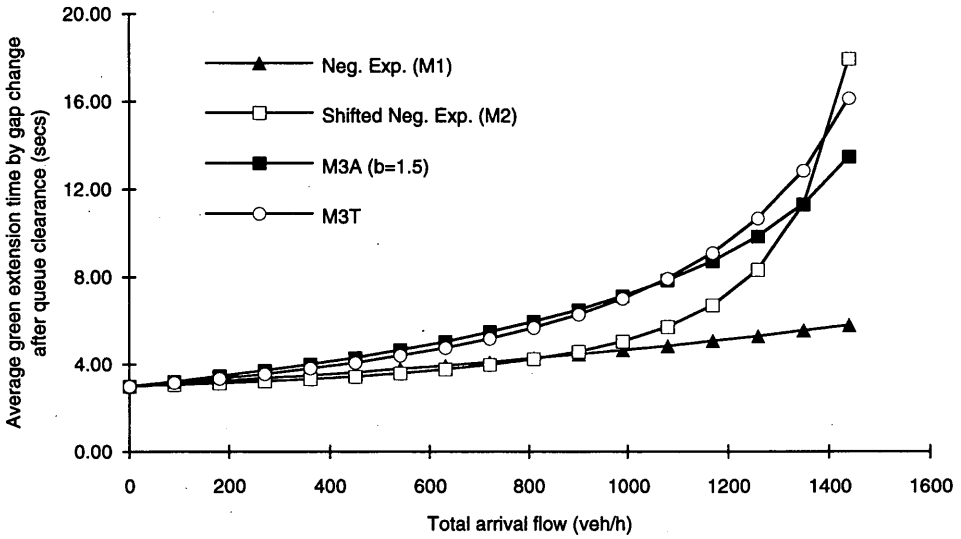


FIGURE 3 Average extension time by gap change after queue clearance (e_g) as a function of the total arrival flow (q) predicted by arrival headway Models M1 ($\Delta = 0$), M2 ($\Delta = 2$), M3A ($\Delta = 2$, $b = 1.5$), and M3T ($\Delta = 2$): single-lane case with $e_o = 3.0$.

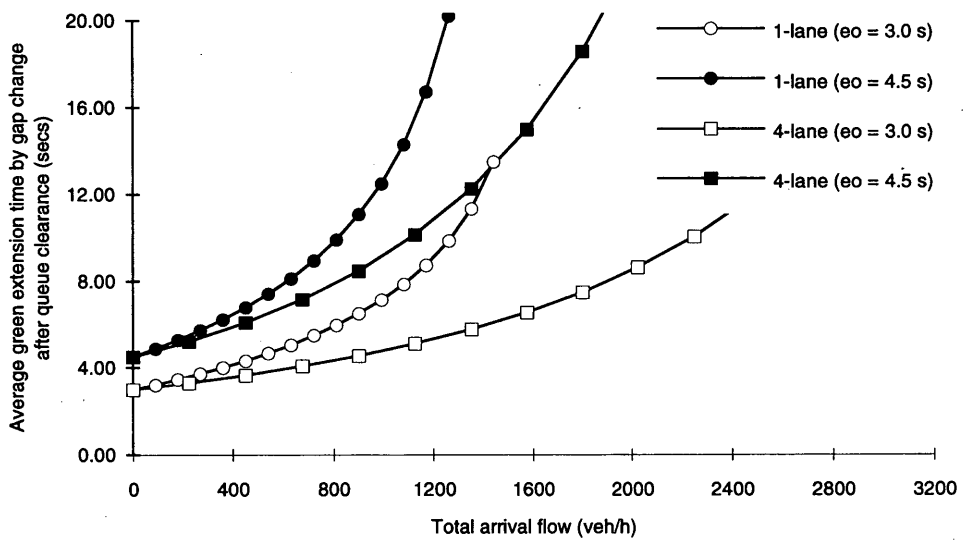


FIGURE 4 Average extension time by gap change after queue clearance (e_g) as a function of the total arrival flow (q) predicted by Model M3A for $e_o = 3.0$ and 4.5: single-lane case with $\Delta = 2$, $b = 1.5$, and multilane case with $\Delta = 0.5$, $b = 1.0$.

e_{tS} = terminating time for side street movements [this can be the gap time (e_o), passage time (e_p), or 0 in the case of maximum change];

l_M (lost time) and g_{minM} (minimum green time) = main road movements.

In this formula, ($l_M + g_{minM}$) is used as an approximation to ($I_M + G_{minM}$) where I_M is the intergreen and G_{minM} is the displayed minimum green time for the main road. Note that there is no maximum green time constraint for the main road.

The average green time for the side street (g_S) can be estimated from Equations 9 through 14.

The saturated portion of the side street green period (g_{sS}) can be estimated by using the red time for the side street (r_S) calculated from

$$r_s = g_M + L \tag{16}$$

where L is the total intersection lost time, which is the sum of lost times for the main road and the side street ($L = l_M + l_S$).

The cycle time is given by

$$c = g_M + g_s + L \tag{17}$$

The operation of pedestrian-actuated signalized pedestrian crossings (without any actuations for vehicle traffic) is similar to the semiactuated signal operations and can be analyzed by simply replacing the side street with the pedestrian movement. If the vehicle stream is also detected, the analysis method for fully actuated signals applies. Pedestrian green time is always $g = g_{minP}$, but g_{minP} can be modified to allow for no pedestrian arrivals during some signal cycles.

Linked Actuated Signals

Although the method for analyzing actuated signal operations described in this paper is applicable to isolated intersections, many aspects of the method also apply to linked actuated operations (i.e., to the operation of actuated signals that are part of a coordinated system). However, several important issues should be taken into consideration; some of them will be discussed briefly.

The headway distribution models described by Equations 1 through 5 are based on the assumption of random arrivals. For platooned arrivals as they occur at linked signals, separate arrival flow rates for main platoons and between main platoons could be used. The arrival headways within platoons can be used as constant headways, but they are not necessarily the same as the minimum arrival headway (Δ).

Similarly, proportions of traffic arriving during green and red periods can be used separately. For example, in a well-coordinated system, a higher proportion of traffic will arrive during green (and in a well-defined platoon). Whereas this high flow rate is relevant to the analysis of extension times, the lower arrival flow rate during the red period is relevant to the determination of minimum green time. Residual queues that can form at the end of the green period because of a particular signal offset should also be considered in determining the minimum green times and extension times.

Linked actuated systems work under a master cycle time with certain amounts of green time preallocated to some phases to guarantee signal offsets for uninterrupted progression of main platoons. Methods given in previous sections need to be extended to deal with this type of operation. One particular extension of the method is the allocation of any excess green time to nominated movements (phases). This excess green time would result from the imposition of a master cycle that may be longer than the cycle time under isolated actuated operation. The green split priority method used in the SIDRA program (*2*) is suitable for this purpose.

The semiactuated signals and pedestrian-actuated crossings provide simple cases for which all excess green time is allocated to the main road. For example, the excess green time (g_x) for a semiactuated signal site can be calculated from

$$g_x = c - (g_s + g_M + L) \tag{18}$$

where c is the master cycle time.

If g_x is positive, the main road green time can be adjusted by this amount ($g'_M = g_M + g_x$ so that $g'_M + g_s + L = c$). This would increase the queue clearance time for the side street (g_{sS}) and therefore the green time would increase. As a result, an iterative method would be required for estimating the green times. A negative g_x indicates that the master cycle time is insufficient for the operation of this site, but this could be accommodated by adjusting the side street green time down, for example, by setting a lower maximum green time.

Similarly, all excess green time at a pedestrian-actuated crossing can be allocated to the main road, but a negative excess time is not acceptable since the pedestrian movement operates at minimum green time.

Cycle Time for Actuated Signals

The green times estimated using the methods described previously can be used to calculate the cycle time. For this purpose, it is necessary to identify the critical movements, that is, the movements that require the longest green (and lost) times. The critical movement identification method described by Akçelik (*1*) for fixed-time signals and implemented in the SIDRA program (*2*) can be adapted for this purpose. The method allows for complex cases of movement overlaps and accommodates the movements whose green times are set to minimum or maximum values.

For application of the critical movement identification method to the case of actuated signals, average green time estimates can be used as required green times. The method compares the sums of required green and lost times for all combinations of movements and determines the critical movements as the set of movements that require the greatest sum of required green and lost times. This total time is equivalent to the cycle time (except in the case of linked actuated signals):

$$c = \Sigma(g_i + l_i) \tag{19}$$

where g_i and l_i are the green time and lost time for ith critical movement.

The formulas given in previous sections for determining the queue clearance time indicate that the required green time for a movement depends on the red time (or the cycle time); this effect is stronger with controllers using gap-reduction methods. Therefore, the green and red times for conflicting movements become interdependent, which requires iterative computations. In fact, even the critical movements may change as cycle time changes. Furthermore, saturation flows may change with green time and cycle time because of such factors as opposed turns, lane blockages and short lanes. This situation also necessitates the use of an iterative method. However, it is no different from the analysis of fixed-time signal operations, and such a method is already implemented in SIDRA (*2*).

The method to estimate the average cycle time at actuated signals can be enhanced by using adjusted minimum green times for all movements to allow for the possibility of no vehicle arrivals during the signal cycle (phase-skipping under low flow conditions), as with pedestrian movements.

For single green periods for all movements (i.e., no residual queues), the following formula can be derived from Equations 12 and 19 for estimating the average actuated signal cycle time as a more direct but limited version of Equation 19:

$$c = \frac{L + G_m + E}{1 - Y'}$$

subject to

$$Y' < 1.0 \tag{20}$$

where

L = intersection lost time (sum of all critical movement lost times)

$= \Sigma l_i$

G_m = sum of minimum and maximum green times ($g_{i\min}$, $g_{i\max}$) for critical movements whose green times are set to $g_i = g_{i\min}$ or $g_i = g_{i\max}$ (therefore not included in the summation for Y' or E)

$$= \Sigma(g_{i\min} + g_{i\max}) \quad (22)$$

E = adjusted extension time for intersection

$$= \Sigma(1 - y_i)\, e_{gi} \quad (23)$$

where $y_i = q_i/s_i$ is the flow ratio for ith critical movement (q_i = arrival flow, s_i = saturation flow) and the summation is for critical movements excluding those with green times set to $g_{i\min}$ or $g_{i\max}$; and

Y' = adjusted flow ratio for intersection

$$= \Sigma f_{qi} y_i \quad (24)$$

where f_{qi} is a queue clearance time calibration factor as in Equations 11 and 12 and the summation is for critical movements excluding those with green times set to $g_{i\min}$ or $g_{i\max}$.

For linked actuated signals, a predetermined cycle time is used and excess time is allocated to specified movements (Equation 18).

Fixed-Time Signal Settings

All of these parameters, except for extension time (e_g), are used for calculating cycle time and green times for fixed-time (pretimed) signals (*1–4*). For comparison, the method for fixed-time signals is also given here.

The practical cycle and green time method for computing fixed-time signal settings (*1,2*) tries to achieve specified target (practical) degrees of saturation (x_p) for critical movements. Usually, the principle of equal degrees of saturation (the same x_p value for all movements) is used. A more general method is to allow for different x_p values to be specified for different movements (e.g., 0.90 for the main road and 0.95 for the side road). This method can be expressed in a form consistent with the method given for estimating actuated signal timings:

$$g_i = u_{pi} c_p$$

subject to

$$g_{i\min} \leq g_i \leq g_{i\max} \quad (25)$$

$$c_p = \frac{L + G_m}{1 - U_p'}$$

subject to

$$U_p' < 1.0 \quad (26)$$

where

c_p = practical cycle time;

G_m = sum of minimum and maximum green times as in Equation 22;

U_p' = adjusted green time ratio for intersection (sum of required green time ratios for critical movements excluding those with green times set to $g_{i\min}$ or $g_{i\max}$):

$$= \Sigma u_{pi} \quad (27)$$

where $u_{pi} = y_i/x_{pi}$ is the green time ratio for ith critical movement (y_i = flow ratio, x_{pi} = practical degree of saturation) and the summation is for critical movements excluding those with green times set to $g_{i\min}$ or $g_{i\max}$.

Analyses of actuated signal timings using the method described in this paper indicate that equal degrees of saturation do not necessarily result in actuated signal cases, and the choice of 0.95 as a practical degree of saturation for actuated signals (x_p = 0.95) recommended by the 1985 HCM (*3*) is not substantiated. The latter point is important in relation to the development of an appropriate delay model for actuated signals.

The cycle time formula given in the HCM is a simpler version of Equation 26, obtained by ignoring the existence of minimum and maximum green times (a major shortcoming) and assuming the use of equal degree of saturation for all movements.

Example

As a very simple example for vehicle-actuated and fixed-time cycle times, a two-phase case is considered with a single movement in each phase. This occurs at the intersection of two one-way streets with three lanes each. Equal lane utilization is assumed, and equal conditions are assumed for both approaches:

Δ = 0.5 sec

b = 1.0 (multilane case)

s = 1,500 veh/hr/lane

l = 5 sec

$g_{\min}$ = 8 sec

$g_{\max}$ = 50 sec

e_o = 3.5 sec

x_p = 0.90

Therefore,

$L = 2 \times 5 = 10$ sec

$c_{\min} = 2 \times g_{\min} + L = 2 \times 8 + 10 = 26$ sec

$c_{\max} = 2 \times g_{\max} + L = 2 \times 50 + 10 = 110$ sec

Figure 5 shows the cycle time (c) as a function of the total intersection flow (twice the approach flow) for vehicle-actuated as well as two fixed-time signal settings (practical cycle time with x_p = 0.90, and minimum delay cycle settings obtained using SIDRA). With all methods in this example, the green times for the

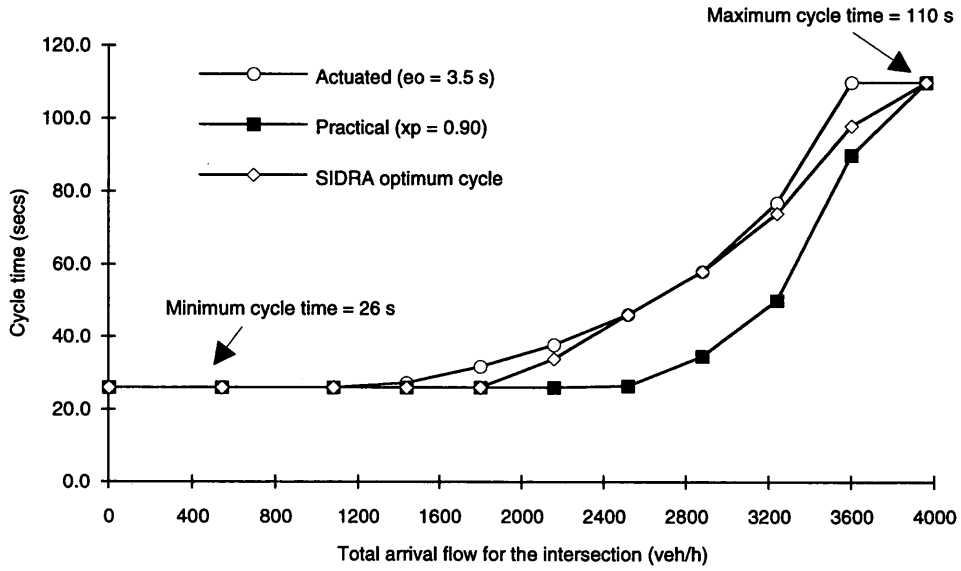


FIGURE 5 Cycle time (c) as a function of the total intersection flow for vehicle-actuated (e_o = 3.5) and fixed-time signals (practical cycle with x_p = 0.90, and SIDRA minimum-delay cycle).

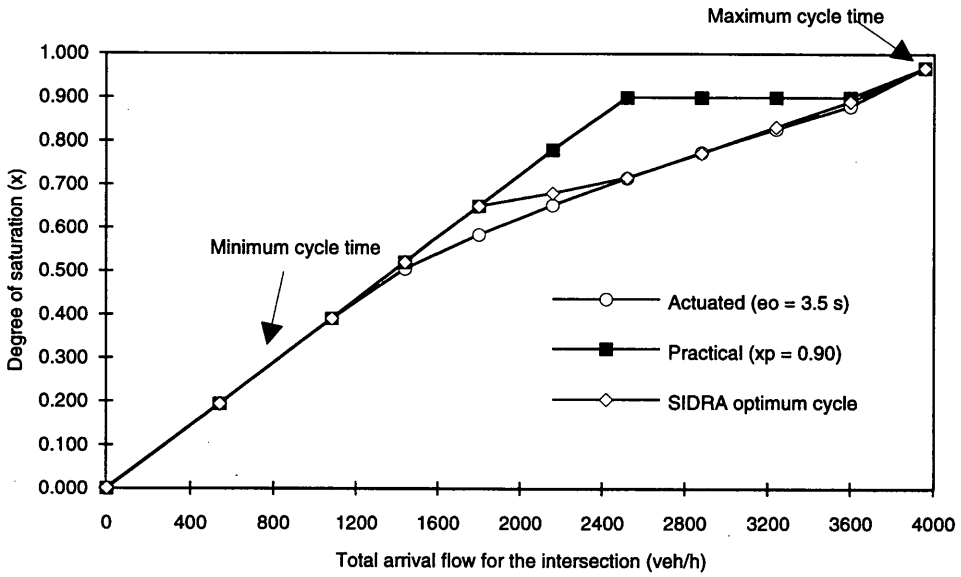


FIGURE 6 Intersection degree of saturation (x) as a function of total intersection flow for vehicle-actuated (e_o = 3.5) and fixed-time signals (practical cycle with x_p = 0.90, and SIDRA minimum-delay cycle).

two phases are equal. The minimum green times are used for low flows, and maximum green times are used for the high flow point in the graph. The intersection degree of saturation (largest x for any movement; $x = q/Q$ where capacity $Q = s_g/c$) corresponding to Figure 5 is shown in Figure 6.

CONCLUSION

An analytical method for estimating average green times and cycle time at actuated signals has been presented. The method can be seen as an extension of the current Australian, HCM, United Kingdom, and similar methods (*1–4*) for the analysis of fixed-time signals. The analysis method presented in this paper can be implemented manually. However, implementation through computer software such as SIDRA (*2*) is useful for dealing with complex intersection geometry and phasing arrangements and for obtaining solutions that require iterations.

The discussion in this paper is limited to the operation of a basic actuated controller that uses passage detectors and a fixed gap time setting (fully and semiactuated control cases). However, the method can be applied to the operation of more complex actuated signal

controllers using various gap reduction techniques with passage and presence detectors (*5*).

Model M3A for the proportion of bunched traffic has been calibrated for single-lane and multilane cases using real-life and simulation data after the writing of this paper (*15*). Validation and calibration of the green time and cycle time formulas given in this paper as well as the formulas for estimating intersection performance (delay, queue length, number of stops) are also needed.

The method given in this paper provides an analytical tool for investigating the optimization of actuated signal operations (to minimize delay, queue length, number of stops, or a performance index). This can be done with relative ease compared with the use of a simulation model, which is the most common method used for this purpose.

Many actuated controller parameters can be considered for optimization, namely, minimum green, gap time, maximum extension (or maximum green), additional parameters for gap reduction, and the location and other characteristics of detectors. Some suggestions are available on the effect of a fixed gap time setting in a basic actuated controller based on a limited amount of research published in the literature (*4,6,16–18*). These are summarized by Akçelik (*5*), and work is in progress to investigate the validity of these suggestions by means of simple examples reported in the literature.

ACKNOWLEDGMENT

The author thanks Ian Johnston, Executive Director of the Australian Road Research Board, for permission to publish this article.

REFERENCES

1. Akçelik, R. *Traffic Signals: Capacity and Timing Analysis*. Research Report ARR 123. Australian Road Research Board, Nunawading, 1981.
2. Akçelik, R. *Calibrating SIDRA*, 2nd ed. Research Report ARR 180. Australian Road Research Board, Nunawading, 1993.
3. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
4. Webster, F. V., and B. M. Cobbe. *Traffic Signals*. Her Majesty's Stationery Office, London, England, 1966.
5. Akçelik, R. *Analysis of Vehicle-Actuated Signal Operations*. Working Paper WD TE 93/007. Australian Road Research Board, Nunawading, 1993.
6. Staunton, M. M. *Vehicle Actuated Signal Controls for Isolated Locations*. An Foras Forbartha, Dublin, Ireland, 1976.
7. Lin, F.-B. Estimation of Average Phase Durations for Full-Actuated Signals. In *Transportation Research Record 881*, TRB, National Research Council, Washington, D.C., 1982, pp. 65–72.
8. Lin, F. B. Predictive Models of Traffic-Actuated Cycle Splits. *Transportation Research*, Vol. 16B, No. 5, 1982, pp. 361–372.
9. Cowan, R. J. Useful Headway Models. *Transportation Research*, Vol. 9, No. 6, 1975, pp. 371–375.
10. Troutbeck, R. J. Average Delay at an Unsignalized Intersection with Two Major Streams Each Having a Dichotomized Headway Distribution. *Transportation Science*, Vol. 20, No. 4, 1986, pp. 272–286.
11. Troutbeck, R. J. *Evaluating the Performance of a Roundabout*. Special Report 45. Australian Road Research Board, Nunawading, 1989.
12. Troutbeck, R. J. Recent Australian Unsignalised Intersection Research and Practices. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, Berlin, Germany, 1991, pp. 239–257.
13. Tanner, J. C. A Theoretical Analysis of Delays at an Uncontrolled Intersection. *Biometrika*, Vol. 49, Nos. 1 and 2, 1962, pp. 163–170.
14. Tanner, J. C. The Capacity of an Uncontrolled Intersection. *Biometrika*, Vol. 54, Nos. 3 and 4, 1967, pp. 657–658.
15. Akçelik, R., and E. Chung. Calibration of the Bunched Exponential Distribution of Arrival Headways. *Road and Transport Research*, Vol. 3, No. 1, 1994, pp. 42–59.
16. Grace, M. J., R. W. J. Morris, and P. G. Pak-Poy. Some Aspects of Intersection Capacity and Traffic Signal Control by Computer Simulation. *Proc., 2nd Australian Road Research Board Conference*, Vol. 2, No. 1, 1964, pp. 274–304.
17. Morris, R. W., and P. G. Pak-Poy. Intersection Control by Vehicle-Actuated Signals. *Traffic Engineering and Control*, Vol. 9, No. 6, 1967, pp. 288–293.
18. Pak-Poy and Associates. *A Comparison of Road Traffic Signal Controllers Using Simulation Techniques*. Research Report ARR 10. Australian Road Research Board, Nunawading, 1975.

The views expressed are those of the author and not necessarily those of the Australian Road Research Board.

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

Overflow Delay Estimation for a Simple Intersection with Fully Actuated Signal Control

JING LI, NAGUI M. ROUPHAIL, AND RAHMI AKÇELIK

Queueing delay at a traffic signal can be generally estimated as the sum of two components, uniform and overflow. The delay formula in the 1985 *Highway Capacity Manual* (HCM) applies primarily to lane groups under pretimed control. Although the HCM contains a method for estimating cycle length and splits under vehicle-actuated operation, the resulting effect on delays has yet to be verified. Furthermore, the HCM assumption of "snappy" operation and its inability to compare pretimed and actuated control have been criticized in the literature. An approach for estimating overflow delays for lane groups under vehicle-actuated control using the current HCM delay model format is presented. An existing cycle-by-cycle simulation model has been modified to produce delay for a basic vehicle-actuated signal operation. Overflow delay is computed as the difference from total simulated delay minus estimated uniform delay for the average cycle conditions. The results indicate that the average cycle and overflow delays are very much related to the controller settings such as minimum and maximum greens and cycles and unit extensions, with longer unit extensions producing higher cycle length and overflow delay. Furthermore, applying the 1985 HCM formula to the simulated signal settings resulted in much higher delays, which implies the need for separate calibration of the second delay term to account for the actuated control effects. The simulation model was executed to produce a calibration data base for an analytical overflow delay model.

In many traffic signal installations, the two most common types of intersection control are pretimed and vehicle actuated. In fact, some modern controllers can implement any combination of both controls depending on the level of traffic demand and the need to provide signal coordination. Actuated control schemes are typically classified into semiactuated, fully actuated, and volume-density control (*1*). In all schemes, phase green time is allocated to the different movements on the basis of the prevailing traffic demand. The three actuated control schemes vary in the amount of detectorization and in the establishment of criteria for phase termination (*1*). In contrast, pretimed control is established on the basis of average demand and, therefore, is often unable to respond adequately to random fluctuations in traffic volumes and demand variations on a cycle-by-cycle basis.

To establish capacity and level of service (LOS) impacts of actuated control operation, the 1985 *Highway Capacity Manual* (HCM) provides recommendations in Appendix 2 of Chapter 9 regarding the method for estimating the "average" cycle length and green splits in the peak 15-min period (*2*). This step is critical to the operational analysis procedure since signal timing settings are known for neither existing (barring actual field observations) nor projected conditions. These estimates are subsequently used to produce stopped delay and LOS.

Two fundamental issues arise in the HCM estimation process: (a) how realistic are the estimates of average cycle and splits for actuated control? and (b) is the current HCM delay equation, and in particular the overflow delay term, valid for both pretimed and fully actuated control, or are separate calibrations warranted? In this work, the focus is on the latter. Results from the literature are also presented that shed more light on the first issue.

The analysis presented in this paper applies to a basic vehicle-actuated signal controller that uses a fixed-time extension (gap time) setting and passage detection. A detailed analytical treatment of this type of controller as well as more sophisticated modern controllers that use gap-reduction and various density techniques is presented by Akçelik (*3*).

REVIEW OF 1985 HCM METHOD

Stopped delay is the principal performance measure for assessing the LOS of signalized intersections. In the case of fully actuated signalized lane groups, the average approach delay per vehicle in the 1985 HCM can be estimated according to the following:

$$d = (d_1 + d_2)\,\mathrm{PF} \tag{1}$$

$$d_1 = \frac{0.5C_{av}(1 - \lambda_{av})^2}{(1 - \lambda_{av} X_{av})} \tag{2}$$

$$d_2 = 900\,TX_{av}^2\left[(X_{av} - 1) + \sqrt{(X_{av} - 1)^2 + \frac{mX_{av}}{QT}}\right] \tag{3}$$

where

d = average approach delay per vehicle;
d_1 = average uniform delay per vehicle;
d_2 = average overflow delay per vehicle;
PF = progression factor;
C_{av} = cycle length (sec);
λ_{av} = g/C, ratio of effective green to cycle length;
X_{av} = v/c, degree of saturation, ratio of arrival flow rate to capacity;
m = calibration parameter (m = 4 in 1985 HCM);
Q = capacity [vehicles per hour (vph)]; and
T = flow period (hr) (T = 0.25 in 1985 HCM).

The progression factor PF = 0.85 reduces the queueing delay to account for the more efficient operation with fully actuated opera-

J. Li, Urban Transportation Center, 117 Lakeshore Court, Richmond, Calif. 94804. N. M. Rouphail, Department of Civil Engineering, North Carolina State University, P.O. Box 7908, Raleigh, N.C. 27695. R. Akçelik, Australian Road Research Board, P.O. Box 156, Nunawading 3131, Australia.

tion when compared with isolated, pretimed control. In upcoming revisions to the HCM Chapter 9 procedures, the progression factor will be applied to the uniform delay term only. Finally, stopped delay, d_s, can be estimated using the approximation $d_s = d/1.30$.

Because delay estimation requires knowledge of signal timings in the average cycle, the HCM provides a simplified estimation method. The average signal cycle length is computed from

$$C_{av} = \frac{LX_c}{X_c - \sum (v/s)_{ci}} \quad (4)$$

where X_c equals critical volume-to-capacity (v/c) ratio under fully actuated control ($X_c = 0.95$ in HCM). For the critical lane groups (ci), the effective green

$$g_i = (v/s)_{ci}(C_{av}/X_c) \quad (5)$$

where s is the saturation flow rate.

Two major differences emerge between the estimation of delays for pretimed and fully actuated lane groups. In the latter, delays are reduced by 15 percent to account for the more efficient operation at the same v/c ratios. More important, the design v/c ratio, X_c, for lane groups under actuated control is higher than that of a comparable pretimed controller. This is the result of the typically shorter phase lengths associated with actuated control, in which right of way is transferred to the conflicting phases soon after the queue dissipates or demand for a conflicting phase preempts the current phase. On the other hand, the lower X_c for pretimed control is meant to provide a margin of safety to accommodate short-term variations in demand.

The assumption of "snappy" operation has been the subject of criticism in the literature. Lin, for example, compared the predicted cycle length from Equation 4 with field observations in Upstate New York (*5*). In all cases, the observed cycle lengths were higher than predicted whereas the observed X_c ratios were lower. Tarnoff (*6*) simulated the operation of fully actuated controllers in NETSIM (*7*). He found that delays were sensitive to and increased with the controller's unit extension, an indication that snappy operation may well depend on the actual controller's parameter settings. The association between delays and controller's parameter settings was also pointed out independently by Akçelik (*8*), Santiago (*9*), and Skabardonis (*10*). In a recent paper, Akçelik (*3*) derived a cycle length formula for vehicle-actuated signal control allowing for minimum and maximum green time settings.

Finally, the proposed delay model in Equation 1 appears to violate a well-known principle in signal systems control: basic fully actuated controllers (i.e., with no skip phasing or gap reduction features) behave as pretimed controllers under very light or very heavy traffic flow conditions. Under light flow conditions, phase green times are dictated by the controller's minimum greens; under heavy flow conditions, all phases "max out." An examination of Equation 1 reveals that regardless of demand level, the actuated controller always outperforms its pretimed counterpart. In reality, delays will be similar between the two types of control for very high and very low v/c ratios assuming that the minimum and maximum signal timing parameters for the actuated control case are equivalent to those for the fixed-time controllers. For intermediate flow conditions, delay benefits can be expected with actuated control, and only when the controller parameters are set properly. This problem is addressed to a certain extent in the revised Chapter 9 method. In this revision, a delay factor is applied to the first (uniform delay) term only, and thus overall delays for pretimed and actuated operation will tend to converge at high v/c ratios, since the second term governs in that region.

In recognition of the existing deficiencies in the 1985 HCM with regard to actuated control operation, NCHRP has initiated a research project to address many of the stated problems (*11*).

METHODOLOGY

Delay Model Framework

The proposed approach uses the delay model format in the 1985 HCM (Equation 1) to estimate delay under fully actuated control, with some notable variations to both the uniform and overflow delay terms.

1. The progression factor is taken out of the formulation of delay model. Since the objective is to study the effect of signal settings on delay estimates, the first term is considered to be identical to the pretimed control, except that it uses the average rather than the fixed signal settings. The effect on the second term is considered in the calibration term m ($m = 4$ and $T = 0.25$ in 1985 HCM Equation 3), as discussed earlier.
2. The multiplier X^2 is taken out of the formulation of the overflow delay term. This is consistent with previous comments regarding the desirability of convergence at high v/c ratios for all types of control. By eliminating this term, the relationship between the steady-state and the time-dependent forms of the delay model using the coordinate transformation method (*12,13*) is preserved. [For more details on this issue, see the work by Akçelik (*3,8,14,15*) and by Akçelik and Rouphail (*16,17*).] Finally, the proposed form allows for direct comparison of the resultant delay models with their pretimed counterpart, calibrated in previous work (*4,18*).

To summarize, the steady-state form of the overflow delay model is derived from the principles of queueing theory, assuming a generalized service time distribution, and a random arrival distribution

$$d_2 = \frac{kX}{Q(1 - X)} \quad (6)$$

where k is a calibration parameter and Q is the lane group capacity. The corresponding time-dependent formulation of the model given in Equation 6, obtained by using the coordinate transformation method, is

$$d_2 = 900T\left[(x - 1) + \sqrt{(X - 1)^2 + \frac{mX}{QT}}\right] \quad (7)$$

of which Equation 3 is a special case with $m = 4$, $T = 0.25$ hr, and the X^2 term is deleted. A detailed treatment of the subject of the coordinate transformation is outside the scope of this paper. Interested readers are referred elsewhere (*12,13*). Akçelik (*14*) noted that the parameters k in Equation 6 and m in Equation 7 are related such that $m = 8k$.

Simulation Model

An existing discrete, macroscopic dynamic cycle-by-cycle simulation model has been adapted to model timings and delays at an inter-

section with two single-lane approaches and two-phase basic fully actuated control. Vehicles are represented as individual (discrete) entities, but delays are computed for groups of vehicles having the same properties (hence macroscopic). Details of the model operation and assumptions under pretimed control at isolated intersections have been discussed in a recent paper (*18*). Here, the authors focus on the variations that were implemented to model actuated control operation.

Vehicle Generation

Because cycle and green times are unknown, the simulated number of arrivals per cycle cannot be determined a priori. In the revised model, arrivals were estimated on the basis of the maximum controller settings for the purpose of establishing an arrival flow rate in each cycle. The arrival rates are then used to determine the appropriate phase lengths.

Basic Phase Length

The basic green time in Cycle i needed to discharge the initial queue as well as the new arrivals in this cycle is estimated from

$$g_{ib} = \frac{EOQ_{i-1} + v_i r_i}{S_i - V_i} \tag{8}$$

where

EOQ_{i-1} = queue length at end of Cycle $i - 1$,
v_i = average arrival rate during Cycle i,
r_i = effective red in Cycle i, and
s_i = saturation flow rate for approach lane during Cycle i.

In other words, the basic green time g_{ib} is equivalent to the saturated portion of the green period.

It is cautioned that while Equation 8 assumes a fixed saturation flow rate, normal variations in queue discharge headways could lead to the premature termination of the phase, particularly when short unit extensions are used. This possibility is not accounted for in the development of the delay models presented here, but it can certainly affect their validity and limit their applicability. Further work on incorporating this effect in the simulation model is planned.

With a two-phase controller, the effective red r_i is equivalent to the effective green of the competing phase plus the total lost time in the cycle (L).

Actual Phase Length

Although in theory the right of way should yield to a competing movement as soon as the basic phase length expires, in reality the green time is extended as long as vehicles are detected at headways that are shorter than the preset unit extension. This assumes, of course, that neither the minimum nor the maximum settings apply. This green extension is, therefore, dependent on both the prevailing headway distribution and the unit extension. For a single-lane case it may be reasonably assumed that headways follow the shifted negative exponential distribution, with a mean headway of $1/v_i$, v_i = arrival flow rate in vehicles per second, and a minimum headway of Δ. When the green extension is measured from the time that the basic green ends, this extension is equivalent to the length of a block of consecutive headways each greater than or equal to the unit extension, followed by a headway greater than the unit extension. Thus, at a minimum, the green will be extended by one unit extension.

Define this green extension for cycle i as E_i. It can be shown that, under the stated assumptions, the average E_i can be estimated from probability theory as

$$E_i = -\left(\frac{1}{v_i} - \Delta\right) + \frac{1}{v_i} e^{\frac{UE-\Delta}{1/v_i-\Delta}} \tag{9}$$

This relationship is depicted graphically in Figure 1 with $\Delta = 2$ sec. It is shown that the higher the unit extension, the more sensitive is the extension time to the prevailing volumes.

To summarize, the phase length for cycle i in the simulation model is expressed as

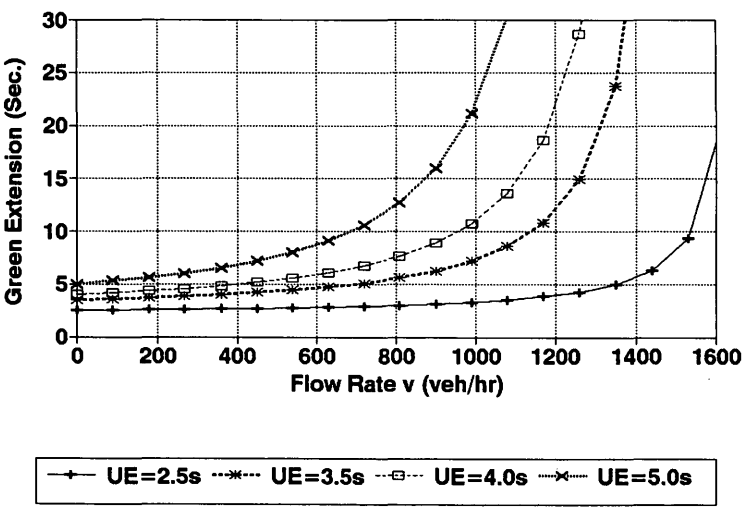


FIGURE 1 Extension green time E_i for indicated flow rate and unit extension.

$$g_i = g_{ib} + E_i = \frac{EOQ_{i-1} + v_i r_i}{s_i - v_i} + E_i \quad (10)$$

The effective phase green g_i is subject to constraints on minimum and maximum greens, which are either internally computed (assuming typical detector setback, speed limit, and unit extension) or entered as input into the simulation model. The cycle length is readily derived as $(r_i + g_i)$ for two phase operation. Note that g_i for the just-completed phase, when added to the lost time per cycle, constitutes the current effective red time for the competing phase. The process is then repeated for the competing phase and reverts back and forth between the two phases until the simulation time expires. A sample output of the simulation is provided in Figure 2. For this study, the minimum headway Δ is set to 2 sec for all simulation runs.

SIMULATION RESULTS

Signal Settings

Figures 3 and 4 depict the simulated cycle length and effective green times over a 2-hr simulation run. In both cases, the minimum green was set at 18 sec, maximum green at 56 sec, and unit extension at 2.5 sec. Equal flows on both approaches were simulated. In Figure 3 the resulting average *v/c* ratio is 0.86, whereas in Figure 4 it is 0.94 (both are less than the HCM's 0.95). The simulated cycle length and green exhibit a large degree of randomness, not too dissimilar in pattern from the field observations gathered by Prevedouros (*19*). In Figure 4 it is evident that many cycles and green times are reaching the maximum settings much more frequently than those depicted in Figure 3.

```
                MAJOR FLOW STATISTICS FOR 2 SIMULATED HOURS

**** INPUT ECHO DATA ****

MINIMUM CYCLE LENGTH      =   44        MAXIMUM CYCLE LENGTH    =  120
CONTROLLER UNIT EXTENSION =   2.5       SATURATION HEADWAY      =  2
FLOW IN VEHICLES PER HOUR =   800       MAXIMUM CAPACITY/CYCLE  =  28

**** SIMULATION RESULTS FROM 85 CYCLES ****

AVERAGE CYCLE LENGTH              =  85.85      S.DEVIATION =  26.65
SIMULATED INTERSECTION V/S RATIO =   .87        S.DEVIATION =  .16

AVERAGE GREEN TIME    =  41.05      S.DEVIATION =  13.82
AVERAGE RED TIME      =  44.79      S.DEVIATION =  15.25
SIMULATED V/C RATIO   =  .95        S.DEVIATION =  .11

 FLOW/CYCLE........MEAN =  19.84    S.DEVIATION =  8.01    I-RATIO=  3.23
 END OVERFLOW Q....MEAN =  1.52     S.DEVIATION =  3.53    END OF PERIOD= 0
 MAXIMUM QUEUE.....MEAN =  22.05    S.DEVIATION =  9.13
 DELAY/VEHICLE.....MEAN =  29.73    S.DEVIATION =  20.59

 PRESS ANY KEY TO CONTINUE
```

FIGURE 2 Sample simulation model output.

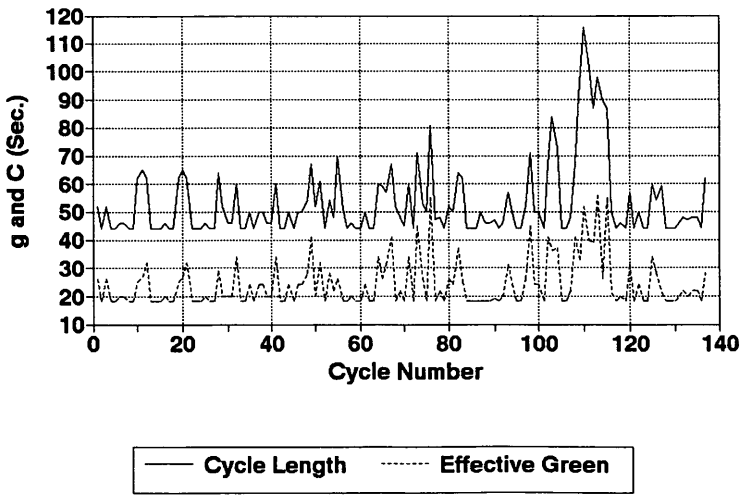


FIGURE 3 Effective green and cycle length for UE = 2.5, average flow rate = 700 vph.

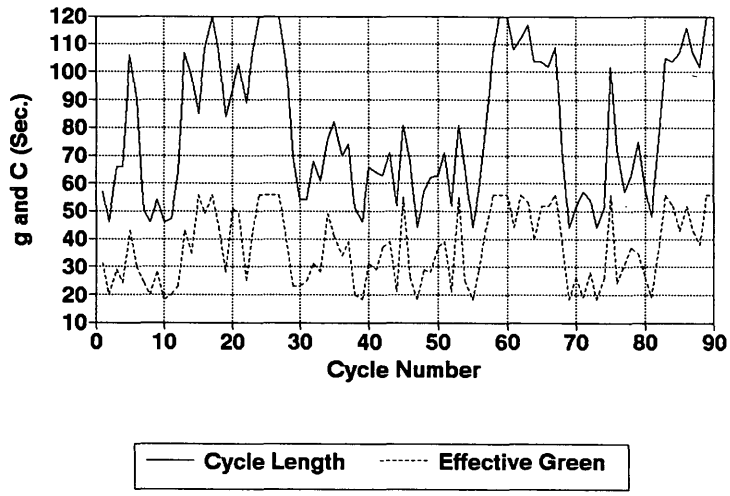


FIGURE 4 Effective green and cycle length for UE = 2.5, average flow rate = 800 vph.

In Figures 5 and 6, the average cycle length obtained from simulation is compared with those estimated from the HCM formula (Equation 4) with $X_c = 0.90, 0.95$, and 1.0, for unit extensions (UEs) of 2.5 and 4.0 sec, respectively. The minimum and maximum cycles were set at 44 and 120 sec for UE = 2.5 sec and at 60 and 120 sec for UE = 4.0 sec. Compared with the HCM formula, the simulated cycle lengths exhibit a more gradual increase in cycle length with intersection flow ratio. Also noted are the much longer cycle lengths associated with the longer unit extension in Figure 6.

Approach Delays

Simulated delays for a 2.5- and 5.0-sec unit extensions are compared in Figure 7. At low *v/c* ratios the delays are comparable, except for the effect of the different minimum green (which are higher for UE = 5.0 sec). Had the minimum greens been set equal for the two cases, there would have been no difference in delay. The two delay curves diverge in the region $0.78 < v/c < 0.95$. As *v/c* approaches 1.0, both curves converge to the maximum settings (and therefore equivalent delays). The results so far have confirmed both expectations and previous results relating delays to unit extensions by Tarnoff (*6*).

Delays were next compared with those estimated from the HCM delay formula for actuated controller [Equation 1, $(d_1 + d_2)$PF where PF = 0.85], using the simulated output values of average cycle, greens, and *v/c* ratios. The results are depicted in Figures 8 and 9 for unit extensions of 2.5 and 4.0 sec, respectively. The HCM overflow delays were adjusted for a 2-hr analysis period (*20*). The graphs also depict the uniform delay component (Equation 2). The HCM uniform and simulated delays appear to be comparable for *v/c* up to 0.78 for UE = 2.5 sec (Figure 8) and for *v/c* up to 0.80 for UE = 4.0 sec (Figure 9). Beyond that value, the HCM formula diverged considerably from the simulated and uniform delays for UE = 2.5 sec but

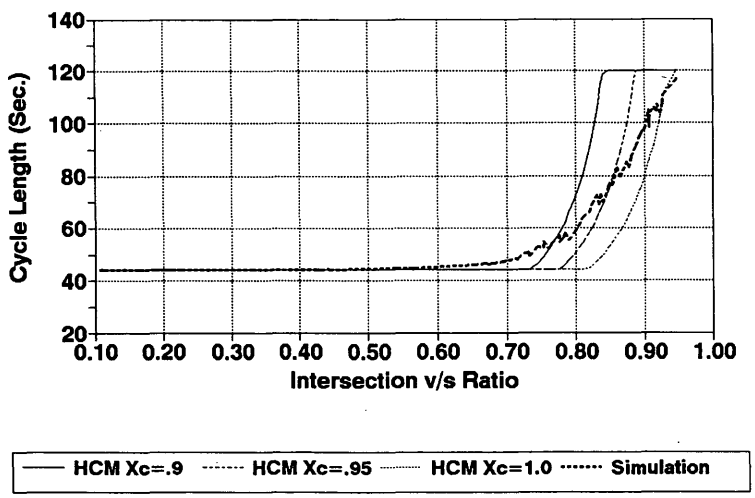


FIGURE 5 Average cycle length: HCM versus simulation for UE = 2.5 sec, L = 8 sec/cycle.

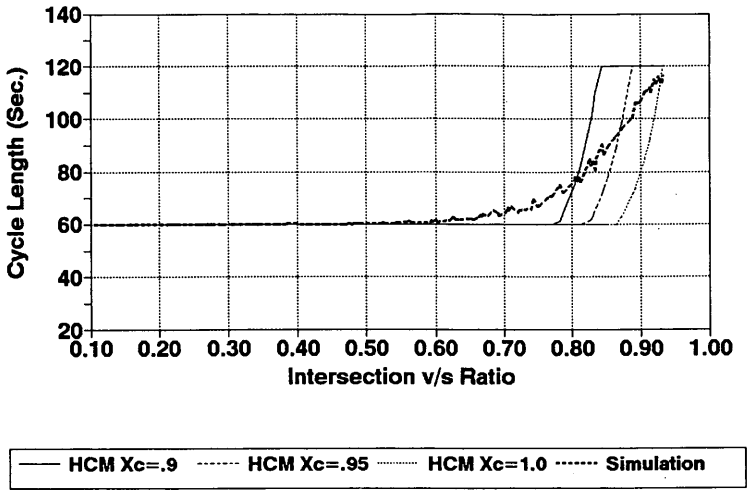


FIGURE 6 Average cycle length: HCM versus simulation for UE = 4.0 sec, L = 8 sec/cycle.

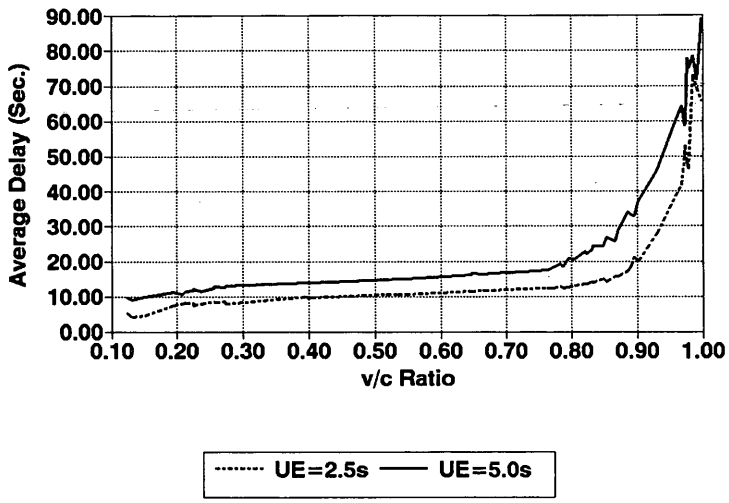


FIGURE 7 Simulated average delay for UE = 2.5 sec versus UE = 5.0 sec.

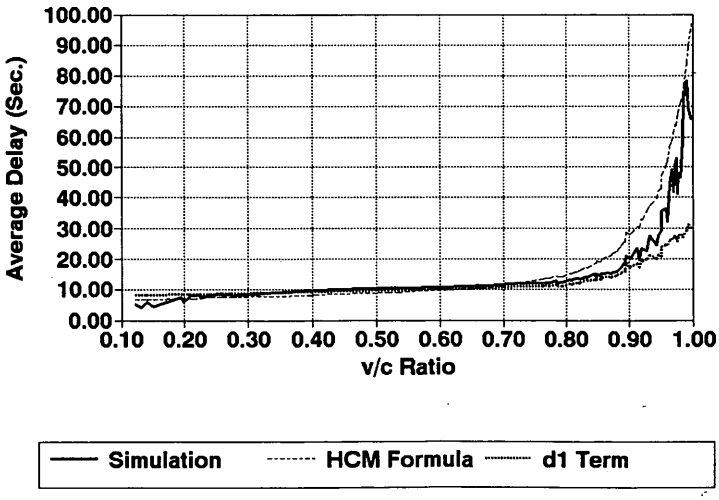


FIGURE 8 Simulated versus HCM delay for UE = 2.5 sec, L = 8 sec/cycle.

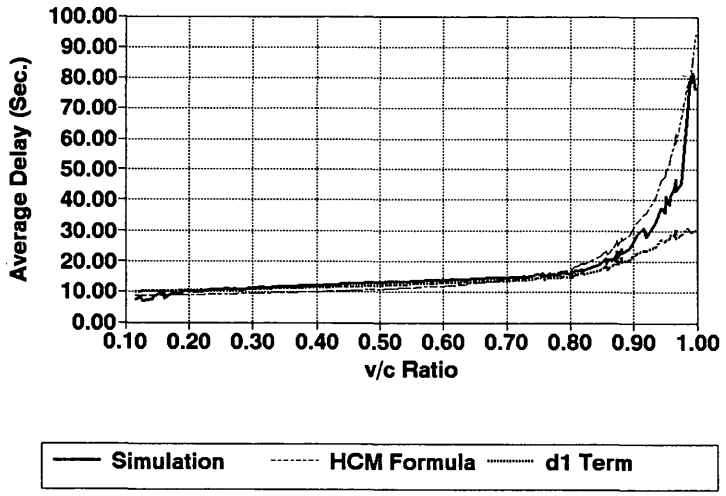


FIGURE 9 Simulated versus HCM delay for UE = 4.0 sec, L = 8 sec/cycle.

only slightly for UE = 4.0 sec. On the other hand, both the simulated and uniform delay curves tracked each other rather well for *v/c* up to 0.90 for UE = 2.5 sec and up to *v/c* = 0.88 for UE = 4.0 sec. The vertical distance between the simulated and uniform delay constitutes the overflow delay value. It is evident from these results that even when supplied with the proper values of average cycle, greens, and *v/c* ratios, the HCM formula tended to overestimate delay compared with the simulated results; the level of overestimation depends on the unit extension setting. In examining Figure 9, it is evident that further increases in the unit extension may actually bring the actuated controller delay close to the HCM delay estimate.

OVERFLOW DELAY MODEL CALIBRATION

Methodology

Referring to the section on delay model framework, the overflow delay model calibration proceeded in two steps. First, the uniform delay term described in Equation 2 was estimated. This requires estimates of the average cycle, green times, and *v/c* ratios. There are three ways of producing these data: first, and preferably, through field observations over a reasonable analysis interval (*19*); second, to derive them analytically using the HCM formula as in Equation 4 or from alternative formulas (*21*); third, to obtain them from a simulation model. Figures 8 and 9 have already indicated that when supplied with simulated values, Equation 2 produced uniform delay estimates that are virtually identical to the simulated delays at low volume conditions (i.e., when the overflow term is actually negligible). Thus, it was decided that Equation 2, using simulated signal timing parameters, is adequate for characterizing the uniform delay term d_1. The overflow delay term was simply estimated as the difference between the simulated approach delay and the computed uniform delay. Because this investigation has so far indicated a strong unit extension effect, separate models were calibrated, one for each unit extension (*22*). All calibrations were performed using the steady-state form of the overflow delay model given by Equation 6. Since only the single parameter k is needed to characterize the model, a simple, no-intercept regression modeling approach was used. Four data sets, each corresponding to a unit extension, were extracted from the simulation. Their characteristics are given in Table 1.

Results

The calibration results for the parameter k along with the overall statistical model evaluation criteria (standard error and R^2) are given in Table 2. The parameter k, which corresponds to pretimed control, calibrated in previous work (*4*) is also presented. It is worth noting that the pretimed steady-state model was also calibrated using the same cycle-by-cycle simulation approach but with fixed signal cycles and splits. The first and most apparent observation is that the pretimed model produced a k-value higher than the actuated models. Second, and as expected, the parameter was found to increase with the size of the unit extension.

TABLE 1 Calibration Data Base Description

Unit Extension(s)	Minimum Cycle(s)	Maximum Cycle(s)	Minimum v/c Ratio	Maximum v/c Ratio	Number of Observations
2.5	44	120	0.131	0.958	78
3.5	54	120	0.126	0.958	123
4.0	60	120	0.116	0.956	144
5.0	70	120	0.104	0.959	159

TABLE 2 Calibration Results for Steady-State Overflow Delay Function

		Traffic Actuated			
		Unit Extension (Second)			
	Pretimed[a]	2.5	3.5	4.0	5.0
Obs.	480	78	123	144	159
k	0.427	0.084	0.119	0.125	0.231
s.e.	NA	0.003	0.002	0.002	0.006
R^2	0.903	0.834	0.909	0.933	0.861
$m = 8k$	3.416	0.672	0.952	1.000	1.848

[a]See Reference (4) for details.

To evaluate the resulting overflow delay model, simple linear regression models were fitted between the predicted (as dependent variable) and simulated (as independent variable) delays for three levels of unit extension, as indicated in Table 3. The first data set, UE = 2.5, did not produce a good fit, with the intercept term significantly higher than 0 and the slope significantly lower than unity. Thus, this model would tend to overestimate delays at low *v/c* ratios and underestimate them at the high *v/c* ratios. On the other hand, the other two data sets produced excellent fits, with intercepts statistically 0 and slopes near unity.

Finally, the derived delay models are compared with the HCM model in the time-dependent form (*12*); they are depicted in Figure 10. Here the HCM formula is expressed by Equation 1; the actuated models apply to UE = 2.5, 3.5, and 5.0 sec and have the general form given by Equation 7. Furthermore, all comparisons are based on an analysis period $T = 0.25$ hr and for a lane group capacity of $Q = 500$ vph. The deterministic oversaturation delay, which applies at very high *v/c* ratios and constitutes the asymptote for all delay models, is also depicted, where

$$d_2 = 1{,}800T(X - 1) \tag{11}$$

It is evident that among all the indicated functions, the one produced by the HCM formula gave the highest overall delays. For the actuated delay functions, the delays were very similar for *v/c* ratios lower than 0.65 and for *v/c* ratios greater than 1.10 (when they all converge to the deterministic model). For values in between, delays were higher for the actuated models with the longer unit extensions.

SUMMARY AND CONCLUSIONS

This paper summarizes a first attempt at developing analytical delay models for traffic under basic actuated signal control using a fixed unit extension (gap time) setting and passage detection. The effort has been guided by what many perceive to be weaknesses in the present HCM methodology with regard to the operational analysis of this type of control. One unanswered question has been the quantification of the effect of actuated control on overflow delay, given that random queues can be better absorbed in an actuated system by virtue of the phase extension feature. Yet the 1985 HCM procedure applies a flat 15 percent delay reduction factor to both delay. Other points of concern include the apparent disconnect between the actuated controller parameters and the resulting signal efficiency, the inability to compare pretimed and actuated control, and methods for estimating the average signal parameters. A nationwide research study aimed at addressing a number of these problems is now under way.

A macroscopic, stochastic simulation model developed in earlier work was adapted for the study of capacity and delays for basic two-phase fully actuated operation. This simulation was previously used in the calibration of a pretimed overflow delay model (*18*). It is capable of modeling and estimating individual cycle lengths, phase times, and delays for up to 400 cycles.

Although the results of the study must be considered preliminary in nature, given the lack of field verification, they nevertheless point to some interesting and consistent trends. As well, many results appear to confirm data and trends found in the literature. To summarize, the following conclusions are offered:

1. The use of a fixed critical X_c ratio in estimating average signal timing parameters for fully actuated operation is not recommended. The appropriate value must be derived from the actual controller settings, such as unit extensions, minimum and maximum greens, and cycles. See the work by Akçelik (*3*), and the preceding paper in this Record) for a detailed derivation of signal parameters.
2. The use of the general overflow delay form given in Equation 7 is recommended. It guarantees convergence to the deterministic oversaturation delay irrespective of the type of control that is implemented. The calibrated models were based on that form.
3. Overflow delay was found to increase with an increase in unit extension, as represented by the parameter m in Equation 7. The increased delay is the consequence of higher cycle lengths and red times, leading to longer queues.
4. Delay differences for various unit extensions in the time-dependent form were not significant for *v/c* ratios lower than 0.65 and for *v/c* ratios greater than 1.10 and were quite close to those experienced under pretimed control. In most cases, the delays at high *v/c* ratios duplicated a pretimed signal at the maximum settings.
5. The calibrated models for actuated control delays yielded lower overflow delay values than the pretimed model. This is a

TABLE 3 Regression Results for Predicted Versus Simulated Delay

Unit Extension(s)	Variable	b_j	$s.e.(b_j)$	T or F	$p > T$ or F	R^2
	Intercept	3.713	0.456	8.133	0.0001	
2.5	Slope	0.830	0.018	86.453[a]	0.0001	0.965
	Intercept	0.065	0.577	0.133	0.9106	
3.5	Slope	1.076	0.027	8.211[a]	0.0053	0.951
	Intercept	0.298	0.696	0.428	0.6695	
5.0	Slope	1.036	0.029	1.465[a]	0.2286	0.909

[a] Test for Slope = 1 (F test).

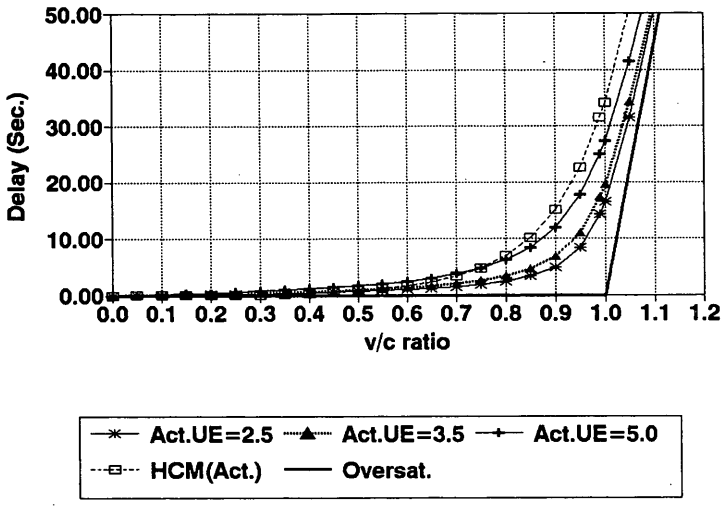


FIGURE 10 Overflow delay comparison time-dependent form, T = 0.25 hr, Q = 500 vph.

result of the actuated controller's ability to operate high *v/c* ratios without incurring substantial random queues and delays.

ACKNOWLEDGMENTS

This research is sponsored by FHWA under a contract to the Texas Transportation Institute and the University of Illinois at Chicago. The authors wish to thank Dan Fambro, principal investigator, for his assistance and leadership on the project, and Cesar Perez, FHWA project monitor, for his continued support.

REFERENCES

1. McShane, W. R., and R. P. Roess. *Traffic Engineering*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1990.
2. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
3. Akçelik, R. *Analysis of Vehicle-Actuated Signal Operations*. Working Paper WDTE93/007. Australian Road Research Board, Nunawading, July 1993.
4. Tarko, A., and N. M. Rouphail. *Overflow Delay at a Signalized Intersection Approach Influenced by an Upstream Signal*. Working Paper. Urban Transportation Center, University of Illinois at Chicago, Dec. 1991.
5. Lin, F.-B. Application of 1985 *Highway Capacity Manual* for Estimating Delays at Signalized Intersections. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 18–23.
6. Tarnoff, P. J. *NCHRP Report 233: Selecting Traffic Signal Control at Individual Intersections*. TRB, National Research Council, Washington, D.C., June 1981.
7. *TRAF User Reference Guide*. Office of Safety and Traffic Operations Research and Development, IVHS Research Division, FHWA, U.S. Department of Transportation, May 1992.
8. Akçelik, R. *Traffic Signals: Capacity And Timing Analysis*. Research Report ARR 123. Australian Road Research Board, Nunawading, 1981.
9. Santiago, A. J., and S. E. Smith. Evaluation of the Highway Capacity Manual Procedure for Signal Design. *Compendium of Technical Papers*, ITE 61st Annual Meeting, Milwaukee, Wisc., Sept. 1991.
10. Skabardonis, A. *Progression Through a Series of Intersections with Traffic Actuated Controllers*. Report FHWA-RD-89-133. Vol. 2, User's Guide, Oct. 1988.
11. *Capacity Analysis for Actuated Intersections*. Research Project Statement, NCHRP Project 3–48, FY '94. TRB, Washington, D.C., 1993.
12. Kimber, R. M., and E. M. Hollis. Peak Period Traffic Delay at Road Junctions and Other Bottlenecks. *Traffic Engineering and Control*, Vol. 19, No. 10, 1978, pp. 442–446.
13. Akçelik, R. *Time-Dependent Expressions for Delay, Stop Rate, and Queue Length at Traffic Signals*. Internal Report AIR 367–1. Australian Road Research Board, Nunawading, 1980.
14. Akçelik, R. The Highway Capacity Manual Delay Formula for Signalized Intersections. *ITE Journal*, Vol. 58, No. 3, March 1988, pp. 23–27.
15. Akçelik, R. *Calibrating SIDRA*, 2nd ed. Research Report ARR 180. Australian Road Research Board, Nunawading, 1993.
16. Akçelik, R., and N. M. Rouphail. Estimation of Delays at Traffic Signals for Variable Demand Conditions. *Transportation Research*, Vol. 27B, No. 2, 1993, pp. 109–131.
17. Akçelik, R., and N. M. Rouphail. Overflow Queues and Delays with Random and Platooned Arrivals at Signalized Intersections. *Journal of Advanced Transportation*, 1993.
18. Tarko, A., N. M. Rouphail, and R. Akçelik. Overflow Delay at a Signalized Intersection Approach Influenced by an Upstream Signal: An Analytical Investigation. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993.
19. Prevedouros, P. D. Actuated Signalization: Traffic Measurements and Capacity Analysis. *Compendium of Technical Papers*, ITE 61st Annual Meeting, Milwaukee, Wisc., Sept. 1991, pp. 234–238.
20. Fambro, D. B., J. Daniel, N. M. Rouphail, S. R. Sunkari, and J. Li. *Highway Capacity Manual*, 3rd ed., revised. TRB, National Research Council, Washington, D.C., Ch. 9 and 11 (in preparation).
21. Rouphail, N. M., and J. Li. *Calibration and Validation of Delay Models Under Fully Actuated Control*. FHWA Project DTFH61-92-R-00071 Working Paper 3. Urban Transportation Center, University of Illinois at Chicago, May 1993.
22. Rouphail, N. M., and J. Li. *Cycle-by-Cycle Analysis of Signal Delay and Parameters at Signalized Intersection Under Fully Actuated Control*. FHWA Project DTFH61-92-R-00071, Working Paper 2. Urban Transportation Center, University of Illinois at Chicago, April 1993.

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

Use of Default Parameters for Estimating Signalized Intersection Level of Service

RICHARD G. DOWLING

The 1985 *Highway Capacity Manual* (HCM) "operations" method for estimating the level of service of signalized intersections can require a large amount of field data: turning movement volumes, lane geometry, signal timing, approach grades, percentage heavy vehicles, number of parking maneuvers per hour, number of buses stopping per hour, peak hour factors, number of conflicting pedestrians per hour, arrival types, and right turns on red. The effects on accuracy of replacing most of these required input data with the default values recommended in Table 9-3 of the HCM are tested. The average stopped delay was calculated for six signalized intersections starting with basic volume (flow rate), lane geometry, and signal timing data. The HCM-recommended default values for grades, heavy vehicles, and such were used in place of the rest of the required input data. The calculations were then repeated several times; each time one or more of the default values were replaced with field data. The resulting delay estimates were then compared with field measurements of delay. The test results indicated that users can obtain reliable estimates of intersection level of service and delay using only field-measured turning movements, lane geometry, and signal timing plus the HCM-recommended defaults for the rest of the required input data. Field measurements of peak hour factors, grades, percentage heavy vehicles, parking maneuvers, number of stopping buses, conflicting pedestrians, and arrival type improved the accuracy of the delay estimates, but the improvements were comparatively minor and did not change the estimated intersection level of service. Delay estimates for intersections with critical volume-capacity ratios of less than 85 percent of capacity were insensitive to additional data on peaking, arrival type, and saturation flow rates.

The 1985 *Highway Capacity Manual* (HCM) recommends an "operations" method for estimating the average stopped delay at signalized intersections (*1*). This method can require a great deal of data collection. Users need to know turning movement volumes, lane geometry, signal timing, approach grades, percentage heavy vehicles, number of parking maneuvers per hour, number of buses stopping per hour, peak hour factors (PHFs), number of conflicting pedestrians per hour, arrival types, and right turns on red. Collecting all of these data would require at least 4 person-hr per intersection and possibly twice that for more complex intersections.

How much of this information is really necessary? How does the use of defaults for most of the required data affect the accuracy of the HCM method? These are the questions that this paper is designed to answer.

Two signalized intersections in Oakland, California, were videotaped for 1 peak hr each. The average stopped delay for each intersection was then calculated using the HCM method starting with basic volume (flow rate), geometric, and signal timing information combined with the default parameter values recommended in Table 9-3 of the HCM. The calculation was repeated several times, and the default values were replaced gradually with more and more field-observed values. The results were then compared with the field-measured average stopped delay to determine how additional field data-improve the accuracy of the HCM operations method.

This procedure was also performed for four more signalized intersections that were contained in the 1982 validation data set for NCHRP Project 3-28[2], "Urban Signalized Intersection Capacity," which was the precursor to the current 1985 HCM method. This older data set was not as detailed, so not all of the tests performed on the two videotaped intersections could be repeated; however, it was possible to reproduce most of the tests.

The results for all six intersections were combined and evaluated to determine how additional data collection might improve the accuracy of the stopped delay and level-of-service estimates produced by the HCM operations method for signalized intersections.

Dowling Associates, 180 Grand Avenue, Suite 995, Oakland, Calif. 94612.

BACKGROUND

During the development of the current HCM procedure for estimating the capacity and level of service of signalized intersections, the validity and accuracy of the proposed method were tested many times.

Reilly developed the NCHRP 3-28[2] procedure using several data sets throughout the United States (*2*). He then validated the proposed procedure against 25 observations made at eight intersections in Arizona and California and found that the mean absolute error (MABS) in the estimate of average intersection stopped vehicle delay was 1.1 sec/vehicle. May later used five of these eight intersections to compare this procedure with other capacity analysis methods (*3*). The delay equation and the method for estimating saturation flow were subsequently modified before being included in Chapter 9 of the 1985 HCM.

Teodorvic tested the HCM method against 16 observations made on the approaches to five intersections in Delaware (*4*) and found that the HCM method overpredicted average stopped delay by an average of 12 sec on an approach basis. The MABS was 18 sec, with a standard deviation of 26 sec.

The actual error may have been higher since Teodorvic eliminated observations in which the calculated volume-capacity ratio (*v/c*) was greater than 1.2. He did keep two observations in which the estimated *v/c* was greater than 1.0, and these two observations accounted for most of the observed error. Since a queue cannot discharge faster than the actual capacity of the approach, one can conclude that the error may be attributed to an underestimation of actual saturation flow rates. This may have been due to a failure to measure ideal saturation flow rates in the field or to the HCM method's underestimation of actual approach capacity at these locations.

Lin compared the HCM estimated average stopped delay with field measurements for 20 approaches at seven intersections in New York State (*5*). He found the MABS to be 9.0 sec (on an approach

```
HCS:  Signalized Intersection    Version 2.1                              1
================================================================================
Center For Microcomputers In Transportation
University of Florida
512 Weil Hall
Gainesville, FL   32611-2083          (904) 392-0378
================================================================================
Streets: (E-W) Twenty-Seventh          (N-S) Harrison
Analyst: RGD                           File Name: HARRIS7C.HC9
Area Type: Other                       5-16-93 8-9
Comment: Vol+Geo+actual timing+PHF+Delay Adj+actual sat (1 lane NB LT)
================================================================================
```

Traffic and Roadway Conditions

	Eastbound			Westbound			Northbound			Southbound		
	L	T	R	L	T	R	L	T	R	L	T	R
No. Lanes	1	> 2	<		> 2	<	1	2	<	1	2	<
Volumes	53	150	120	81	218	125	286	418	36	81	1411	110
PHF or PK15	0.92	0.92	0.92	1.00	1.00	1.00	0.94	0.94	0.94	0.83	0.83	0.83
Lane Width	12.0	12.0			12.0		12.0	12.0		12.0	12.0	
Grade		0			0			0			-3	
% Heavy Veh	1	1	1	1	1	1	2	2	2	0	0	0
Parking	(Y/N)	Y	0	(Y/N)	Y	1	(Y/N)	Y	11	(Y/N)	N	
Bus Stops			0			0			7			4
Con. Peds			26			71			63			75
Ped Button	(Y/N)	N		(Y/N)	N		(Y/N)	Y	17.5 s	(Y/N)	Y	20.5
Arr Type	3	3	3	2	2	2	3	3	3	4	4	4
RTOR Vols			0			0			0			0

Signal Operations

Phase combination	1	2	3	4		5	6	7	8
EB Left	*				NB Left	*			
Thru	*				Thru	*	*		
Right	*				Right	*	*		
Peds	*				Peds	*	*		
WB Left	*				SB Left		*		
Thru	*				Thru		*		
Right	*				Right		*		
Peds	*				Peds		*		
NB Right					EB Right				
SB Right					WB Right				
Green	17P				Green	15P	39P		
Yellow/A-R	3				Yellow/A-R	3	3		
Lost Time	3.0				Lost Time	3.0	3.0		

```
Cycle Length:  80 secs  Phase combination order: #1 #5 #6
```

FIGURE 1 Harrison and 27th input data sheet.

basis). He investigated ways in which the progression adjustment factor and the second term of the delay equation contributed to this error, particularly for actuated signals. However, he did not go on to investigate ways in which the other parameters used in the HCM method to determine saturation flow would influence the result.

From these results, it appears that the current HCM method can be expected to predict average stopped delay with an average error of from 9 to 18 sec/vehicle on each approach. Data on mean error for entire intersections, however, are not available.

Several of the investigators are also unclear as to whether they used exclusively field-measured data or substituted some of the defaults recommended in Table 9-3 in their computations. There are no data on how the use of default parameters in lieu of some field data might significantly worsen the performance of the HCM method.

METHODOLOGY

Overview

The impact of default parameters on the accuracy of the HCM operations method was evaluated at six intersections in the San Francisco Bay Area of California. Four of these intersections were taken from the 1982 validation data set used by May and Reilly for testing the validity of the NCHRP 3-28[2] procedure. Another two were videotaped in Oakland in 1993 and the results combined with the results of the older data set evaluated by May.

The peak 15-min capacity and level of service were calculated for each intersection using the McTrans software, HCS2.1, released in 1990 (*6*) (Figures 1 through 6). The HCS-estimated average stopped delay for the entire intersection was compared with the field-

```
HCS:  Signalized Intersection    Version 2.1                              1
================================================================================
Center For Microcomputers In Transportation
University of Florida
512 Weil Hall
Gainesville, FL   32611-2083          (904) 392-0378
================================================================================
Streets: (E-W) Webster                 (N-S) Grand
Analyst: RGD                           File Name: WEB7.HC9
Area Type: Other                       3-10-93 4:42PM
Comment: vol + geo + actual timing+PHF+delay adj.+ actual sat
================================================================================
```

Traffic and Roadway Conditions

	Eastbound			Westbound			Northbound			Southbound		
	L	T	R	L	T	R	L	T	R	L	T	R
No. Lanes					> 2 <		1	2 <			> 2 <	
Volumes				110	280	66	127	336	18	25	747	256
PHF or PK15				0.83	0.83	0.83	0.97	0.97	0.97	0.90	0.90	0.90
Lane Width					12.0		12.0	12.0			12.0	
Grade					0			0			0	
% Heavy Veh				0	0	0	2	2	2	0	0	0
Parking				(Y/N)	Y	3	(Y/N)	Y	1	(Y/N)	Y	4
Bus Stops						0			4			13
Con. Peds			0			118			29			68
Ped Button				(Y/N)	Y	17.5 s	(Y/N)	Y	11.5 s	(Y/N)	Y	13.0
Arr Type				2	2	2	4	4	4	4	4	4
RTOR Vols						0			0			0

Signal Operations

Phase combination	1	2	3	4		5	6	7	8
EB Left					NB Left	*			
Thru					Thru	*	*		
Right					Right	*	*		
Peds	*				Peds		*		
WB Left	*				SB Left		*		
Thru	*				Thru		*		
Right	*				Right		*		
Peds	*				Peds		*		
NB Right					EB Right				
SB Right					WB Right				
Green	27P				Green	10P	34P		
Yellow/A-R	3				Yellow/A-R	3	3		
Lost Time	3.0				Lost Time	3.0	3.0		

Cycle Length: 80 secs Phase combination order: #1 #5 #6

FIGURE 2 Webster and Grand input data sheet.

measured average stopped time delay. Approach stopped delay was also evaluated in the tests.

Tests

The tests started with the minimum necessary field data needed to estimate level of service using all of the defaults suggested in Table 9-3 and Equation 9-8 of the HCM (Table 1). Each subsequent test then replaced selected default values with field-observed values, building on the field observations until all field observations had been included in the level-of-service analysis (Figure 7). Table 2 presents the parameters used in each test. The tests proceeded in the following sequence:

- *Test 1: Turning Movement and Basic Geometric Data, plus Observed Signal Timing.* Test 1 included hourly turning movement volumes (flow rates) for each approach plus basic geometric information (lanes and parking location) obtained from the field. Signal phasing sequences and minimum pedestrian times were also measured in the field and included in this test. Signal timing (cycle lengths and phase lengths) was estimated using SOAP84 (*7*). Defaults were used for all other data (grade, heavy vehicles, parking maneuvers, local buses, conflicting pedestrian volumes, arrival type, and PHF). This test was performed for only the two videotaped intersections in Oakland (Harrison and Webster).
- *Test 2: Turning Movements, Geometry, plus Observed Signal Timing.* In Test 2, the optimal signal timing used in Test 1 was replaced with observed phase and cycle lengths. Actual field-observed signal timing data were used for the two videotaped intersections (Harrison and Webster). Observed phase lengths were not reported for the four intersections evaluated by May. Consequently, estimated phase lengths (based on an equal degree of saturation solution, given the observed flow rates, cycle length, and

```
HCS:  Signalized Intersection    Version 2.1                              1
================================================================================
Center For Microcomputers In Transportation
University of Florida
512 Weil Hall
Gainesville, FL   32611-2083          (904) 392-0378
================================================================================
Streets: (E-W) Rose                    (N-S) Grove (MLK)
Analyst: Rgd                           File Name: GROVE7.HC9
Area Type: Other                       6-4-93
Comment: vol+geo+cycle+PHF+delay adj.+actual sat
================================================================================
```

Traffic and Roadway Conditions

	Eastbound			Westbound			Northbound			Southbound		
	L	T	R	L	T	R	L	T	R	L	T	R
No. Lanes		> 1	<		> 1	<		> 1	<		> 1	<
Volumes	30	89	31	39	110	21	18	272	60	23	407	20
PHF or PK15	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Lane Width		12.0			12.0			12.0			12.0	
Grade		0			0			0			0	
% Heavy Veh	5	5	5	1	1	1	4	4	4	1	1	1
Parking	(Y/N)	Y	2	(Y/N)	Y	0	(Y/N)	Y	4	(Y/N)	Y	4
Bus Stops			0			0			3			3
Con. Peds			4			79			51			56
Ped Button	(Y/N)	Y	11.5 s	(Y/N)	Y	11.5 s	(Y/N)	Y	11.5 s	(Y/N)	Y	11.5
Arr Type	3	3	3	2	2	2	5	5	5	3	3	3
RTOR Vols			0			0			0			0

Signal Operations

Phase combination	1	2	3	4	5	6	7	8
EB Left	*				NB Left	*		
Thru	*				Thru	*		
Right	*				Right	*		
Peds	*				Peds	*		
WB Left	*				SB Left	*		
Thru	*				Thru	*		
Right	*				Right	*		
Peds	*				Peds	*		
NB Right					EB Right			
SB Right					WB Right			
Green	16P				Green	43P		
Yellow/A-R	3				Yellow/A-R	3		
Lost Time	3.0				Lost Time	3.0		

Cycle Length: 65 secs Phase combination order: #1 #5

FIGURE 3 Grove and Rose input data sheet.

phase sequence) were used for these four intersections instead of field data.

- *Test 3: Turning Movements, Geometry, Signal Timing, plus PHF.* In Test 3, the default 0.90 peak 15-min factor (PHF) was replaced with the observed PHF. A single intersectionwide PHF was used for the intersections evaluated by May because of the lack of reported data. Approach-specific PHFs were applied to the two videotaped intersections.
- *Test 4: Turning Movements, Geometry, Timing, PHF, plus Arrival Type.* In Test 4 the default Arrival Type 3 was replaced with actual approach arrival types based on field measurements of R_p (platoon ratio) for the two videotaped intersections. The reported arrival types were used for the other four intersections evaluated by May.
- *Test 5: Turning Movements, Geometry, Timing, PHF, Arrival Type, plus Saturation Adjustment Factors.* Field measurements of the percentage of heavy vehicles (F_{hv}), grade (F_g), parking maneuvers (F_p), local buses stopping per hour (F_{bb}), and conflicting pedestrians per hour were used to replace the default values used in the saturation adjustment process. The data set did not permit testing of the lane width factor and area type factor since all lanes were 12 ft wide and all intersections were located outside of central business district areas.
- *Test 6: Turning Movements, Geometry, Timing, PHF, Arrival Type, Saturation Adjustment Factors, plus Ideal Saturation Flow.* The default 1,800 ideal saturation flow rate was replaced with a field-measured 1,900 vehicles per hour green per lane at the two Oakland intersections. This test was not performed on the four intersections reported by May because the necessary data on ideal saturation flows in these areas were lacking.
- *Test 7: Turning Movements, Geometry, Timing, PHF, Arrival Type, Saturation Adjustment Factors, Ideal Saturation Flow, plus Field-Measured Saturation Flows.* Actual saturation flow rates were measured for selected approaches. These were generally the more congested movements at each intersection. The ideal saturation flow entry was modified manually in the HCS2 software for

```
HCS:  Signalized Intersection    Version 2.1                              1
================================================================================
Center For Microcomputers In Transportation
University of Florida
512 Weil Hall
Gainesville, FL   32611-2083          (904) 392-0378
================================================================================
Streets: (E-W) McDonald                (N-S) San Pablo
Analyst: Rgd                           File Name: MCDONAL7.HC9
Area Type: Other                       6-3-93
Comment: vol+geo+cycle+PHF+delay adj.+actual sat
================================================================================
```

Traffic and Roadway Conditions

	Eastbound			Westbound			Northbound			Southbound		
	L	T	R	L	T	R	L	T	R	L	T	R
No. Lanes	1	1	1	1	1 <		1	2	1	1	2	1
Volumes	320	210	260	113	131	56	210	905	95	108	374	78
PHF or PK15	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Lane Width	12.0	12.0	12.0	12.0	12.0		12.0	12.0	12.0	12.0	12.0	12.0
Grade		0			0			0			0	
% Heavy Veh	1	1	1	0	0	0	2	2	2	1	1	1
Parking	(Y/N)	N		(Y/N)	N		(Y/N)	N		(Y/N)	N	
Bus Stops			4			0			0			0
Con. Peds			0			11			13			4
Ped Button	(Y/N)	Y	23.5 s	(Y/N)	Y	23.5 s	(Y/N)	Y	14.5 s	(Y/N)	Y	17.5
Arr Type	3	3	3	3	3	3	3	3	3	3	3	3
RTOR Vols			0			0			0			0

Signal Operations

Phase combination	1	2	3	4		5	6	7	8
EB Left	*				NB Left	*	*		
Thru	*				Thru		*	*	
Right	*				Right		*	*	
Peds	*				Peds		*	*	
WB Left		*			SB Left	*			
Thru		*			Thru			*	
Right		*			Right			*	
Peds		*			Peds			*	
NB Right					EB Right				
SB Right					WB Right				
Green	23A	21A			Green	7A	8A	21A	
Yellow/A-R	3	3			Yellow/A-R	3	3	3	
Lost Time	3.0	3.0			Lost Time	0.0	3.0	3.0	

Cycle Length: 95 secs Phase combination order: #1 #2 #5 #6 #7

FIGURE 4 McDonald and San Pablo input data sheet.

these approaches until the resulting HCM saturation flow calculation resulted in the field-measured saturation flow.

- *Test 8: Field-Measured Average Stopped Delay.* Test 8 merely documents the average stopped delay values measured in the field for each intersection. The field-measured average total delay reported by May was converted back to average stopped delay using the same 1.3 conversion factor used by May to convert stopped delay to total delay. A volume-weighted average of the approach delay reported by May was used to obtain average stopped delay for the entire intersection.

Field measurements of average stopped delay for the two videotaped intersections in Oakland were obtained using the "point sampling" method with 15-sec sampling periods as described by Reilly in his paper on delay estimation techniques (*8*). Average stopped delay was estimated for the peak 15-min volume (flow rate) period for each intersection.

Queues due to vehicles failing to clear the previous cycle were not present at the beginning or ending of the 1-hr sample period. There were short queues of vehicles (fewer than 10 vehicles) on the intersection approaches not receiving a green indication at the start and end of the sample period.

The field-measured delay was assumed to be the true delay for the purposes of this evaluation. Note that Reilly estimated that the point sampling method appears to be a slightly biased estimator of true stopped delay, overestimating delay by about 8 percent (*6*).

ROBUSTNESS OF DATA SET

Every data set is by definition a small sample of the universe of real-world data. No sample can be expected to cover all possible conditions in the field, but by comparing the range of the parameters contained in the data set with the range of values for the parameters in

```
HCS:  Signalized Intersection    Version 2.1                                1
================================================================================
Center For Microcomputers In Transportation
University of Florida
512 Weil Hall
Gainesville, FL   32611-2083          (904) 392-0378
================================================================================
Streets: (E-W) Central                 (N-S) Carlson
Analyst: Rgd                           File Name: CARL7.HC9
Area Type: Other                       6-3-93
Comment: vol+geo+cycle+PHF+Delay Adj+ actual sat
================================================================================

                        Traffic and Roadway Conditions
--------------------------------------------------------------------------------
|          |    Eastbound    |    Westbound    |    Northbound   |    Southbound   |
|          | L    T    R     | L    T    R     | L    T    R     | L    T    R     |
|----------|---- ---- ----    |---- ---- ----    |---- ---- ----    |---- ---- ----    |
|No. Lanes |     > 2  <      |     > 2  <      | 1    2  <       | 1    2  <       |
|Volumes   |  78  522  200   |  16  270   74    | 210  430   40    |  59  164   37    |
|PHF or PK15|0.85 0.85 0.85   |0.85 0.85 0.85    |0.85 0.85 0.85    |0.85 0.85 0.85    |
|Lane Width|     12.0        |      12.0        |12.0 12.0         |12.0 12.0         |
|Grade     |         0       |          0       |         0        |         0        |
|% Heavy Veh|  2    2    2   |   1    1    1    |   1    1    1    |   1    1    1   |
|Parking   | (Y/N) N         | (Y/N) N          | (Y/N) N          | (Y/N) N          |
|Bus Stops |              7  |               0  |              0   |              1   |
|Con. Peds |              0  |               7  |              1   |              5   |
|Ped Button|(Y/N) Y 20.5 s   |(Y/N) Y 20.5 s    |(Y/N) Y 17.5 s    |(Y/N) Y 17.5      |
|Arr Type  |  3    3    3   |   3    3    3    |   4    4    4    |   3    3    3   |
|RTOR Vols |              0  |               0  |              0   |              0   |
--------------------------------------------------------------------------------

                              Signal Operations
--------------------------------------------------------------------------------
Phase combination 1     2     3     4  |           5     6     7     8
EB  Left          *                    |NB  Left    *
    Thru          *                    |    Thru          *
    Right         *                    |    Right         *
    Peds          *                    |    Peds          *
                                       |
WB  Left          *                    |SB  Left    *
    Thru          *                    |    Thru          *
    Right         *                    |    Right         *
    Peds          *                    |    Peds          *
                                       |
NB  Right                              |EB  Right
SB  Right                              |WB  Right
                                       |
Green           36P                    |Green      16P   19P
Yellow/A-R       3                     |Yellow/A-R  3     3
Lost Time      3.0                     |Lost Time  3.0   3.0

Cycle Length:  80 secs  Phase combination order: #1 #5 #6
```

FIGURE 5 Carlson and Central input data sheet.

HCM and in the field, one can gain a sense of how completely the data set does represent the real world.

Lane Geometry

The data set consists of six intersections with 22 approaches. About a quarter of the approaches are single-lane approaches, and the rest are two-lane approaches with and without exclusive turn lanes. All single-lane approaches are opposed by single-lane approaches in this data set. One- and two-lane approaches are well-represented in this sample, but approaches of three lanes and wider are missing. The data set also does not contain an example of a single-lane approach opposed by a multilane approach.

Signal Timing and Left Turn Treatment

The intersections included in this analysis ranged from two-phase fixed-time signals up to six-phase fully actuated signals. Fixed-time signals account for 80 percent of the sample. Actuated signals are represented by only one intersection in this sample.

Cycle lengths range from 65 to 95 sec. Longer cycle lengths that might be typical of wider intersections in suburban locations are not represented in this sample data set.

About two-thirds of the approaches have permitted left-turn phasing. Half of the approaches have exclusive left-turn lanes. The only combination not represented in this data set is protected left turns from a shared turn lane.

Other Characteristics

The maximum average intersection delay observed at the sample intersections was 35 sec (Level of Service D). The highest critical *v/c* (X_c) was 96 percent. Most of the six intersections and 22 approaches in the data set tended to be uncongested (X_c less than 85 percent of capacity); however, the sample data set does include

```
HCS:  Signalized Intersection    Version 2.1                                1
================================================================================
Center For Microcomputers In Transportation
University of Florida
512 Weil Hall
Gainesville, FL   32611-2083          (904) 392-0378
================================================================================
Streets: (E-W) Dwight                  (N-S) Sacramento
Analyst: Rgd                           File Name: DWIGHT7.HC9
Area Type: Other                       6-3-93
Comment: vol+geo+phf+delay Adjust+actual sats
================================================================================
```

Traffic and Roadway Conditions

	Eastbound			Westbound			Northbound			Southbound		
	L	T	R	L	T	R	L	T	R	L	T	R
No. Lanes		> 1	<		> 1	<	1	2	<	1	2	<
Volumes	55	335	50	47	262	81	84	586	50	125	455	40
PHF or PK15	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Lane Width		12.0			12.0		12.0	12.0		12.0	12.0	
Grade		0			0			0			0	
% Heavy Veh	2	2	2	2	2	2	2	2	2	2	2	2
Parking	(Y/N)	Y	20	(Y/N)	Y	20	(Y/N)	Y	20	(Y/N)	Y	20
Bus Stops			0			0			0			0
Con. Peds			50			50			50			50
Ped Button	(Y/N)	Y	20.5 s	(Y/N)	Y	20.5 s	(Y/N)	Y	11.5 s	(Y/N)	Y	11.5
Arr Type	3	3	3	3	3	3	3	3	3	3	3	3
RTOR Vols			0			0			0			0

Signal Operations

Phase combination	1	2	3	4		5	6	7	8
EB Left	*				NB Left	*			
Thru	*				Thru	*			
Right	*				Right	*			
Peds	*				Peds	*			
WB Left	*				SB Left	*			
Thru	*				Thru	*			
Right	*				Right	*			
Peds	*				Peds	*			
NB Right					EB Right				
SB Right					WB Right				
Green	40P				Green	24P			
Yellow/A-R	3				Yellow/A-R	3			
Lost Time	3.0				Lost Time	3.0			

Cycle Length: 70 secs Phase combination order: #1 #5

FIGURE 6 Dwight and Sacramento input data sheet.

TABLE 1 Default Values Used in Level-of-Service Analysis (*1*)

Parameter	Value
Ideal saturation flow	1,800 veh/hr/lane
Conflicting pedestrian flow	Low: 50 pedestrians/hr
Percentage heavy vehicles	2%
PHF	0.90
Grade	0%
Number of stopping buses	0 buses/hr
Number of parking manuevers	20 manuevers/hr
Arrival type	3

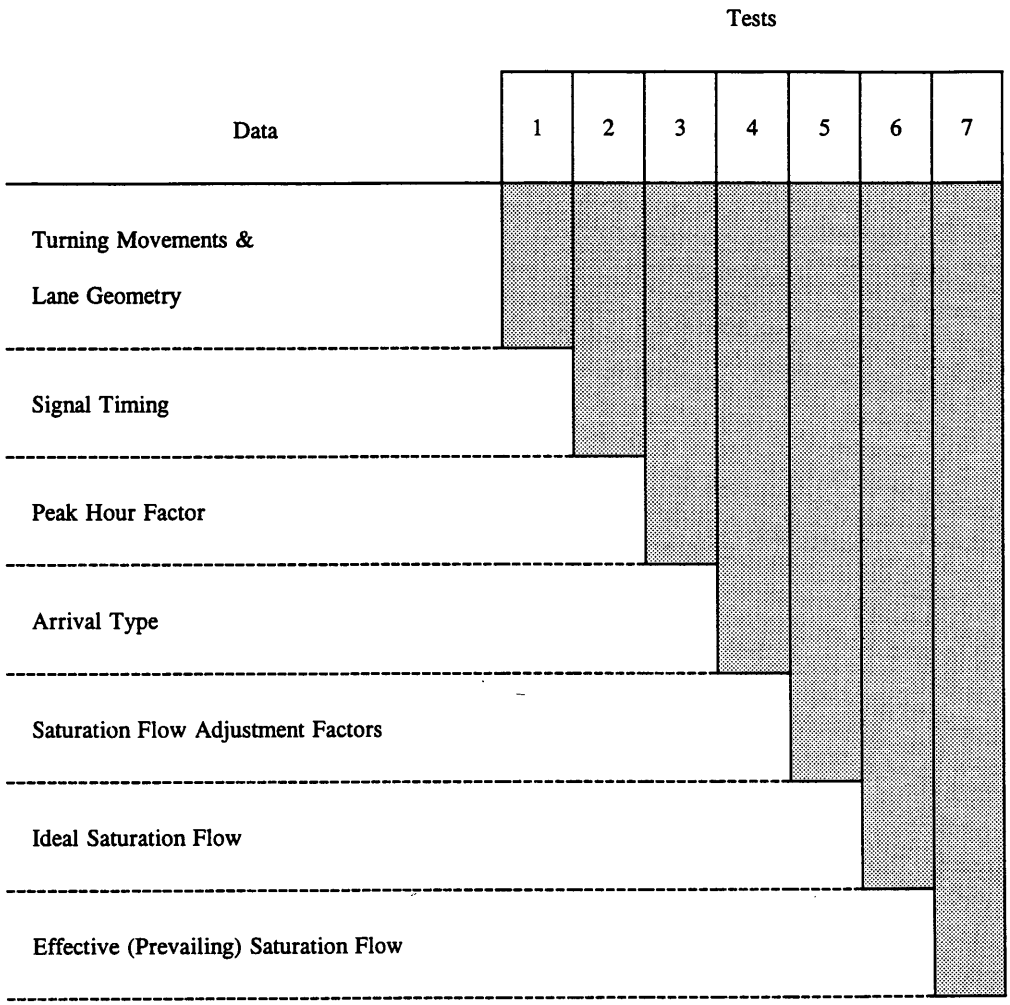


FIGURE 7 Test structure.

4 approaches with *v/c*'s greater than 85 percent of capacity and two intersections (Harrison and McDonald) with X_c's greater than 85 percent.

Variation of Field-Measured Parameters from HCM Defaults

The field-measured parameters in the data set (PHF arrival type, ideal saturation flow, grade, heavy vehicles, parking buses, and pedestrians) did vary from the default values contained in the HCM, but in most cases they fell close to the default values recommended in the HCM (Table 3).

The approach PHFs varied from 0.83 to 1.00 in the field. The maximum range for the PHF is 0.25 to 1.00. The mean PHF observed in the field was 0.87, which is close to the HCM-recommended default of 0.90. Grades were generally flat (under 3 percent) in the data set. The mean observed grade was 0 percent, which is the same as the HCM default.

Heavy vehicles ranged from 0 to 5 percent in the data set. The mean percentage heavy vehicles was 1 percent, which is close to the HCM-recommended default of 2 percent. Parking maneuvers ranged from 0 to 11 per hour—significantly lower than the HCM-recommended default of 20.

Stopping local buses ranged from 0 to 13 per hour. The mean was two buses per hour, which is close to the HCM-recommended default of 0. Pedestrian volumes ranged from 0 to 118 pedestrians per hour; the mean was 32, which is less than the HCM-recommended default of 50.

RESULTS

The results for each test are given in Tables 4 through 7. Table 4 presents the MABS in the estimated average stopped delay per vehicle aggregated for each intersection. Table 5 gives the variation in the estimated critical *v/c* (X_c) for each intersection. Table 6 gives the effect of each test on the estimated intersection level of service. Table 7 presents the estimated average stopped delay per vehicle (in seconds) by approach, sorted by *v/c*.

Results for Entire Intersections

The 1985 HCM operations method for estimating signalized intersection level of service was found to be able to estimate average

TABLE 2 Parameter Values Used in Tests

Parameter Values																							
Parameter	Harrison				Webster			Grove				McDonald				Carlson				Dwight			
	E	W	N	S	W	N	S	E	W	N	S	E	W	N	S	E	W	N	S	E	W	N	S
Test #1 - SOAP Signal Timings																							
Cycle	70"				70"																		
Split - Thrus	20"	20"	50"	39"	21"	38"	38"																
Split - Lefts	-	-	11"	-	-	11"	-																
Test #2 - Actual Signal Timings																							
Cycle	80"				80"			65"				95"				80"				70"			
Split - Thrus	20"	20"	60"	42"	30"	37"	37"	19"	19"	46"	46"	26"	24"	35"	24"	39"	39"	22"	22"	43"	43"	27"	27"
Split - Lefts	-	-	18"	-	-	13"	-	-	-	-	-	-	-	21"	10"	-	-	19"	19"	-	-	-	-
Test #3 - Peak Hour Factor																							
	0.92	1.05	0.94	0.83	0.83	0.97	0.90	0.85															
Test #4 - Arrival Type																							
	3	2	3	4	2	4	4	3	2	5	3	3	3	3	3	3	3	4	3	3	3	3	3
Test #5 - Saturation Adjustment																							
Grade (%)	0	0	0	-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Heavy Veh.(%)	1	1	2	0	0	2	0	5	1	4	1	1	0	2	1	2	1	1	1	2	1	1	1
Parking	0	1	11	None	3	1	4	2	0	4	4	None				None				1	1	1	1
Buses	0	0	7	4	0	4	13	0	0	3	3	4	0	0	0	7	0	0	1	0	0	11	0
Peds	26	71	63	75	118	29	68	4	79	51	56	0	11	13	4	0	7	1	5	9	51	0	0
Test #6 - Ideal Saturation Flow per Lane																							
	1900	1900	1900	1900	1900	1900	1900																
Test #7 - Effective Saturation Flow Per Lane (measured in field)																							
Thru Sat				1912			1731	1286	1525	1104	1475			1626	1602	1386				1494	1457		
Left Sat			972											1446				1636					

TABLE 3 Parameter Ranges

Parameter	HCM Default	Range in Field Data
Pedestrians	50	0–118
Heavy vehicles	2%	0–5%
PHF	0.90	0.83–1.00
Grade	0%	0–3%
Number of stopping buses	0	0–13
Number of parking manuevers	20	0–11
Arrival type	3	2–5

TABLE 4 Average Stopped Delay by Intersection

Average Stopped Delay By Intersection (secs/veh)								
Test	Harrison(1)	Webster	Grove	McDonald	Carlson	Dwight	Mean	Mean Absolute Error
1	22.1	13.1					17.6	7.8
2	31.5	16.5	8.8	26.1	17.1	13.8	19.0	2.7
3	52.5	16.6	9.4	28.5	17.7	15.0	23.3	5.1
4	48.6	16.6	9.3	28.5	16.3	15.0	22.4	4.7
5	48.1	18.3	8.9	28.5	16.3	13.7	22.3	4.6
6	33.6	16.6					25.1	0.6
7	29.1	20.0	8.8	34.6	17.3	13.4	20.5	2.0
Field Measured	33.3	17.5	9.0	35.2	20.7	14.4	21.7	0.0

(1) Northbound, Southbound, and Westbound legs only

intersection stopped delay with a MABS of 3 sec when using all of the default parameters recommended in Table 9-3 of the HCM. Field measurements of the default parameters plus measurements of ideal and effective saturation flows on critical approaches reduced the MABS to 2 sec (Table 4).

Importance of Observing Signal Timing (Tests 1 and 2)

Field observations of signal timing were found to improve significantly the accuracy of the estimated intersection delay. The MABS dropped from 7.8 sec in Test 1 to 2.7 sec in Test 2. A more accurate comparison is obtained if only the two intersections that were included in both tests are considered (Harrison and Webster). In this case, the MABS still improves significantly, dropping from 7.8 to 1.4 sec. The MABS has dropped from approximately 30 percent to only 12 percent of the mean stopped delay at these two intersections.

Lane geometry and turning movements by themselves were not sufficient to estimate the correct level of service at the intersections of Harrison and Webster. The use of SOAP to estimate signal timings resulted in levels of service one level better than actual for both intersections (see Test 1 in Table 6).

Accuracy of Using Only Defaults (Test 2)

The MABS was found to be 2.7 sec (12 percent of the mean delay) when the HCM operations method was applied using turning movements, lane geometry, and signal timing as the only field-collected data. This accuracy was sufficient to give the correct level of service at all intersections (see Test 2 results in Table 6).

Effect of PHF and Arrival Type (Tests 3 and 4)

The addition of field measurements of the PHF and arrival type generally did not improve the average delay estimates at most intersections. The MABS actually increased to about 5 sec for both tests (Table 4).

TABLE 5 Critical *X v/c* as Estimated by HCM Method

Critical "X" Volume Capacity Ratio (As Estimated by HCM Method)							
	Test	Harrison(1)	Webster	Grove	McDonald	Carlson	Dwight
1	Volumes + Geometry + SOAP Timings	0.900	0.701				
2	Volumes + Geometry + Signal Timing	0.889	0.691	0.597	0.802	0.660	0.721
3	Volumes + Geometry + Signal Timing + PHF	0.910	0.702	0.636	0.848	0.703	0.770
4	Volumes + Geometry + Signal Timing + PHF + Rp	0.910	0.702	0.636	0.848	0.703	0.770
5	Vols + Geo + Signal + PHF + Rp + Grade + HV + Pkg + Bus + Peds	0.988	0.723	0.577	0.844	0.706	0.703
6	Test 5 plus Ideal Saturation Flow	0.936	0.685				
7	Test 6 with Actual Saturation selected moves	0.907	0.729	0.562	0.879	0.757	0.663

(1) For Northbound, Southbound, Westbound legs only

TABLE 6 Intersection Level of Service

Intersection Level of Service							
	Test	Harrison(1)	Webster	Grove	McDonald	Carlson	Dwight
1	Volumes + geometry + SOAP Timing	C	B				
2	Volumes + Geometry + Signal Timing	D	C	B	D	C	B
3	Volumes + Geometry + Signal Timing + PHF	E	C	B	D	C	B
4	Volumes + Geometry + Signal Timing + PHF + Rp	E	C	B	D	C	B
5	Vols + Geo + Signal + PHF + Rp + Grade + HV + Pkg + Bus + Peds	E	C	B	D	C	B
6	Test 5 with ideal saturation flows	D	C				
7	Test 6 with Actual Saturation selected moves	D	C	B	D	C	B
8	Field Measured Delay	D	C	B	D	C	B

(1) For Northbound, Southbound, Westbound legs only

TABLE 7 Average Stopped Delay per Vehicle by Approach

Average Stopped Delay Per Vehicle by Approach										
Intersection And Approach		v/c	Field Measured Delay	Test						
				1	2	3	4	5	6	7
Carlson	SB	0.30	20.0		19.3	19.4	19.4	19.4		19.4
Carlson	WB	0.36	13.0		10.9	11.1	11.1	11.1		11.1
Webster	WB	0.51	25.9	19.7	16.7	17.1	23.2	22.6	22.2	22.3
Grove	WB	0.54	15.0		19.0	20.1	24.5	24.5		23.4
Grove	SB	0.55	3.0		5.5	5.9	5.9	5.2		4.9
Grove	EB	0.56	36.0		17.7	18.4	18.4	17.4		17.9
Grove	NB	0.56	2.0		4.5	4.7	2.5	2.4		2.8
Dwight	WB	0.57	7.0		10.0	11.3	11.3	9.2		8.0
McDonald	WB	0.58	35.0		23.0	23.3	23.3	23.2		23.2
Webster	NB	0.59	12.1	10.7	12.2	11.8	9.2	9.2	8.3	8.3
Dwight	EB	0.62	8.0		11.5	13.6	13.6	10.4		8.6
Harrison	WB	0.68	24.8	21.5	31.9	26.8	32.7	30.6	29.4	29.4
Dwight	SB	0.69	30.0		15.2	16.1	16.1	15.7		16.1
McDonald	SB	0.71	25.0		24.5	25.1	25.1	25.0		25.5
Dwight	NB	0.73	22.0		16.0	16.7	16.7	16.1		16.7
Carlson	NB	0.76	19.0		23.5	24.5	20.1	20.0		20.4
Carlson	EB	0.79	26.0		13.8	14.4	14.4	14.6		16.8
Harrison	NB	0.83	28.2	17.2	10.8	10.7	10.7	28.6	23	14.7
McDonald	EB	0.91	50.0		30.5	34.0	34.0	34.0		34.0
Webster	SB	0.94	15.8	11.2	18.4	18.4	16.6	20.1	17.5	24.0
McDonald	NB	1.02	30.0		24.7	27.9	27.9	27.9		41.8
Harrison	SB	1.03	37.2	24.2	39.9	75.2	67.6	59.9	38.8	35.0
Mean		0.67	22.05	17.42	18.16	20.30	20.20	20.32	23.20	19.29
Mean Absolute Error			0.0	6.6	6.8	8.0	7.6	6.4	3.4	6.3

The delay estimate deteriorated significantly for one seriously congested intersection (Harrison). The estimation error was greatly increased at Harrison because opposite errors generated by the use of default parameters no longer canceled each other out. Harrison has one saturated approach (southbound) operating at a *v/c* of 1.00 that is extremely sensitive to the estimated saturation flow. The addition of field measurements of the PHF (without saturation flow measurements) caused the HCM method in this case to underestimate significantly the true capacity of this approach, thus causing the increased error.

These results indicate that the additional refinements provided by PHF and arrival type are not warranted in the absence of accurate field data on saturation flows. In fact, the data may worsen the result.

Effect of Measuring Some But Not All Saturation Flow Estimation Parameters (Test 5)

Precise field measurements of the percentage heavy vehicles, parking maneuvers, local buses, and pedestrians did not much improve

the accuracy of the estimated average intersection delay. The MABS remained relatively unchanged between Tests 4 and 5 (Table 4). This conclusion is still true even for intersections with high *v/c*'s (such as Harrison and McDonald, where the *v/c*'s exceed 85 percent of capacity) as well as for those with low *v/c*'s (Figure 8).

Value of Measuring Ideal Saturation Flow in the Field (Test 6)

Test 6 was performed only at the two videotaped intersections, Harrison and Webster, because of the lack of the necessary data for the other intersections.

Field measurements of ideal saturation flow were found to contribute significantly to the accuracy of the HCM method at Harrison. The estimated average delay was within 1 sec of field observations. This intersection has one critical approach that operates at a *v/c* of 1.00 during the peak 15 min. Field measurements of ideal saturation flow did not make any significant contribution to the accuracy of the delay estimate at Webster because of the lack of congestion at this location.

Measuring Effective Saturation Flow in the Field (Test 7)

Field measurements of ideal and effective saturation flow greatly improved the accuracy of the delay estimates for the two intersections with *v/c*'s of more than 85 percent (compare Tests 5 and 7 in Figure 8). The difficulty of accurately measuring and applying effective or prevailing saturation flows in the HCM method, however, made the results less satisfactory than simply measuring ideal saturation flow (compare Tests 6 and 7 in Figure 8). Field observations of saturation flow made no significant improvement in the accuracy of the delay estimates at less congested intersections. The MABS for all intersections were cut in half (from 4.6 under Test 5 to 2.0 under Test 7).

It is much more difficult to measure accurately effective or prevailing saturation flow since many more observations are required over a longer period to adequately represent the average conditions affecting saturation flow over an entire hour. Effective saturation flow measurements consequently tend to be less accurate than ideal saturation flow measurements. This effect is demonstrated in the worsening of the delay estimates for a couple of key approaches when effective saturation flow rate is used rather than ideal saturation flow rate. The MABS, however, remains among the best of the tests.

Results for Individual Intersection Approaches

The previous results are examined briefly at the more detailed approach level to determine if aggregating approach delay to intersection delay may have masked some of the effects of the tests. The results by approach are given in Table 7.

The MABS by approach is generally higher but shows less variation among the tests than for the individual intersections. The error varied relatively little for Tests 1, 2, 3, and 4 (between 7 and 8 sec). Maximum data collection (PHF, arrival type, saturation flow parameters, ideal saturation flow, and prevailing saturation flows) reduced the MABS to 6 sec (about 33 percent of the mean approach delay) (see Test 7 in Table 7).

Test 6 (using field-measured ideal saturation flows) showed a significantly improved MABS of 4 sec, but this was for a smaller sample size of 6 of the total 22 approaches. The results did vary significantly among approaches with high and low *v/c*'s (Figure 9).

The four approaches where *v/c*'s exceeded 85 percent of capacity were extremely sensitive to the lack of field data for all of the parameters used to estimate saturation flow and delay. Tests using partial field data for the parameters (Tests 3, 4, and 5) provided worse delay estimates than Test 2, which relied on only turning movment, geometry, and signal timing data collected in the field.

The remaining 18 approaches, for which *v/c*'s were less than 85 percent, were relatively insensitive to field-collected data on PHFs, arrival type, and saturation flow parameters.

Again, the difficulty of measuring effective saturation flows in the field caused the results of Test 7 to be less satisfactory than those of Test 6.

CONCLUSIONS

The test results for the six intersections suggest the following:

1. The estimate of intersection average stopped delay is insensitive to precise estimates of PHF, arrival type, and saturation flow if the *v/c*'s on all the approaches are less than 85 percent of capacity. There is no value to gathering additional field data beyond traffic counts, lane geometry, and signal timing in order to estimate

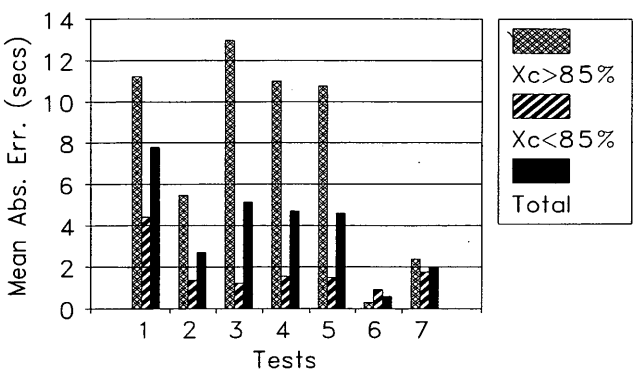


FIGURE 8 Intersection delay results.

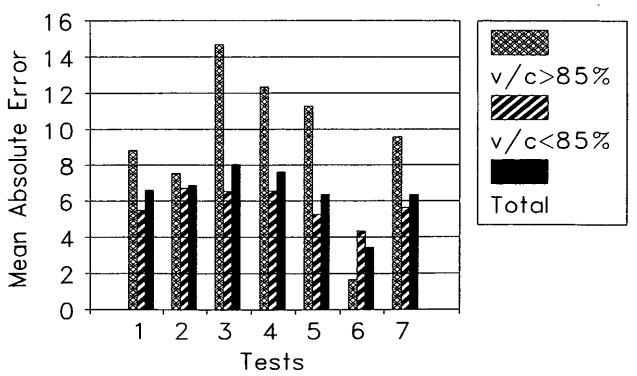


FIGURE 9 Approach delay results.

accurately the intersection level of service when there is little congestion.

2. There is some value to gathering additional field data to more precisely determine the intersection and approach delay if one or more of the approaches has a *v/c* in excess of 85 percent of capacity. However, the analyst must make a full data collection effort covering PHF, arrival type, saturation flow adjustment factors, and ideal saturation flow. Partial data collection efforts that measure only some of these parameters may result in less accurate results.

3. The basic data collection effort of counts, lanes, and signal timing should yield the correct letter grade level of service for most all situations. The MABS will be about 7 sec for the approaches and about 3 to 5 sec for the intersection as a whole.

4. The precision with which the saturation flow adjustment parameters are presented in the HCM may not be warranted in terms of their effect on the estimate of intersection delay. For example, the grade saturation flow adjustment factor might be reported for "low," "medium," and "high" grades rather than by specific grade percentage.

These conclusions apply for situations in which the PHF, arrival type, and ideal saturation flow do not vary significantly (±10 percent) from the default parameter values in the HCM; this was the approximate range of the data set analyzed in this paper. Extrapolations of the results to more extreme situations would require a more extensive data set.

Other investigators (Lin and Teodorvic) have measured higher MABS for approach delay (9 to 18 sec) for data sets where the ideal saturation flow rate is significantly higher (Teodorvic) than the HCM default or the signal control type is predominantly traffic-actuated (Lin). In these and other situations in which conditions vary significantly from the default conditions, additional field data should be collected to determine the intersection level of service with accuracy.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. Reilly, J. R., J. H. Kell, M. L. Gallagher, R. P. Pfeffer, and A. Sorton. *Urban Signalized Intersection Capacity.* NCHRP 3-28[2] Report, TRB, National Research Council, Washington, D.C., 1982.
3. May, A. D., E. Geduzlioglu, and L. Tai. Comparative Analysis of Signalized Intersection Capacity Methods. Presented at 62nd Annual Meeting of the Transportation Research Board, Washington, D.C., 1983.
4. Teodorvic, D., S. Kikuchi, P. Chakroborty, and V. Perincherry. *Analysis of Delay and Level of Service at Signalized Intersections in Delaware.* Report 90-DTC-4. Delaware Transportation Center, University of Delaware, Dover, 1990.
5. Lin, F.-B. Applications of 1985 Highway Capacity Manual for Estimating Delays at Signalized Intersections. Presented at 68th Annual Meeting of the Transportation Research Board, Washington, D.C., 1989.
6. *Highway Capacity Manual Software*, Version 2.1. McTrans Center for Microcomputers in Transportation, University of Florida, Gainesville, 1990.
7. *SOAP84.* FHWA, U.S. Department of Transportation; McTrans Center for Microcomputers in Transportation, University of Florida, Gainesville, Jan. 1985.
8. Reilly, W. R., C. C. Gardner, and J. H. Kell. *A Technique for Measurement of Delay at Intersections.* Report FHWA-RD-135, 136, 137. FHWA, U.S. Department of Transportation, 1976.

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

Permitted Left-Turn Capacity of Exclusive Lanes: Simulation-Based Empirical Method

GANG-LEN CHANG, LEIMIN ZHUANG, AND CESAR PEREZ

An exploratory procedure for analyzing the permitted left-turn capacity with exclusive lanes is presented, including several empirical models for opposing queue length prediction, permitted saturation flow of mixed traffic, and effect of bay length on the left-turn capacity. Some critical factors, such as the number of opposing lanes and the interactions between upstream and downstream green time–cycle length ratios, have been incorporated in the proposed procedures for capacity estimation. A discrete choice modeling methodology has been applied to predict the fraction of time in a cycle during which the through queue length may be over a certain distance. Such a model enables traffic engineers to determine the left-turn bay length from a cost-benefit perspective. It should be noted that all proposed empirical models are grounded on the simulation experiments with TRAF-NETSIM. Hence, adjustments or modifications may be necessary after extensive field observations are conducted to calibrate TRAF-NETSIM.

The presence of left-turning vehicles at signalized intersections tends to cause excessive delay, increase accident potential, and lower intersection capacity. Hence, accommodating left-turning vehicles with effective signal control strategies has long been a source of concern for traffic engineers. In practice, depending on the use of shared or exclusive lanes for left-turning vehicles, traffic engineers must select a left-turn phasing that best satisfies the left-turn demand and minimizes the operational difficulties incurred by left turns. An appropriate tool or procedure to evaluate the proposed design strategies (i.e., permitted, protected, protected/permitted) thus becomes essential.

Over the past several decades, although highway agencies and research institutions have developed various guidelines for analyzing left-turn capacity, the most widely used are the procedures included in Chapter 9 of the 1985 *Highway Capacity Manual* (HCM). In fact, the 1985 HCM has been used by more traffic and transportation engineers in the past 7 years since it was published than the 1965 HCM was in 20 years.

However, because of both the limited resources and the lack of sufficient empirical validation in their developments, many procedures or models recommended by the 1985 HCM are subject to revision. This is particularly true of Chapter 9, "Signalized Intersections." In many situations, the output from an analysis of left-turn capacity either does not agree with field observations or yields vastly different results.

In view of various technical deficiencies identified with given applications for using the HCM signalized intersection methodology, a number of attempts have been made to modify or enhance the current procedures. This is one of several such projects sponsored by FHWA, with an emphasis on the operational analysis of exclusive left-turn lanes.

G.-L. Chang and L. Zhuang, Department of Civil Engineering, University of Maryland, College Park, Md. 20742. C. Perez, Federal Highway Administration, 6300 Georgetown Pike, HSR-10, McLean, Va. 22101.

LITERATURE REVIEW

Most existing methods for left-turn analysis start with the estimation of left-turn saturation flow rate. The capacity under various conditions can thus be obtained with appropriate adjustments of the effective green time, cycle length, and other related factors. Prominent studies in this area include the Illinois method (*1*), the revised HCM draft (*2,3*), the Canadian methods (*4*), the U.K. method (*5*), the Swedish approaches (*6,7*), and Australian Road Research Board procedures (*8*). Despite the increasing attention on improving the accuracy for left-turn analysis, existing methods still face some of the following critical issues:

1. The trade of theoretical rigorousness with analytical tractability, such as using simplified assumptions or ignoring some vital elements, in deriving a convenient closed-form solution;
2. The representation of complex population data with limited field observations, such as fitting an empirical model from selected location data without reliable parameter stability analyses; and
3. The demand for very extensive field data, such as directly applying a simulation program for capacity estimation. A detailed review of these methods or procedures has been conducted by a research team at the University of Maryland, and is available elsewhere (*9*).

One of the promising ways to circumvent the aforementioned difficulties is to develop empirical models from a well-calibrated simulation model. Conceivably, such models may not be so appealing as analytical formulations in terms of their mathematical elegance, but they can realistically incorporate related critical factors and their complex interactions though the results of simulation experiments. The stochastic nature of a traffic system as well as the impact of driver behavior on the resulting capacity can also be explored with a proper design of simulation experiments. Hence, this study, as recommended by FHWA, intends to take full advantage of TRAF-NETSIM in the development of operationally convenient yet theoretically reliable models for estimating the capacity of exclusive left-turn lanes.

FRAMEWORK FOR EXCLUSIVE LEFT-TURN CAPACITY ESTIMATION UNDER PERMITTED PHASING

As indicated, the proposed method for estimating left-turn capacity intends to maximize the use of traffic simulation models so that the complex interactions among driver behavior, geometric conditions, and signal control strategies can be fully considered. The simulation-based analyses also allow for assessment of various

input data quality on the capacity estimation. To facilitate the illustration, the entire process for analyzing the exclusive left-turn capacity (ELTC) under permitted phasing is divided into the following principal steps (Figure 1), including both the empirical models and the computation procedures.

Before each step is described in detail, it should be noted that all employed regression models have been through rigorous evaluation, including the following tests for their required properties:

- The residual of any proposed regression model is a random variable.
- The mean value of the residuals in any particular period is 0.
- The variance of the residuals is constant in each period.
- The residuals follow a normal distribution of 0 mean and constant variance.
- All residuals are independent.
- The model residuals are independent of any explanatory variables.
- The model explanatory variables are not perfectly linearly correlated.
- The macrovariables are correctly aggregated.
- All model parameters are independent of the selected sample size (i.e., stability).

An in-depth discussion of these tests is not within the scope of this paper but is available in most econometrics books. Definitions of all variables used in the following analyses are given in Table 1. Note that all empirical equations presented hereafter are based on the simulation experiments of uncoordinated, pretimed intersections with no queue spillback to the upstream intersection.

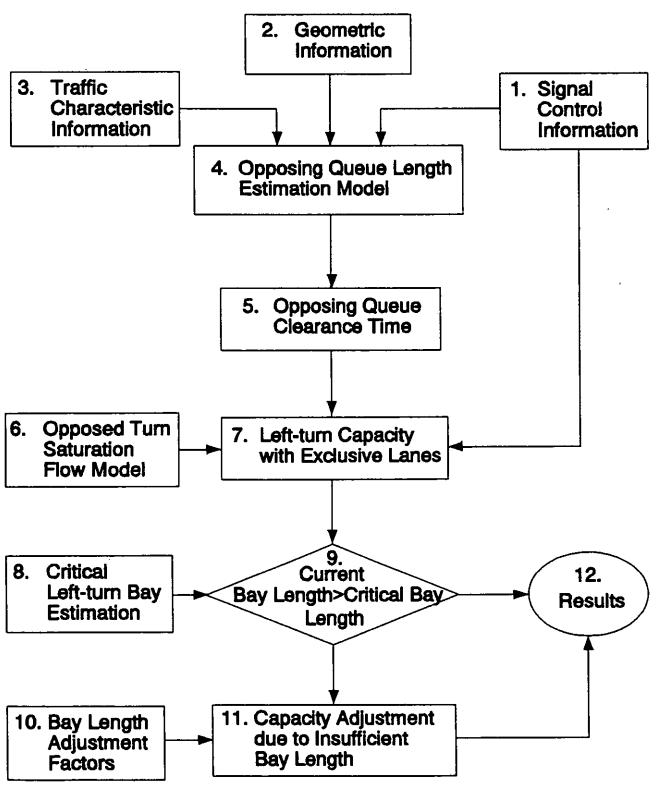


FIGURE 1 Permitted left-turn capacity estimation process (exclusive lane).

Steps 1–3: Preparation of Input Data

As required in the HCM, the first three steps are designed to provide all necessary information for capacity estimation, including signal control plans, geometric conditions, traffic volume, and flow characteristics.

Step 4: Opposing Queue Length Estimation

Since all left-turn vehicles under permitted phasing will be blocked by the opposing discharging vehicles, it is essential to have an accurate estimation of the queue length under the given environment. The available portion of green time can thus be computed according to the observed queue discharging headway. Conceivably, the maximum opposing queue length varies with the arrival traffic patterns, discharge rate, and signal control strategies at both the upstream and the target intersections. A realistic representation of their interactions with analytical formulations would be too complex for use in practice. Hence, the following hybrid model, which is based on extensive simulation experiments, is proposed for this study:

$$N_q = \left[X_5 \cdot (1 - X_3) \cdot \left(\frac{X_1}{X_4 \cdot 3{,}600 \cdot X_2}\right)\right]^{0.6755} \cdot \left(\frac{X_2}{X_3}\right)^{0.7951} \cdot X_1^{0.2235} \cdot X_3^{0.4044} \cdot X_5^{-0.2597} \qquad R^2 = .94, N = 352 \tag{1}$$

where

N_q = number of queue vehicles to be discharged at beginning of green phase;
X_1 = total opposing flow rate per hour (vph), 200 vph $\leq X_1 \leq$ 2,700 vph;
X_2 = green time–cycle length ratio (G/C) for through movement at upstream intersection, $0.3 \leq X_2 \leq 0.8$;
X_3 = G/C for through movement at target intersection i, $0.3 \leq X_2 \leq 0.8$;
X_4 = number of opposing through lanes to discharge queue vehicles, $1 \leq X_4 \leq 3$; and
X_5 = cycle length of target intersection, 60 sec $\leq X_5 \leq$ 120 sec.

Note that the first term approximates the platoons entering the link during the upstream green phase and arriving at the target intersection during the red phase. The effects of G/C at both upstream and downstream intersections and the cycle length on the traffic patterns are then incorporated in the multiplicative adjustment terms.

Step 5: Computation of Opposing Queue Clearance Time

Given the estimated queue length, N_q, from Equation 1, the total opposing queue discharging time can thus be computed by

$$g_0 = N_q \times \overline{H} \tag{2}$$

where $\overline{H}$ is the average queue discharging headway obtained from either field observations or a default empirical value. The unsaturated portion of a green phase for permitted left turns is thus given by

$$g_e = g - g_0 - t_L - a_m$$

TABLE 1 Definition of Model Variables

Variable	Definition
X_1	The total opposing flow rate per hour (vph)
X_2	The G/C ratio at the upstream intersection
X_3	The G/C ratio at the target intersection
X_4	The number of opposing lanes to discharge queue vehicles
X_5	The cycle length of the target intersection
X_6	The total flow rate for the through movement
X_7	The number of lanes for the through movement

where

g = allocated green time,
t_L = loss time, and
a_m = yellow time.

Step 6: Estimation of Permitted Left-Turn Saturation Flow Rate

The primary purpose of Step 6 is to estimate the maximum left-turn flow rate during the effective green phase that has unsaturated opposing flows. Conceivably, factors associated with the maximum permitted left turns include the opposing flow rate, number of opposing lanes, and upstream *G/C* that captures, to some extent, the arrival patterns. To take advantage of TRAF-NETSIM's capabilities, the authors have conducted extensive simulation experiments and have produced the following model for estimating the saturation flow of permitted left turns:

$$S_{pm} = 1{,}723.47 + 0.00017(X_1^*)^2 - 1.0627X_1^* - 300.45X_4 \qquad R^2 = .90, N = 547 \tag{3}$$

where $S_{pm} \geq 0$ and X_1^* is the effective opposing volume to left-turning vehicles, rather than the average opposing flow, and is defined as follows:

$$X_1^* = [Z - Q] \cdot [3{,}600/g_e]$$
$$Z = X_1/[3{,}600/C],\ Q = \alpha \cdot N_4 \tag{3a}$$

where

Z = average opposing vehicles per cycle,
Q = total number of queue vehicles per cycle that exhibit some relation with maximum queue length (N_q), and
α = parameter to capture interrelation between average and maximum queue length per cycle.

The key notion underlying Equation 3a is that after clearing the initial queue on each lane, $(Z - Q)$ vehicles per cycle arrive at the intersection and thus block the left-turning vehicles; if the intersection is not oversaturated, all opposing $(Z - Q)$ vehicles will be discharged during the effective green period. Hence, the actual average gap available for left-turning vehicles is $g_e/(Z - Q)$, and the equivalent opposing flow conflicted with left-turning vehicles under such a condition is X_1^* rather than X_1.

Note that this specification, selected from nine possible function forms, captures the nonlinear relation between the opposing flows and the allowable left-turning vehicles; all parameters are statistically significant at the 0.001 level. Such a specification satisfies all assumptions not only for multivariate regression but also for the stability test (i.e., the estimated results are independent of the selected sample size). Hence, even though TRAF-NETSIM may need to update its key parameters from field observations, the exploratory analysis results remain promising.

Step 7: Computation of Left-Turn Capacity Under Permitted Phasing

Given the opposing queue discharging time and saturation flow rate from Steps 5 and 6, the left-turn capacity under permitted phasing is given by

$$\mathrm{CA}_{pm}(\text{capacity}) = \left(\frac{3{,}600}{C}\right)\left[(g_e) \cdot \left(\frac{S_{pm}}{3{,}600}\right) + N_f\right] = S_{pm}\left(\frac{g_e}{C}\right) + \left(\frac{3{,}600}{C}\right) \cdot N_f \tag{4}$$

where g_e denotes the effective green time for permitted left-turns (i.e., after discharging the opposing queue), or the unsaturated portion of the green phase for opposing vehicles, and N_f is the number of sneakers per cycle.

Note that the aforementioned procedures apply only for estimating the left-turn capacity with an exclusive lane and under non-coordinated signals. Additional adjustments will be necessary if a left-turn bay, instead of lanes, is used. As such, the authors have proposed the following five steps to account for the impact of bay length on the available left-turn capacity.

Step 8: Estimation of Required Bay Length for Permitted Left-Turn

The primary purpose of Step 8 is to ensure that the available capacity for permitted left-turns can be achieved with the given bay length. Hence, it should be considered from both the "demand" and "supply" sides. To some extent, the available permitted capacity, based on the opposing traffic conditions, can be viewed as the supply-side maximum permitted left-turn flows. The maximum allowable arriving vehicles for left turns, on the other hand, function like the demand-side flows. With a simple deterministic analysis, the approximate left-turn bay length under permitted phasing can be computed with the following variables:

CA_{pm} = permitted left-turn capacity with an exclusive lane,
S_T = saturation flow rate for through lane,
Q_T = arriving flow rate for through vehicles in left-turn lane,
Q_L = arriving flow rate for left-turn vehicles, and
$\bar{L}$ = average occupied space per vehicle.

The relations between S_T, Q_T, and Q_L are illustrated in Figure 2.

Basically, the required bay length from the capacity perspective is given by

$$L_{pm} = \max(L_S, L_D) \tag{5}$$

where

$$L_S = \bar{L} \cdot \left[CA_{pm} \cdot \left(\frac{C}{3{,}600} \right) \right] \tag{6}$$

$$L_D = \bar{L} \cdot \left[\frac{Q_L}{3{,}600} \cdot C \right] \quad \text{if } Q_L < S_T - Q_T \tag{7}$$

$$L_D = \bar{L} \left[\left(\frac{S_T - Q_T}{3{,}600} \right) \cdot C \right] \quad \text{if } Q_L > S_T - Q_T \tag{8}$$

Note that Equation 5 represents the average queue length necessary for taking advantage of available left-turn gaps, where C denotes the cycle time. The left-turn capacity may not be fully used if the queue length during the green phase is less than the minimum required size. On the other hand, the bay length may be unnecessarily long if the approach has only very few left-turn vehicles that cannot fully use the available bay length. Hence, Equation 7 represents the required bay length based on the arriving left-turn vehicles per cycle, if the through lanes have enough capacity to accommodate the arriving left-turn vehicles. However, under some conditions, not all left-turn vehicles can merge to the left-turn bay because of the larger number of through vehicles. The required bay length is thus given by Equation 8.

As illustrated in Figure 2, the required bay length to use all of the available left-turn capacity depends not only on its own supply and demand levels, but also on the through flow rate. A left-turn vehicle may be blocked by the through queue vehicles and thus miss the turning opportunity. Hence, the left-turn bay must also be longer than the maximum through queue length per cycle. More specifically, considering the impact of the through flow rate, Equation 5 should be restated as follows:

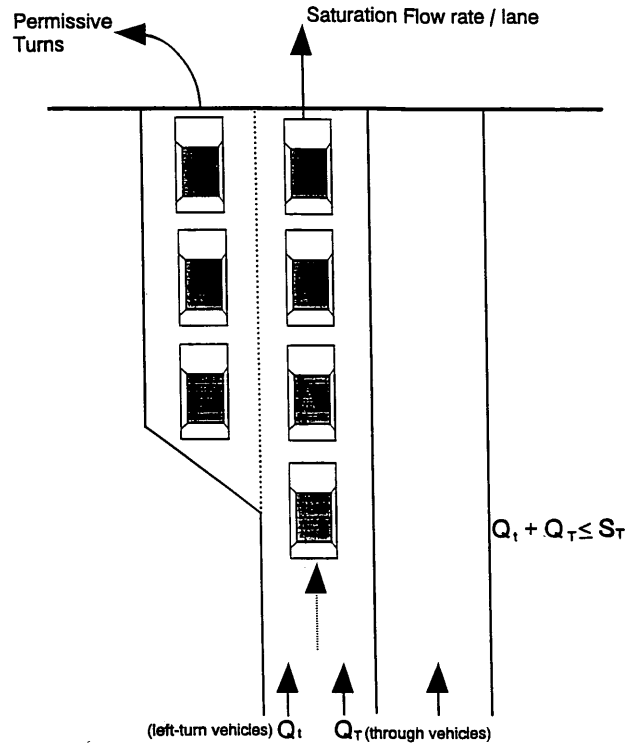


FIGURE 2 Graphical representation of interrelations among bay length and through and left-turn vehicles.

$$L_{pm} = \max(L_S, L_D, L_T) \tag{9}$$

where L_T is the maximum queue length of through vehicles per cycle over 1 hr. In most cases, Equation 1 can be used to predict the maximum queue length as long as the network is not oversaturated.

Step 9: Comparison Between Actual and Required Bay Length

In principle, the left-turn bay can be viewed as a left-turn lane if it is longer than the required length (i.e., $L_A > L_{pm}$). Otherwise some adjustments will be necessary, as the actual usable capacity will be less than the capacity estimated on the basis of opposing traffic conditions. A discrete model to generate the approximate adjustment factor is thus proposed in the next step.

Step 10: Bay Length Adjustment Factors

The purpose of Step 10 is to estimate the fraction of green time in a cycle during which the through queue length is so long that it blocks the left-turn vehicles from entering the left-turn bay. One can then adjust the available capacity on the basis of total blocked duration. With extensive simulation experiments, the authors have developed a discrete choice model for prediction of such a blocked period:

$P(\lambda \geq 120 \text{ ft})$

$$= \frac{\exp\left(a \cdot \frac{X_6}{X_7} + b \cdot X_2 + c \cdot X_3 + d \cdot X_5 + e \cdot \frac{X_2}{X_3}\right)}{\exp(f) + \exp\left(a \cdot \frac{X_6}{X_7} + b \cdot X_2 + c \cdot X_3 + d \cdot X_5 + e \cdot \frac{X_2}{X_3}\right)} \quad (10)$$

where

$a = 0.0011$ ($t = 3.1$),
$b = 2.6464$ ($t = 2.3$),
$c = -1.3985$ ($t = 1.1$),
$d = -0.0005$ ($t = 1.9$),
$e = -0.0870$ ($t = 1.0$),
$f = 2.7285$ ($t = 0.32$), and
$\rho^2 = 0.44$

$P(\lambda \geq 160 \text{ ft})$

$$= \frac{\exp\left(a \cdot \frac{X_6}{X_7} + b \cdot X_2 + c \cdot X_3 + d \cdot X_5 + e \cdot \frac{X_2}{X_3}\right)}{\exp(f) + \exp\left(a \cdot \frac{X_6}{X_7} + b \cdot X_2 + c \cdot X_3 + d \cdot X_5 + e \cdot \frac{X_2}{X_3}\right)} \quad (11)$$

where

$a = 0.0028$ ($t = 5.0$),
$b = 1.1475$ ($t = 1.8$),
$c = -3.6854$ ($t = 2.0$),
$d = 0.0119$ ($t = 3.2$),
$e = 0.2550$ ($t = 6.1$),
$f = 3.7108$ ($t = 3.4$), and
$\rho^2 = 0.63$.

$P(\lambda \geq 200 \text{ ft})$

$$= \frac{\exp\left(a \cdot \frac{X_6}{X_7} + b \cdot X_2 + c \cdot X_3 + d \cdot X_5 + e \cdot \frac{X_2}{X_3}\right)}{\exp(f) + \exp\left(a \cdot \frac{X_6}{X_7} + b \cdot X_2 + c \cdot X_3 + d \cdot X_5 + e \cdot \frac{X_2}{X_3}\right)} \quad (12)$$

where

$a = 0.0028$ ($t = 81.5$),
$b = 1.5931$ ($t = 289$),
$c = -4.4462$ ($t = 276$),
$d = 0.0072$ ($t = 1{,}200$),
$e = 1.3047$ ($t = 361$),
$f = 5.2287$ ($t = 27.8$), and
$\rho^2 = 0.97$

where

$P(\lambda \geq \alpha \text{ ft})$ = fraction of time in a given cycle during which through queue length is longer than α ft,
X_6 = total flow rate for through movement,
X_7 = total number of lanes for through movement, and
ρ = goodness-of-fit indicator for discrete models.

With these functions, one can predict the fraction of time during which the queue exceeds a certain distance.

Step 11: Capacity Adjustment

Given an insufficient bay length, L^*, its permitted left-turn capacity can thus be computed according to the following expression:

$$CA_{pm}(L^*) = CA_{pm} \times [1 - P(\lambda \geq L^*)] \quad (13)$$

where

$CA_{pm}(L^*)$ = left-turn capacity under permitted phasing and bay length of L^* ft;
CA_{pm} = same capacity with a full left-turn lane;
$P(\lambda \geq L^*)$ = total fraction of time in a cycle during which through queue length exceeds given left-turn bay.

NUMERICAL EXAMPLES

To evaluate the performance of the proposed models and procedures, the following four test scenarios have been designed:

- Scenario A:
  - –Number of opposing lanes = 1,
  - –Cycle length = 100 sec,
  - –$G/C = 0.5$,
  - –G/C^* at the upstream intersection = 0.5,
  - –Opposing volume: from 100 to 700 vph (seven cases).
- Scenario B:
  - –Number of opposing lanes = 2,
  - –Cycle length = 100 sec,
  - –$G/C = 0.5$,
  - –$G/C^* = 0.5$, and
  - –Opposing volume: from 100 to 1,000 vph (10 cases).
- Scenario C:
  - –Number of opposing lanes = 3,
  - –Cycle length = 100 sec,
  - –$G/C = 0.5$,
  - –$G/C^* = 0.5$, and
  - –Opposing volume: from 100 to 1000 vph (10 cases).
- Scenario D:
  - –Number of opposing lanes = 4,
  - –Cycle length = 100 sec,
  - –$G/C = 0.5$,
  - –$G/C^* = 0.5$, and
  - –Opposing volume: from 100 to 1000 vph (10 cases).

Since the G/C at both up- and downstream intersections may contribute to the variation of traffic patterns, the authors have also investigated additional 30 cases of similar scenarios but different G/C's.

Field data collection is not the focus of research at this stage, so it is assumed that TRAF-NETSIM is capable of yielding a reasonably reliable capacity estimation, and thus its results are used as the reference base for evaluation.

With such a criterion, the proposed model, as shown in Figures 3 through 6, outperforms the HCM approach in 7 out of 7 cases in Scenario A, 7 out of 10 cases in Scenario B, 7 out of 10 cases in Scenario C, and 8 out of 10 cases in Scenario D. Of the 69 cases overall, the proposed model yielded results better than the HCM in 55 cases.

The research team recognizes that the left-turn capacity obtained with TRAF-NETSIM needs to be validated with field data and that some adjustments may be necessary. An extensive design of experiments will also be needed to examine the proposed procedures under various conditions. Nonetheless, the preliminary performance results indeed indicate the promising future of the proposed model as well as procedures. Hence, with rigorous data validation,

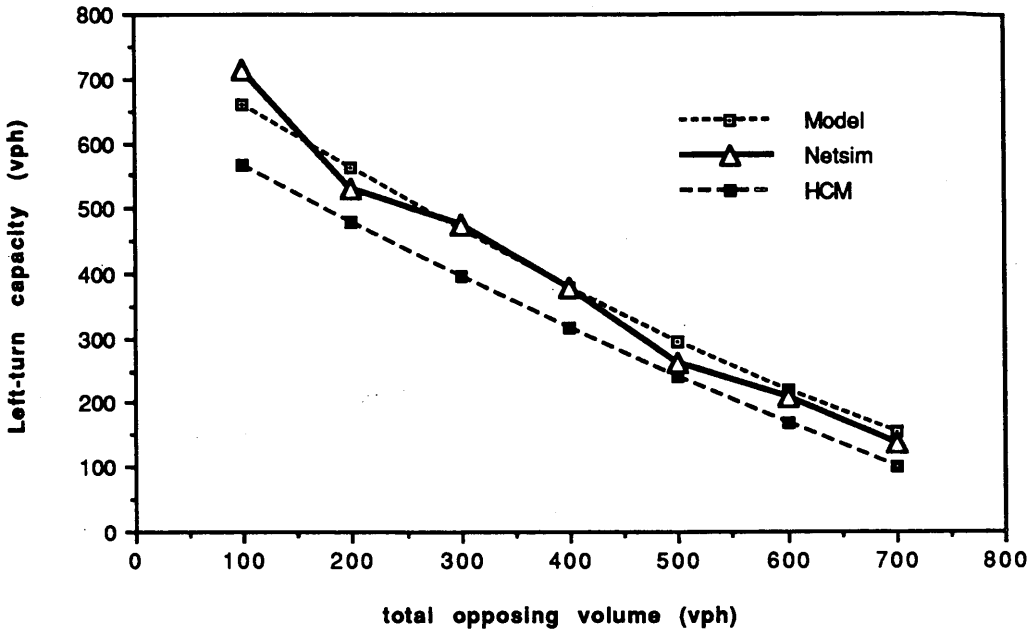


FIGURE 3 Results of Scenario A.

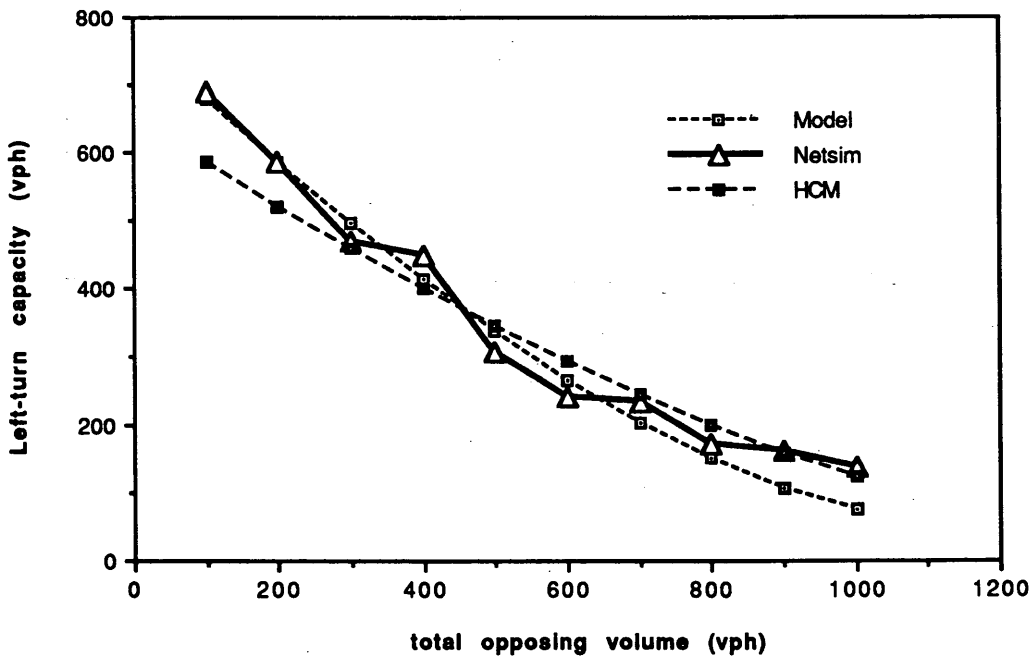


FIGURE 4 Results of Scenario B.

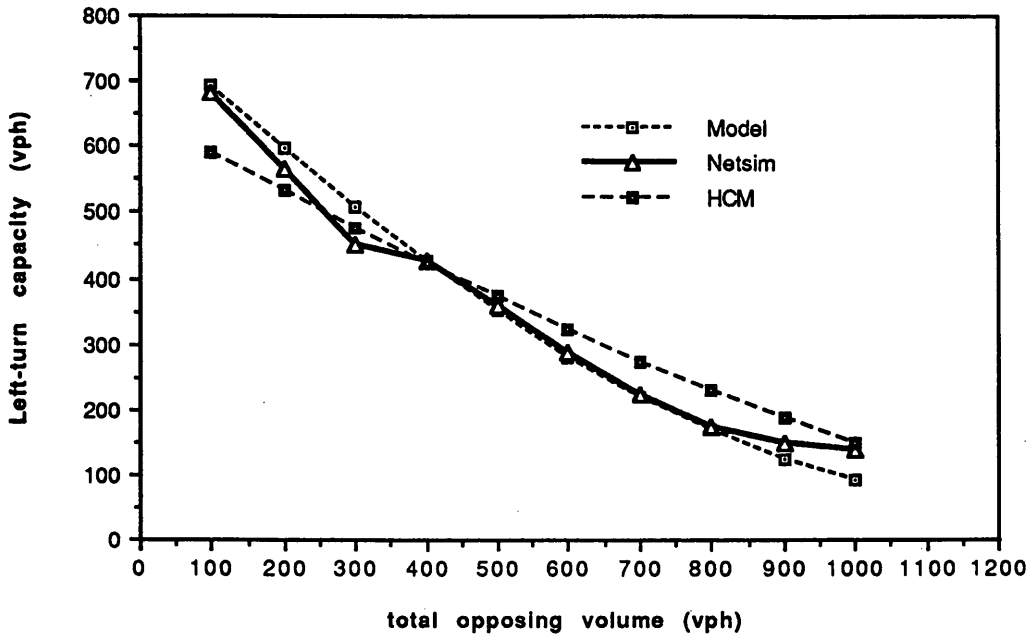


FIGURE 5 Results of Scenario C.

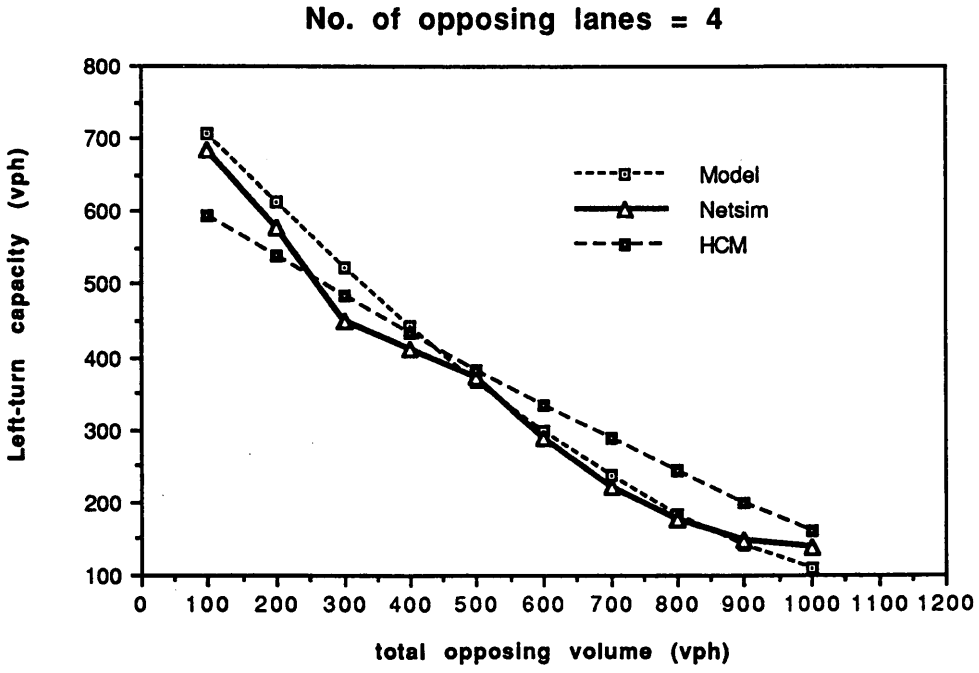


FIGURE 6 Results of Scenario D.

a convenient yet reliable empirical model for left-turn analysis may be achieved.

CONCLUSION

This paper has presented an exploratory procedure for analyzing the permitted left-turn capacity with exclusive lanes, including several empirical models for opposing queue length prediction, permitted saturation flow of mixed traffic, and effect of bay length on left-turn capacity. Some critical factors such as the number of opposing lanes and the interactions between upstream and downstream *G/C*'s have been incorporated in the proposed procedures for capacity estimation. A discrete choice modeling methodology has been applied to predict the fraction of time in a cycle during which the through queue length may be over certain distance. Such a model enables traffic engineers to determine the left-turn bay length from a cost-benefit perspective.

It should be noted that all proposed empirical models are grounded on the simulation experiments with TRAF-NETSIM. Hence, adjustments or modifications may be necessary after extensive field observations have been conducted to calibrate TRAF-NETSIM, which is one of the major tasks in the research project. Some capacity-related parameters, such as discharging headway and truck left-turn processing time, can be estimated from the field data.

Ongoing research along this line includes the development of (a) an analytical model for permitted saturation flow, considering both platoon size and the number of lanes; (b) an operational analysis procedure for protected/permitted and permitted/protected control; and (c) some guidelines for selection of critical capacity-related variables from field data.

REFERENCES

1. Rouphail, N., M. Magnuson, and V. Sisiopiku. *Validation of 1985 HCM Procedures for Capacity and Los of Left Turn Lanes in Illinois, Final Report.* Report FHWA/IL/RC-012. FHWA, U.S. Department of Transportation, May 1991.
2. Roess, R. P., J. M. Ulerio, and V. N. Papayannoulis. Modeling the Left-Turn Adjustment Factor for Permitted Left Turns Made from Shared Lane Groups. In *Transportation Research Record 1287*, TRB, National Research Council, Washington, D.C., 1990, pp. 138–150.
3. Roess, R. P. Development of Analysis Procedures for Signalized Intersections in the 1985 Highway Capacity Manual. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987, pp. 1–16.
4. Teply, S., and A. M. Jones. Saturation Flow: Do We Speak the Same Language? In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 144–153.
5. *United Kingdom Method.* U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1986.
6. Peterson, B. E., A. Hansson, and K. L. Bang. Swedish Capacity Manual. In *Transportation Research Record 667*, TRB, National Research Council, Washington, D.C., 1978, pp. 1–28.
7. *Swedish Highway Capacity Manual.* National Swedish Road Administration, 1977.
8. Akçelik, R. *Traffic Signal: Capacity and Timing Analysis.* Research Report ARR 123. Australian Road Research Board, Nunawading, 1989.
9. Chang, G. L. *A Study to Develop Operational Analysis Procedures for Exclusive Left-Turn Lanes.* FHWA Interim Report; DTFH 61-92-00109. FHWA, U.S. Department of Transportation, 1993.

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

Operational Characteristics of Triple Left Turns

JOHN D. LEONARD II

A study of the characteristics of triple left turns was performed. Five triple-left-turn sites in Orange County, California, were identified. Manual saturation flow rate studies at each site were performed using electronic counter boards. Queue discharge times and vehicle arrivals were compiled for all vehicles by lane and by signal cycle. Across all five sites, a sample consisting of 4,742 lane cycles and 34,898 vehicles was compiled for analysis. On average, these triple left turns supported flows of 795 vehicles per hour, received a 19 percent split of the total cycle time and spent 57 percent of that split time servicing the queue. Computing the saturation flow rate using the method suggested by the 1985 *Highway Capacity Manual*, the average saturation flow rate observed was 1,930 vehicles per hour of green per lane. Variations in the saturation flow rate observed at the triple-left-turn sites were explored. Results reveal no significant differences in saturation flow rates when categorized by site, by weekdays (e.g., Monday through Friday), or by observer. Significant differences were observed between lanes (e.g., inner and middle versus outer), time of day (e.g., morning, midday, and evening) and weekday versus weekend.

A triple left turn is a left-turning movement with three lanes available to the turning vehicles. U-turns may or may not be allowed from the innermost lane, and through movements may or may not be allowed from the outermost lane. In theory, a triple left turn provides additional capacity at the stop line, providing for a greater discharge of turning vehicles over a shorter amount of time. This may result in shorter cycle times at the intersection and additional green time for other traffic movements within a fixed time budget.

Traffic professionals have expressed several concerns over the installation of triple left turns. In general, these concerns may be categorized into safety issues and operational issues. Safety-related questions focus on driver traversal of the triple left turn. For example, special road bumps ("cat tracks") have been installed to channel the drivers correctly through the left turn and into the adjoining approach. Other safety concerns include truck and bus negotiation through such turns.

Operational concerns related to triple left turns focus on the perceived increase in capacity at the stop line. Stop line capacity is usually described as "saturation flow." Teply and Jones (*1*) present an excellent review of different descriptions of saturation flow. For purposes of this paper, the definition presented in Chapter 9 of the *Highway Capacity Manual* (HCM) (*2*) is adopted.

Although it is recognized that a triple left turn provides an absolute increase in the number of turn lanes for the left-turning movement, this increased capacity may be offset by a reduction in the discharge (saturation flow) rates due to driver "cautiousness" during turn traversal, especially of drivers in the center lane. For example, the HCM (*2*, Table 9–12) suggests that adjustment factors of 0.95 and 0.92 be applied to the ideal saturation flow rate to exclusive, protected single and dual left-turning movements, respectively.

The principal goal of this research is to provide an assessment of the current operational experience with triple left turns in California. The objectives of this research are to document operating characteristics associated with triple left turns, including flows serviced by the turns, saturation flow rates, and various signal timing characteristics.

DATA

Five triple-left-turn sites were selected on the basis of discussions with the California Department of Transportation (Caltrans); all of the sites were located in Orange County, in Southern California. Table 1 presents a summary of these sites along with other pertinent data. Figures 1 through 4 present aerial photographs of four of the triple left turn sites; aerial photographs of the Paseo De Valencia at Los Alisos site (Site 3) were not available. Sites 1, 4, and 5 are located within 2 km of each other and are located on the Pacific Coast Highway (PCH) near Newport Bay within the city of Newport Beach. As-built plans were available for these three sites; lane widths for these three sites are given in Table 1. Geometric plans for Sites 2 and 3 were unavailable. Lane usages were obtained through field observations at each site.

Signal phasings at each site are varied. All sites present protected-only signal phasing to the triple left turns. Sites 2 and 3, because of restricted geometrics (e.g., freeway off-ramp and T-intersection), display five and four signal phases, respectively. Site 5 uses six signal phases, displaying green to combined left and through movements to the triple-left-turn approach (i.e., split phase). Sites 1 and 4 display eight traffic signal phases.

Field observations were collected during daylight hours; the weather during all observation periods was clear and dry. Approaches at all sites are generally level (e.g., zero grade.) On-street parking is prohibited at all sites. All sites are located in generally suburban regions. Truck and other heavy vehicle percentages were not explicitly recorded, although all sites could be best characterized as having relatively low percentages (e.g., 0 to 2 percent) of large vehicles.

METHODOLOGY

Data were collected manually using an electronic counting board (*3*) capable of recording keystroke-time stamp combinations to a resolution of 1/64 sec; using this information, it is possible to construct the time histories of vehicles discharging from the stop line. These types of counting boards also enable one observer to record

School of Civil Engineering, Georgia Institute of Technology, Atlanta, Ga. 30332.

TABLE 1 Summary of Triple-Left-Turn Sites

Site Code	Description	City	Short Name	State Route?	Signal Phases	Lane Usage	Lane Widths		
							Inner	Outer	Middle
1	EB Dover to SB Pacific Coast Highway	Newport Beach	PCH/Dover	Yes	8	L-L-L	11'	11'	11'
2	SB Interstate 5 to EB Lake Forest Road	Lake Forest	I-5/LakeF	Yes	5	L-L-LT	.	.	.
3	NB Paseo De Valencia to WB Los Alisos	Laguna Niguel	PDV/LosAl	No	4	L-L-L	.	.	.
4	EB Pacific Coast Highway to NB Jamboree Blvd	Newport Beach	PCH/Jamb	Yes	8	L-L-L	11'	11'	11'
5	NB Bayside Drive to WB Pacific Coast Highway	Newport Beach	PCH/BaySi	Yes	6	L-L-LT	11'	13'	10'

* Lane widths for Sites 2 and 3 were unavailable.

the discharge patterns of all three lanes of the triple left turn movement simultaneously. Results of the field observations may then be translated into cycle-by-cycle summaries of the discharge patterns of the vehicles.

The overall project methodology is as follows:

1. Perform field observations at each site,
2. Construct cycle-by-cycle summaries from the raw data files, and
3. Perform statistical analysis of data.

Field Observation Methodology

The data collection procedure adopted for this field survey is very similar to the manual method described in the HCM (*2*), the principal differences being that a single observer may record data over several lanes and that the discharge times of all vehicles may be recorded. For this field survey, a custom data collection procedure was designed and custom software for data reduction was developed.

Reference point selection and observer positioning follow the recommendations of the HCM (*2*, Chapter 9, Appendix IV). The observer first selects a reference point near the stop bar that is common to all lanes of the triple left turn and then positions himself or herself at the intersection such that the reference points in each lane are visible. For this study of triple left turns, the rear bar of the pedestrian crosswalk was selected as the reference point.

Construction of Cycle-by-Cycle Summaries

Given that a series of keystroke data files have been uploaded from the electronic counter board, the next step of the study methodology involves the construction of cycle-by-cycle data summaries. (The author will provide a detailed description of this procedure on request.) These summaries will be used as input into the next step (statistical analysis). For purposes of this study, a cycle is defined

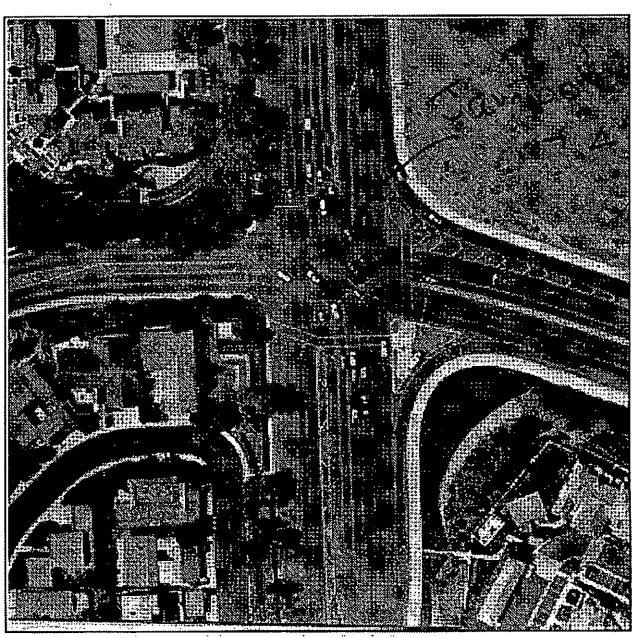


FIGURE 1 Pacific Coast Highway at Jamboree Road.

FIGURE 2 Pacific Coast Highway at Bayside Drive.

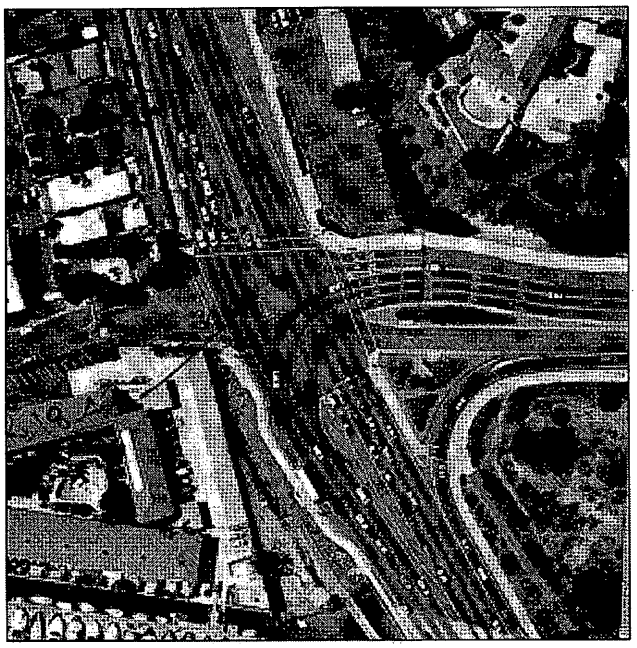


FIGURE 3 Pacific Coast Highway at Dover Drive.

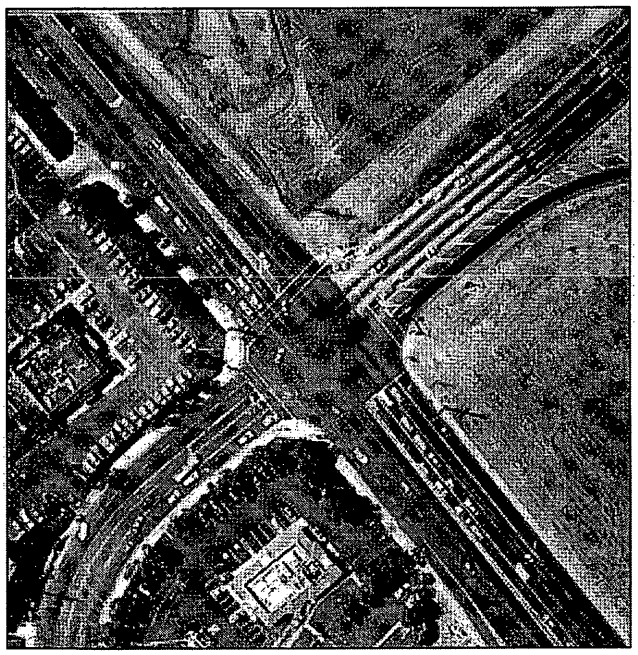


FIGURE 4 Southbound Interstate 5 at Lake Forest Road.

as the period starting at the onset of green and ending with the onset of the next green. Each cycle is divided into three time intervals: green, yellow, and red. Queues are discharged during the green and are accumulated during the yellow and red. Vehicles arriving during the red portion of a cycle are accumulated and then discharged during the succeeding cycle.

Some errors may be introduced during the data collection process. These errors generally result from unobserved, late, or miscoded events (e.g., the observer missed a vehicle arriving into the queue, pressed the button marking the onset of yellow, or pressed the wrong button upon the onset of red). During the construction of the event stream, an attempt is made to identify these types of errors. If the errors are identified, what appear to be good data are salvaged. If inconsistencies cannot be resolved, the data associated with the current cycle are not stored.

Statistical Methodology

The cycle-by-cycle data file constructed may be analyzed using any of a number of standard statistical packages. Two-tailed tests of statistical significance were evaluated using a 95 percent confidence interval.

RESULTS

Observed Vehicle and Cycle Summaries

Field observations were collected at five triple-left-turn sites during the morning, midday, and evening peak travel periods. Twenty-nine field surveys were performed.

A total of 34,898 vehicles over 4,742 lane cycles were observed at all triple-left-turn sites. A lane cycle is defined as the period associated with a traffic signal cycle for a single lane of traffic. For example, three lane cycles are observed at a triple-left-turn site for each traffic signal cycle, one lane cycle for each lane. However, there does not necessarily exist an exact 3:1 correspondence between the number of lane cycles and signal cycles. During periods of low traffic demand (e.g., just before or just after a peak period), one or more of the lanes of the triple left turns would not develop queues. In some cases (e.g., data collected during the first few days of observation) queues with fewer than two or three vehicles queued were arbitrarily omitted by the observer. In later field surveys the observer recorded events for all cycles observed, whether a vehicle was present or not; lane cycles without queued vehicles are designated as missing values for all tabulations. (Unless otherwise stated, use of the word "cycle" is understood to represent a "lane cycle" throughout the rest of the paper.)

Table 2 presents a tabulation of the cycle observations by lane (e.g., inner, outer, and middle) and by time of day (e.g., morning, midday, and evening, or AM, MD, and PM, respectively). Cycles are classified into time-of-day categories using the clock time at onset of the green: AM, before 10:00 a.m.; PM, after 3:00 p.m.; and MD, from 10:00 a.m. to 3:00 p.m. One site (PCH and Jamboree) provides data collected over all three periods, and one site (PCH and Dover) provides data collected over two periods (AM and MD); data at the remaining sites were collected during one period only.

Cycle observations are distributed equally across all three lanes of travel. For the inner and middle lanes, 1,583 and 1,587 valid lane cycles were observed. Cycles without queues were omitted from this sample. Differences in the number of observed cycles would suggest that vehicles would tend to queue first by using the inner and middle lanes and then by using the outer lane.

Table 3 categorizes the cycle observations by day of week. At one site (PCH and Jamboree) data were collected on all days of the week and during all times of day. Data at other sites are not as complete. One reason is that some sites did not exhibit any queueing at specific times of day (e.g., weekend mornings.) A limited project budget also contributed to this reduced coverage. However, all sites

TABLE 2 Cycle Summaries by Lane and Time of Day

Site	Total Vehicles Observed	Lane			Time-Of-Day			Total
		Inner	Outer	Middle	AM	MD	PM	
PCH/Dover	1219	56	49	55	68	92	.	160
I5/LakeF	2370	70	69	76	.	215	.	215
PDV/LosAl	4541	206	210	211	.	.	627	627
PCH/Jamb	26392	1226	1221	1222	1525	772	1372	3669
PCH/Baysi	376	25	23	23	.	.	71	71
All Sites	34898	1583	1572	1587	1593	1079	2070	4742

TABLE 3 Cycle Summaries by Day of Week

Site Name	Day Of Week							Total
	Sun	Mon	Tue	Wed	Thu	Fri	Sat	
PCH/Dover	92	.	.	68	.	.	.	160
I5/LakeF	.	.	132	.	83	.	.	215
PDV/LosAl	.	.	.	290	250	87	.	627
PCH/Jamb	246	618	541	670	752	752	90	3669
PCH/BaySi	.	.	71	.	.	.	.	71
All Sites	338	618	744	1028	1085	839	90	4742

TABLE 4 Queueing Characteristics

Site Name	Vehicles In Queue at Onset of Green												Total
	1	2	3	4	5	6	7	8	9	10	11	12+	
PCH/Dover	.	.	.	19	42	45	31	19	3	1	.	.	160
I5/LakeF	.	1	.	5	33	47	44	35	25	21	4	.	215
PDV/LosAl	35	73	85	89	113	74	56	37	24	14	8	19	627
PCH/Jamb	184	346	513	544	605	479	440	319	121	82	23	13	3669
PCH/BaySi		.	2	46	16	6	1	.	.	.	.		71
All Sites	219	420	600	703	809	651	572	410	173	118	35	32	4742

Mean: 5.16 Vehicles, Mode: 5 Vehicles, Median: 5 Vehicles, Maximum: 17 Vehicles

include at least one period of data collected on a midweek day (e.g., Tuesday, Wednesday, or Thursday.)

Table 4 presents a tabulation of the frequency of lane cycles categorized by numbers of vehicles in the queue. Lane cycles are assigned categories on the basis of number of vehicles that were observed in the standing queue at the onset of green. Each entry in the table represents the number of cycles with the observed queue size. Across all sites, the mode queue size is 5 vehicles, the mean queue size is 5.16 vehicles, the median queue size is 5 vehicles, and the maximum queue size is 17 vehicles.

Saturation Flow Rates

Using the HCM definition, saturation flow rate is the flow in vehicles per hour that could be accommodated by a lane group assuming that the green phase was always available. In general, the maximum flow rate (e.g., saturation flow) may be observed during queue discharge from the stop bar. The HCM defines the period of saturation flow as beginning when the rear axle of the fourth vehicle in the queue crosses the reference point and ending when the rear axle of the last queued vehicle at the beginning of the green crosses the same reference point. For this study of triple left turns, given that the time at discharge was recorded for each vehicle observed during each cycle and that the backs of queues were tracked, alternative saturation flow rate computations using the fields observations may be evaluated.

Table 5 presents a summary of saturation flow rates at each site computed using alternative calculation techniques. Each calculation method is associated with a designation "Drop X," where X represents the number of vehicles dropped from the front of the queue when determining the mean discharge headway. For example, the mean discharge headways for scenario Drop 4 are computed by subtracting the discharge time recorded for the fourth vehicle from the discharge time recorded for the last vehicle to clear and dividing this total headway by the number of gaps between the fourth through the last vehicles—specifically, $n - 4$ where n represents the total vehicles in the queue at the onset of green.

The "None" scenario represents the saturation flow rates computed using all vehicles in the queue and the total discharge time

TABLE 5 Alternative Saturation Flow Rate Computations

Site Name	Vehicles Dropped from Front of Queue						
	None	Drop 1	Drop 2	Drop 3	Drop 4*	Drop 5	Drop 6
PCH/Dover	1673	1754	1835	1871	1939	1981	2077
I5/LakeF	1651	1703	1775	1824	1877	1923	1942
PDV/LosAl	1963	1831	1874	1951	1989	2020	2083
PCH/Jamb	1797	1805	1829	1868	1921	1952	2023
PCH/Baysi	1539	1632	1739	1823	1997	2157	1477
All Sites	1804	1799	1831	1875	1928	1959	2024

*Drop 4 corresponds with the suggested HCM calculation method.

from the onset of green. These values are typically lower than the other values, which may be attributed to start-up loss times associated with vehicles near the beginning of the queue. As one would expect, as these initial vehicles are discarded from the calculation, the saturation flow rate increases. The Drop 4 scenario corresponds with the HCM definition of saturation flow rate period, specifically, the period measured from the fourth through the last vehicle in the queue at onset of green.

As additional vehicles are ignored from the calculation, saturation flow rates averaged over all sites increase from about 1,800 to 2,030 vehicles per hour of green per lane (vphgpl). The HCM-suggested method (e.g., Drop 4) results in an observed saturation flow rate of about 1,930 vphgpl over all sites.

Using the Drop 4 value in Table 5, saturation flow rate values at each site range from 1,880 vphgpl at Site 2 (the freeway off-ramp) to 2,000 vphgpl at Site 5. However, an analysis of variance (Table 6) reveals that these means are not significantly different from the overall group mean of 1,930 vphgpl.

The observed value of 1,930 vphgpl is much larger than the "ideal" saturation flow rate of 1,800 vphgpl suggested by the HCM. As such, computation of a left-turn adjustment factor using only these data is not possible. Additional data collected at the same sites for single and double left turns and for through movements would allow explicit identification of this adjustment factor. However, by accepting several conservative assumptions one may estimate the value of the ideal saturation flow rate at these sites. Assume a lane width of 11 ft, a heavy truck percentage of 2 percent, and the same left-turn adjustment factor applied to exclusive double-left-turn lanes. The HCM saturation flow rate equation (2, Equation 9–8) may be written as

$$s_{lt3} = s_o * f_w * f_{HV} * f_{LT}$$

where

s_{lt3} = saturation flow rate for triple left turns under prevailing conditions,
s_o = ideal saturation flow rate,
f_w = lane width adjustment factor (0.97),
f_{HV} = heavy vehicle adjustment factor (0.99), and
f_{LT} = left-turn lane adjustment factor (0.92).

Under these assumptions, and using the factors obtained from the appropriate tables of the HCM, the ideal saturation flow rate for these five sites would be approximately 2,180 vphgpl.

Variations in Saturation Flow

The influences of site, lane, time of day, and day of week on the observed saturation flow rates of triple left turns were investigated. Table 6 gives a summary of the results of this series of analysis of variance tests. The table presents the explanatory variable under study, the sample size used in the calculation of the F-scores, the degrees of freedom (e.g., categories in explanatory variable less 1), the F-score and a level of significance determined using the F-score, and the degrees of freedom within the sample. All tests of significance are evaluated at the .05 level.

TABLE 6 Saturation Flow Rate Analysis of Variance

Primary Effect	Sample Size	Deg.Of Freedom	F-Score	Significance of F	Reject Null?
Site Code (1-5)	2784	4	2.063	0.083	No
Lane (Inner,Outer,Middle)	2784	2	4.113	0.016	Yes
Time of Day (AM,MD,PM)	2874	2	15.344	0.000	Yes
Weekdays Only (Monday-Friday)	2441	4	1.964	0.097	No
Weekday/Weekend (M-F vs. Sat-Sun)	2784	1	21.769	0.000	Yes
Observer (Two alternates)	2784	1	0.142	0.706	No

All Secondary Interaction Effects (e.g., Lane with TOD, etc.) are not significant at 0.05 Level.

Lane Utilization

Saturation flow rates categorized by lane ranged from 1,890 vphgpl in the outer lane to about 1,950 for the inner and middle lanes (Table 7). These differences are statistically significant and suggest that the outer lane of a triple-left-turn group will exhibit a reduced capacity from the inner and middle lanes. A lane utilization factor based on the mean value of saturation flow for the entire triple-left-turn lane group may be proposed. Let

$$s_{lg} = f_u * s_{lt3}$$

where s_{lg} is the saturation flow rate of the lane group within a triple left turn and f_u is the lane utilization factor. Substituting values of 1,930 vphgpl for the saturation flow rate under prevailing conditions and 1,950 and 1,890 vphgpl for inner/middle and outer lane groups, respectively, lane utilization factors of 1.01 for the inner/middle lane group and 0.98 for the outer lane group may be computed.

One might attribute the lower saturation flow rate of the outer lane to the fact that some sites share left and through movements from this lane. However, an examination of Site 4, which consists of three exclusive left-turn lanes, each with similar lane widths, reveals the same distribution of saturation flow rates between lanes.

Time-of-Day Variations

The data suggest statistically significant variations in saturation flow rates by time of day (e.g., morning, midday, and evening). Values of 1,990, 1,860, and 1,920 vphgpl were observed for the morning, midday, and evening periods (Table 7). A time-of-day adjustment factor may be proposed as

$$s_t = f_t * s_{lt3}$$

where s_t is the saturation flow rate for a specific time-of-day period and f_t is the adjustment factor. Corresponding adjustment factor values of 1.03, 0.96, and 0.99 may be computed for the morning, midday, and evening periods.

These results support the hypothesis that the population of drivers may be classified into two subpopulations: commuting and noncommuting, with each subpopulation exhibiting significantly different driving characteristics (e.g., acceptable discharge headways as measured by saturation flow rate). Different times of day would be composed of different proportions of commuting and noncommuting drivers. Commuting drivers concentrate their activity during the morning and evening peaks; they tend to be more aggressive and accept smaller headways. Noncommuting drivers (e.g., shoppers and tourists) concentrate their activity in the midday and evening periods; they tend to be less aggressive and accept larger headways. The a.m. peak period (with an observed mean saturation flow rate of 1,990 vphgpl) consists of primarily commuter drivers, the midday period (1,860 vphgpl) consists of primarily noncommuting drivers, and the p.m. peak consists of a mix of commuting and noncommuting drivers and exhibits a saturation flow rate of 1,920 vphgpl, approximately midway between the morning and midday periods.

Day-of-Week Variations

The data also suggest significant variations in saturation flow rates observed on weekends (Saturday and Sunday) versus weekdays (Monday through Friday). The weekday saturation flow rates over all sites was observed to be 1,940 vphgpl, whereas the group mean for weekends was observed to be 1,810 vphgpl. Differences in saturation flow rates between individual weekdays and the overall weekday group mean were not statistically significant. A day-of-week utilization factor may be proposed:

$$s_d = f_d * s_{lt3}$$

where s_d is the saturation flow rate for a particular day of week (e.g., either weekday or weekend) and f_d is the day-of-week adjustment factor. Adjustment factor values of 1.01 and 0.94 are calculated for weekdays and weekend days, respectively.

Other Variations

As a consistency check, the saturation flow rates measured by two different observers were compared. Results suggest that the saturation flow rates measured by each observer are not significantly different from the population mean estimated for the entire sample.

All secondary interaction effects (e.g., lane utilization by time of day, or time of day by day of week) were not significant at the 0.05 level. The lack of significance between interactions of these variables simplifies implementation of the proposed adjustment factors.

TABLE 7 Observed Saturation Flow Rates by Lane and Time of Day

	Lane			Time-Of-Day			
Site	**Inner**	**Outer**	**Middle**	**AM**	**MD**	**PM**	**Overall**
PCH/Dover	1938	1894	1979	1977	1908	.	1939
I5/LakeF	1888	1913	1834	.	1877	.	1877
PDV/LosAl	1954	2005	1994	.	.	1989	1989
PCH/Jambo	1948	1868	1954	1992	1838	1888	1921
PCH/Baysi	2209	1942	1655	.	.	1997	1997
All Sites	1946	1891	1950	1991	1856	1921	1928

Values computed using *n-4* vehicles in queue (HCM suggested)

TABLE 8 Signal Timing Characteristics of Triple Left Turns

Site Name	Flow (vphgpl)	Sat Flow (vphgpl)	Green (secs)	Yellow (secs)	Cycle (secs)	Split (%)	Busy (%)
PCH/Dover	274	1939	21	3.3	100	21	65
I5/LakeF	382	1877	29	3.8	104	28	56
PDV/LosAl	298	1989	21	4.1	87	25	46
PCH/Jamb	253	1921	18	3.0	103	17	59
PCH/BaySi	188	1997	15	2.8	103	15	71
All Sites	265	1928	19	3.2	101	18	57

It should be noted, however, that during the weekend mornings, significant queues did not form at any of the sites. As a result one should restrict application of the time-of-day factor to weekdays only.

Signal Timing Characteristics

By using the discharge times of vehicles in the queue in conjunction with the observed times of onset of the green, yellow, and red timing intervals, summary statistics of the signal timing characteristics of the observed triple left turns may be compiled.

Table 8 presents a summary of the signal timing characteristics for triple left turns. For each site the observed flow, saturation flow rate, observed green time, observed yellow time, observed cycle time, observed split time, and observed busy time are presented. Values of flow represent averaged per-lane volumes. Saturation flow rates are measured in vehicles per hour of green per lane and represent the per-lane saturation flow rates determined earlier. Values of green, yellow, and cycle represent the mean observed times of the respective signal timing intervals. For the sample of triple left turns, average interval lengths of 19, 3, and 101 sec were observed for the green, yellow, and cycle times, respectively. The observed split time (in percentage) represents the ratio of green time to cycle time.

The busy time represents the proportion of the green time spent discharging the queue. It is very similar to the degree of saturation, the main difference being that the degree of saturation is computed using all vehicles discharged during the cycle, whereas the busy time is computed using only the vehicles discharging from the queue. In general, the busy time will be less than the degree of saturation value, as some vehicles pass through the intersection after the queue has completely discharged. In the limiting cases, when all vehicles arrive on the red and discharge from the queue, the busy time will equal the degree of saturation. Conversely, when all vehicles arrive on green and no queue was accumulated during the preceding cycle, the busy time will be 0.

For the sample of triple left turns observed, the triple-left-turning movement services 795 vehicles per hour (vph) while receiving only on average 19 percent of the total cycle time. This demonstrates the potential capacity improvement at intersections without adversely affecting the length of the cycle. A busy time of less than 60 percent suggests the possibility of further reducing the split time without hurting the throughput of the triple left turn.

SUMMARY AND CONCLUSIONS

A study of the characteristics of triple left turns was performed. Five sites in Orange County, California, were identified. Manual saturation flow rates at each site were studied using electronic counter boards. Queue discharge times were collected for all vehicles by lane and by cycle.

Across the five sites, a sample consisting of 4,742 lane cycles and 34,898 vehicles was compiled for analysis. On average, these triple left turns supported flows of 795 vph, received a 19 percent split of the total cycle time, and spent 57 percent of that split time servicing the queue. The average saturation flow rate observed over all sites under prevailing conditions was approximately 1,930 vphgpl.

Variations in the saturation flow rate observed at the triple-left-turn sites were also investigated. Results reveal no significant differences between saturation flow rates by site, between weekdays, or by observer. Significant differences were observed between lanes (e.g., inner and middle versus outer), time of day (morning, midday, and afternoon), and time of week (weekday versus weekend). Appropriate saturation flow rate adjustment factors have been suggested.

ACKNOWLEDGMENTS

This research was made possible with the cooperation and assistance of the staff members at the various district offices of Caltrans. Special thanks are extended to Fred Rooney, who served as project monitor for Caltrans headquarters. The Institute of Transportation Studies provided administrative and clerical support for the research: Ziggy Bates coordinated the word processing and Anne-Marie Defeo provided contract administration. Fred Kim and Bill Shao did outstanding jobs of data collection. Financial support was provided by the State of California Department of Transportation and FHWA.

The author wishes to also thank the reviewers whose comments and suggestions have improved the quality of this paper.

REFERENCES

1. Teply, S., and A. Jones. Saturation Flow: Do We Speak the Same Language? In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991.
2. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
3. *IMC-IV User's Manual.* Jamar Sales Company, Ivyland, Pa., 1989.

The views expressed in this work remain the author's and do not necessarily reflect those of the sponsors; the author is solely responsible for any errors.

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

Saturation Headways at Stop-Controlled Intersections

MICHAEL KYTE, ZONGZHONG TIAN, JULIA KUHN, HEIDI POFFENROTH, MARC BUTORAC, AND BRIAN ROBERTSON

Capacity analysis procedures for stop-controlled intersections require saturation headways or related parameters as inputs. Unfortunately, no data base currently exists for these parameters, including critical gaps and follow-up times for two-way stop-controlled (TWSC) intersections and saturation headways for all-way stop-controlled (AWSC) intersections, for conditions found in the United States. The results of a set of recent measurements of these parameters are reported, and several important issues are explored: (a) how are the critical gap and follow-up gap measured at a TWSC intersection and what is the relationship between them? (b) what is the saturation headway (i.e., follow-up gap) for a TWSC intersection? (c) what is the effect of turning movements on the saturation headway at an AWSC intersection? and (d) should other cases be considered, in addition to the standard four cases, when measuring the saturation headway at an AWSC intersection? For TWSC intersections, a relationship was found between the critical gap and the follow-up time. In addition, the importance of the directional movement of the major street vehicle terminating a gap as it affects the critical gap was determined. A new set of categories was developed for saturation headway cases for AWSC intersections, and the importance of the movement direction of the subject approach vehicle on the saturation headway was determined.

The basic parameter used to estimate capacity at a signalized intersection is saturation headway. Ideal saturation headway is the difference in the passage time at the intersection stop line between two consecutive vehicles once the queue is moving in a stable manner. The 1985 *Highway Capacity Manual* (HCM) (*1*) notes that the saturation headway is "estimated as the constant average headway between vehicles which occurs after the 6th vehicle in the queue and continues until the last vehicle in the queue clears that intersection." Field measurements must consider the start-up lost time, or that time at the beginning of the green phase that is required for the queue to begin to move. The capacity procedures given in Chapter 9 of the HCM provide a standard value for the ideal saturation headway of 2.0 sec/veh, which yields an ideal saturation flow rate of 1,800 vehicles per hour (vph) of green. The procedure provides adjustments to this ideal value to consider the effects of intersection geometry, opposing traffic flow, signal timing parameters, and pedestrian flows.

The capacity analysis procedure for unsignalized intersections is given in Chapter 10 of the HCM. A new version of Chapter 10 is planned for release in 1994, with an improved procedure for two-way stop-controlled (TWSC) intersections based on a capacity methodology developed by Siegloch and described by Brilon et al. (*2*). The chapter also includes a procedure for estimating the capacity of an all-way stop-controlled (AWSC) intersection based on *Transportation Research Circular 373* (*3*).

Department of Civil Engineering, University of Idaho, Moscow, Idaho 83843.

Both of the capacity procedures for stop-controlled intersections use the concept of saturation headway. The TWSC intersection procedure is defined in terms of the critical gap and the follow-up gap. The critical gap is the minimum time gap in the major traffic stream needed by a minor stream vehicle to merge into or travel through the major stream. The follow-up gap is the minimum headway between the first vehicle and the second vehicle, and subsequent vehicle pairs, as they enter the same major stream gap, when a continuous queue exists on the minor street approach. In effect, the follow-up gap is the saturation headway for the minor traffic stream when the conflicting major stream flow is zero.

Table 10-2 in the new version of Chapter 10 gives critical gaps ranging from 5.0 sec for major stream left-turning traffic to 6.5 sec for minor stream left-turning traffic. Follow-up gaps range from 2.1 sec for left-turning traffic from the major street to 3.4 sec for left-turning traffic from the minor stream. The capacity on the minor stream approach, based on Siegloch's work, is a function of the major stream flow rate (v_c), the critical gap (t_g), and the follow-up gap (t_f). The capacity equation is given in Equation 1.

$$c_p = \frac{3{,}600}{t_f} e^{-v_c t_0/3{,}600} \quad (1)$$

One of the problems with this procedure, however, is that it has not been validated with data collected from sites in the United States. Data in Table 10-2 were measured first in Germany and then slightly modified on the basis of studies of critical gap for a very limited number of sites in the United States. None of these U.S. studies attempted to measure the follow-up gap and assumed only the fixed relationship between the critical gap and the follow-up gap given in Equation 2:

$$t_g = 0.6\, t_f \quad (2)$$

A further complication is the inherent difficulty in measuring the critical gap. The HCM defines the critical gap as the median time headway between two successive vehicles in the major street traffic stream that is accepted by drivers in a subject movement that must cross or merge with the major street flow. Several researchers [e.g., Kittelson and Vandehey (*4*)] have pointed out the difficulty in using this definition. In fact, the formulation of the Siegloch equation is based on a very specific description of the gap acceptance process that may yield estimates of the critical gap that are different from those produced by the HCM definition. According to the Siegloch formulation, one vehicle will accept a major stream gap that is greater than the critical gap but less than the sum of the critical gap and the follow-up gap. Two vehicles will use a gap that is greater than the sum of the critical gap and the follow-up gap but less than the sum of the critical gap and twice the follow-up gap. To

measure the critical gap in this way, a continuous minor stream queue is required. Brilon et al. recommend the use of either the maximum likelihood technique or Ashworth's method if a continuous queue is not present on the minor street approach (*5*).

The AWSC intersection capacity procedure is based on a set of four saturation headways, each defined according to the conditions faced by the subject approach driver. Table 10-5 in the new version of Chapter 10 gives values of 3.5 sec/veh when the subject vehicle is faced with neither opposing nor conflicting stream vehicles and 9.0 sec/veh when the subject vehicle is faced with both opposing and conflicting approach vehicles. Table 1 presents the saturation headway from Table 10-5 of the new version of Chapter 10.

The capacity of an approach is based on the mix of traffic conditions faced by the subject approach driver and is defined in terms of the volume proportions of each of the intersection approaches. The capacity of an approach varies from 1,100 vph when the subject driver faces no opposing or conflicting vehicles to 525 vph when the subject driver faces a continuous queue of vehicles on both the opposing and conflicting approaches.

The four headway cases given in Table 1 do not consider directly the effects of turning traffic. The Case 2 headway, which is a subject vehicle faced by an opposing vehicle and no conflicting vehicles, does not consider the effects of the interaction of one or both of the vehicles turning and not traveling straight through the intersection. The value of 5.5 sec given in Table 1 is assumed to cover the range of combinations that actually make up Case 2: for example, pairs of through vehicles with no turning conflicts, one through vehicle opposed by a left-turning vehicle, one through vehicle opposed by a right-turning vehicle, and so on. Although the capacity equation given in the new version of Chapter 10 does provide an adjustment for turning movements, it is based only on the overall proportions of turning movements and not on the microscopic or vehicle-by-vehicle interactions that actually reflect the impedance resulting from turning vehicle conflicts.

STUDY OBJECTIVES

The purpose of this paper is to report on a study of saturation headway measurements made at stop-controlled intersections in order to explore several questions raised in the previous discussion; these issues include the following:

1. How are the critical gap and follow-up gap measured at a TWSC intersection? What is the relationship between the follow-up gap and the critical gap?
2. What is the saturation headway (i.e., follow-up gap) for a TWSC intersection?
3. What is the effect of turning movements on the saturation headway at an AWSC intersection?
4. Should other cases be considered, in addition to the standard four cases, when measuring the saturation headway at an AWSC intersection?

This paper also investigates one other issue important in the formulation of the capacity analysis procedure for TWSC intersections. The gap acceptance mechanism that is the basis for the TWSC intersection capacity analysis procedure assumes a priority among the various traffic streams at a TWSC intersection. Traffic streams assumed to conflict with each minor stream movement are identified, and the degree of conflict is specified. For traffic on the stop-controlled approach, right-turning vehicles arriving from the left on the major street are weighted by a factor of 0.5, indicating that although this group of major street vehicles affects the operation of the minor street traffic, the effect is less than that for the through major street traffic. However, the factor of 0.5 is based not on empirical data but on judgment only. This paper provides a procedure that may help to validate this relationship.

DATA COLLECTION AND REDUCTION METHODS

Data were collected at two sites for this study, at one AWSC intersection site and one TWSC intersection site. The AWSC intersection site is located in suburban westside Portland, Oregon. It has four legs with a single lane on each approach. One video camera was used to record traffic flow through the intersection. The camera was located so that all vehicles entering the intersection could be viewed and so that the queue activity on one approach could be viewed also. The TWSC intersection site is located in Pullman, Washington. It is a T-intersection, with two lanes (one each for left-turning and right-turning vehicles) on the stop-controlled approach. The major street has single lanes on each approach. One camera was used to record traffic operations, again recording all vehicle movements through the intersection as well as the queue activity on the stop-controlled approach. Since a continuous queue was present only for the minor street left-turn approach, only this movement was used for the analysis described later in this paper.

Vehicle passage times through the conflict point at the intersection were recorded using the Traffic Data Input Program (*6*) operating on an IBM-compatible personal computer. While observing the videotape of traffic traveling through the intersection, the pro-

TABLE 1 Saturation Headway Data for AWSC Intersections

Condition	Mean Saturation Headway, sec/veh			
	Case 1	Case 2	Case 3	Case 4
All data	3.5	5.5	6.5	9.0
Single lane approach sample sites	3.9	5.6	6.5	9.0
Multi lane approach sample sites	1.5	4.3	6.3	9.3

Notes:
Case 1: Subject vehicle does not face either opposing or conflicting vehicles.
Case 2: Subject vehicle faces only an opposing vehicle.
Case 3: Subject vehicle faces only conflicting vehicles.
Case 4: Subject vehicle faces both opposing and conflicting vehicles.

gram operator presses a key to record the desired events. The events of interest include the passage times of all vehicles as well as the times that each vehicle on the subject stop-controlled approach enters the end of the queue, arrives at the stop line, and enters the intersection. This effort produces a raw data file for each of the two intersections.

For the TWSC intersection, the raw data file was used to create a second file with the following variables for each subject approach (minor street left-turning) vehicle: the time that the vehicle entered the queue, the time that the vehicle arrived first in line at the stop line, the time that the vehicle left the stop line, and the passage times through the intersection of each higher-priority vehicle seen by the minor stream vehicle. This latter information was used to construct the gaps that were accepted and rejected by the minor stream vehicle and the pair of higher-priority vehicles that defined the beginning and end of each gap. A third data set was also created on the basis of the number of minor stream vehicles using each major traffic stream gap. Only data that were collected during the existence of a continuous queue on the minor street were used in creating the data sets.

For AWSC intersections, the raw event data file was used to create a record for each vehicle on the subject stop-controlled approach that included the following variables: the time that the vehicle arrived in the queue, the time that the vehicle arrived first in line in the queue, the time that the vehicle entered the intersection, and a list of opposing and conflicting vehicles that entered the intersection since the departure of the previous subject approach vehicle. This latter information allowed the determination of the saturation headway as well as conditions faced by the subject approach driver. Only those subject approach vehicles that were a part of a continuous queue were included in the data base.

TWSC INTERSECTION DATA ANALYSIS

Determination of Critical Gap and Follow-Up Gap

Gap acceptance theory defines the critical gap and the follow-up gap in a clear manner. Figure 1 illustrates these definitions for a critical gap of 5.0 sec and a follow-up gap of 2.5 sec. The theory states that one minor stream vehicle will use a gap that is greater than the critical gap and less than the sum of the critical and the follow-up gaps. As stated previously, the follow-up gap is just the saturation headway for the minor stream, since each time the major stream gap increases by the follow-up gap, one additional minor stream vehicle can be absorbed into the major traffic stream. The primary requirement for this mechanism to be used as the basis for field measurements is that the minor stream must have a continuous queue.

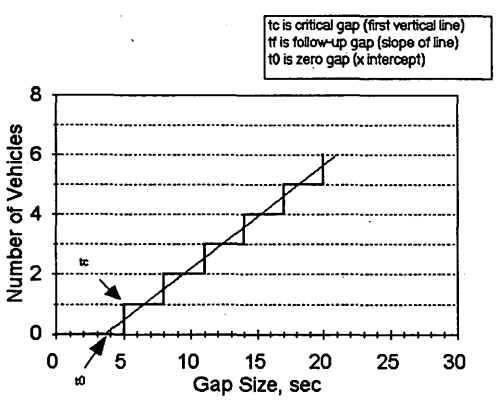


FIGURE 1 Gap acceptance mechanism.

Table 2 presents the range of major stream gap sizes used by various numbers of left-turning minor stream vehicles at the TWSC intersection used for this study during periods of continuous queueing on the minor stream approach. Figure 2 shows a plot of the individual gap sizes versus the number of vehicles using each gap. The mean gap size for each vehicle number is also shown. These mean values are used to estimate a regression line, whose parameters are then used to estimate the various gap parameters.

In Table 2 some of the vehicles-per-gap cells included only a few observations, even two or fewer. The regression line was plotted using the data for a range of one vehicle to four vehicles per gap, cells that included three or more observations. This line is shown in Figure 3. Several parameters of interest can be derived from the equation that forms the basis for the line. The follow-up gap is the reciprocal of the slope of the line. The zero gap is the x-intercept. The critical gap is the zero gap plus half the follow-up gap.

The parameter estimates that were developed from the regression line are as follows:

Gap	*Estimate (sec)*
Zero	3.0
Follow-up	3.3
Critical	4.7

Two comparisons can be made with respect to these parameters. In this case, the follow-up gap is equal to 0.70 of the critical gap. This compares with the value of 0.60 assumed in the current version of Chapter 10 of the HCM and a computed value of 0.52 using data provided in the new version of Chapter 10. The saturation flow rate, the reciprocal of the follow-up gap, is 1,090 vph. This compares with a value of 1,060 vph from the new version of Chapter 10.

Effect of Major Stream Right-Turn Vehicles on Critical Gap and Follow-Up Gap

Table 10-3 in Chapter 10 of the HCM gives the traffic streams that have priority over each minor traffic stream at a TWSC intersection. The table further describes the manner in which these conflicting volumes are to be summed in order to provide an estimate of the total conflicting volume faced by a given subject traffic stream. For example, the conflicting volume for the left-turning traffic on the minor traffic stream includes half of the major street right-turning volume from the left. The use of the one-half in this term has been justified as follows: although the minor stream left-turning traffic does not have to share intersection space with the major stream right-turning traffic arriving from the left, it is affected by this stream. But it is often difficult to know if a major stream vehicle will indeed turn right even if it has so indicated with its turn signal. This uncertainty means that the major stream right-turning movement does affect the behavior of the minor stream left-turning traffic. Using only half of this traffic volume recognizes the fact that the effect is not as great as that for the through major stream traffic. Again, the value of one-half is based on judgment only.

Data collected in this study allow the development of a procedure for the quantification of this effect. For the left-turning minor

TABLE 2 Number of Vehicles Accepting Gaps of Various Sizes

Number of Vehicles Using Gap	Mean Gap, sec	Standard Deviation	Maximum Value, sec	Minimum Value, sec	Obs
1	5.93	2.26	11.10	1.43	85
2	10.05	2.38	14.28	4.28	27
3	13.93	2.62	19.88	11.42	10
4	15.08	4.42	18.89	8.89	3
5	23.04	4.15	27.19	18.89	2
6	28.24	-	28.24	28.24	1
7	-	-	-	-	0
8	46.91	-	46.91	46.91	1

Note:
1. Obs is the number of observations.
2. The data shown in this table are for the left turning traffic from the minor street.

traffic stream vehicles, each gap that was accepted is classified into one of two categories: the first category includes those gaps that are terminated by a major street right-turning vehicle from the left; the second category includes all other gaps accepted by these left-turning minor stream vehicles. Table 3 shows a clear difference between these two cases. When a gap is terminated by a major stream right-turning vehicle from the left, more minor stream vehicles are likely to use a gap of a given size. This is also indicated in the size of the critical gap for these two cases. If a gap is terminated by a major stream right-turning vehicle from the left, the critical gap is estimated to be 3.2 sec. For all other gap termination combinations, the critical gap is estimated to be 50 percent higher, or 4.8 sec.

The significance of this relationship is more clear when the capacity equation is examined further. The issue under consideration here can be stated mathematically as follows. If t_{c1} is the overall critical gap for all minor stream left-turning vehicles (regardless of the conflicting vehicle that terminates the gap) and if t_{c2} is the critical gap for minor stream vehicles when the gap is terminated by a major street right-turning vehicle from the left, the correct adjustment to the conflicting volume equation is given by α in Equation 3:

$$\alpha\, v_{RT}\,(t_{c1} - 0.5\, t_f) = v_{RT}\,(t_{c2} - 0.5\, t_f) \quad (3)$$

where α is currently given as 0.5 in the HCM procedures, v_{RT} is the major street right-turning volume approaching from the left, and each side of Equation 3 is the exponent in the Siegloch capacity equation. If t_{01} and t_{02} are the zero gaps for the two cases described earlier, this relationship can be simplified by solving for t_{01} in terms of t_{02}, as given in Equation 4.

$$t_{02} = \alpha\, t_{01} \quad (4)$$

In this case, α is equal to 1.3 divided by 2.8, or 0.46. This is nearly equal to the factor of 0.5 now used in the conflicting volume equation. This method can be used to check the assumptions of conflicting volume used for other minor stream movements as given in Figure 10-3 of the HCM.

AWSC INTERSECTION DATA ANALYSIS

The new version of the HCM Chapter 10 describes a capacity analysis procedure based on a set of conditions faced by drivers on the

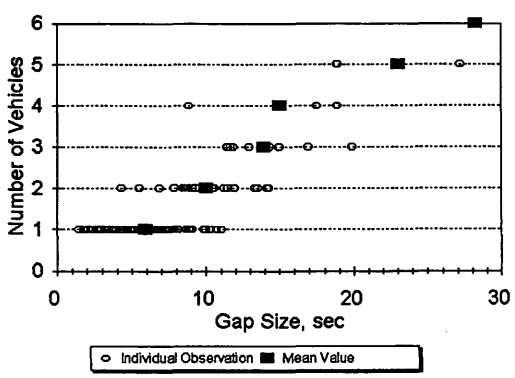


FIGURE 2 Gap size versus number of vehicles using gap (individual observation and mean value).

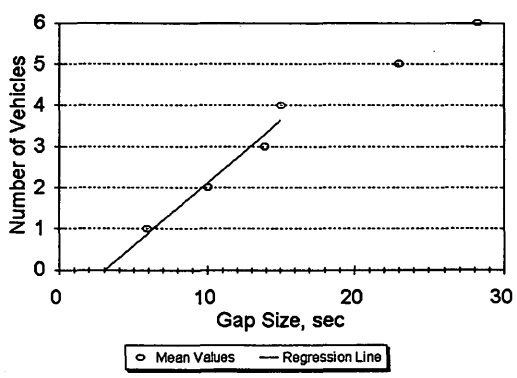


FIGURE 3 Gap size versus number of vehicles using gap (mean values and regression line).

TABLE 3 Effect of Vehicle Movement Terminating Accepted Gap

Vehicle Movement Terminating Accepted Gap	Mean Gap Size for Various Number of Vehicles Using Gap, sec			t_c, sec
	1 Vehicle	2 Vehicles	3 Vehicle	
Major street RT vehicle from the Left	4.9	9.6	12.2	3.2
All other major street vehicles	6.7	10.4	14.4	4.8

subject approach. The four cases, along with the saturation headways measured for each, are described in Table 1.

Although these data led to a more comprehensive capacity analysis procedure than was previously available, the procedure does have some obvious limitations. Most important, the four cases provide only a very simplified classification of the conditions actually faced by the subject approach driver. Case 2, for example, is the condition in which the subject approach driver is faced by a driver on the opposing approach. The turning movements of either driver, clearly important factors in the resulting saturation headway, are not considered.

Saturation headway data were collected for one approach of an AWSC intersection to determine if there were subsets of these four basic cases that could be established so that the capacity estimation procedure given in Chapter 10 could be improved. For each subject approach vehicle that was a part of a continuous queue, the saturation headway was measured and the conditions faced by the driver were identified, including the turning movement directions for all vehicles.

Two separate series of tests were conducted. First, for each of the four cases, the effect of the direction of the subject approach vehicle was determined. Second, subsets of Cases 3 and 4 were identified and tested.

Effect of Subject Approach Vehicle Movement

Table 4 gives a summary of the saturation headway data for each of the four cases according to the directional movement of the subject approach driver. The difference-in-means test was used to determine if there was a significant difference between the mean value of the saturation headway as a function of the turning movement direction of the subject vehicle. Since there was a small number of left-turning vehicles in each case, only through and right-turning vehicles could be compared.

The difference-in-means test compares the mean and standard deviation for two samples, with the hypothesis that the two samples are drawn from the same population. The null hypothesis (that the saturation headways for the through and right-turning vehicles are from the same population) for Cases 3 and 4 can be rejected at a confidence level of 0.99. The null hypothesis can be rejected for Cases 1 and 2 at a 0.95 level. Thus it can be concluded that the directional movement of the subject vehicle has an effect on the saturation headway.

Table 5 presents the computed capacities using the saturation headways for the through and right-turning movements for each of the four cases. Separation of the saturation headways by turning movement results in considerably different capacity estimates, with capacity differences ranging from 27 to 50 percent between the through and the right-turning movement capacities. Since the capacity equation now includes only an additive factor to account for turning movements, some future adjustment clearly is required so that a more accurate estimate of approach capacity is available.

Consideration of Case Subsets

Another way of improving the AWSC intersection capacity procedure is to determine if the four cases can be divided into subsets that better reflect the conditions faced by the subject vehicle. For example, Case 3 states that the subject vehicle is faced by vehicles on the conflicting approach and not on the opposing approach. But this case can include one or two conflicting vehicles, one from the left and one from the right, or both.

Several subsets were considered for Cases 3 and 4 to determine if additional cases are justified. Table 6 presents these subsets. Table

TABLE 4 Effect of Turning Movement Direction on AWSC Intersection Saturation Headways

Case	Mean, sec			Standard Deviation			Observations			Test Statistic
	LT	TH	RT	LT	TH	RT	LT	TH	RT	
1	1.6	3.0	2.1	0.03	1.52	0.97	2	14	37	2.13
2	3.2	4.2	2.8	-	1.53	1.25	1	14	12	2.49
3	6.6	6.3	4.9	2.71	1.43	1.83	4	57	25	3.37
4	8.2	7.9	6.2	2.44	2.10	1.91	15	164	25	4.11

Note: The test statistic is computed using the difference in means test.

TABLE 5 Effect of Turning Movement on Approach Capacity of AWSC Intersection

Case	Headway, sec		Capacity, veh/hr		Percent Difference
	TH	RT	TH	RT	
1	3.0	2.1	1200	1714	+43
2	4.2	2.8	857	1286	+50
3	6.3	4.9	571	735	+29
4	7.9	6.2	456	581	+27

7 shows the saturation headways that were measured for each of the six subsets. Tables 8 and 9 give the difference-in-means test statistics that resulted in the comparisons between the subsets. Several conclusions can be made with respect to the data presented in these tables.

First, there is no statistically significant difference between Cases 3a (5.1 sec) and 3b (5.6 sec). That is, from the standpoint of the subject approach driver, it makes no difference if a conflicting vehicle approaches from the left or the right, as long as there is only one conflicting vehicle.

But there is a significant difference between Case 3a or 3b and Case 3c (6.8 sec). Thus if one conflicting vehicle is present on both the left and the right approaches, the saturation headway for the subject vehicle is different, in this case longer, than if the subject vehicle were faced by only one conflicting vehicle.

There are also some differences in the three Case 4 subsets. Similar to the results for Cases 3a and 3b, there does not appear to be a significant difference between the Case 4a and 4b subsets (6.8 and 6.9 sec). But there are differences between the subsets of Cases 4a and 4b and of Case 4c (8.4 sec). Thus, even though the direction of approach on the conflicting approach does not make a difference, the number of conflicting vehicles is significant.

Future versions of the capacity model for AWSC intersections should consider more than just the four cases now included. This paper has shown that at least two additional cases are warranted.

SUMMARY AND CONCLUSIONS

The results of a study of the saturation headway and related data for stop-controlled intersections have been presented. Data collected from one TWSC intersection and one AWSC intersection have been used to illustrate several important aspects about the saturation headway, and thus the capacity, of these two types of intersection.

For TWSC intersections,

- The theoretical definitions of the critical gap and the follow-up gap that underlie gap acceptance theory were described, and values for the two parameters were computed on the basis of data collected at the study site. The relationship between these two parameters was given.
- The importance of the directional movement of the major stream vehicle terminating a gap was illustrated for gaps that were rejected and accepted by minor stream left-turning vehicles. The

TABLE 6 Subsets for Saturation Headway Cases

Subset	Description
3a	One conflicting vehicle from the right
3b	One conflicting vehicle from the left
3c	One conflicting vehicle from both the left and the right
4a	One conflicting vehicle from the left and one opposing vehicle
4b	One conflicting vehicle from the right and one opposing vehicle
4c	One conflicting vehicle from both the left and right, and one opposing vehicle

TABLE 7 Saturation Headways for Subsets for AWSC Intersections

Subset	Mean Headway, sec	Standard Deviation	Observations
3a	5.1	1.60	31
3b	5.6	1.38	20
3c	6.8	1.67	36
4a	6.9	1.62	32
4b	6.8	1.57	62
4c	8.4	2.39	82

TABLE 8 Test Statistics for Case 3

	Case 3a	Case 3b	Case 3c
Case 3a	-	-1.115	-4.149
Case 3b	1.115	-	-2.864
Case 3c	4.149	2.864	-

TABLE 9 Test Statistics for Case 4

	Case 4a	Case 4b	Case 4c
Case 4a	-	-0.229	-3.85
Case 4b	0.229	-	-4.777
Case 4c	3.85	4.777	-

technique described here allows a quantification of the conflicting vehicle equations now given in Figure 10-3 of the HCM.

For AWSC intersections,

- The effect of the turning movement direction of the subject approach vehicle on the saturation headway was determined to be significant. This effect must be considered in future versions of the capacity equation.
- The classification of the four basic saturation headway cases for AWSC intersections into a new set of subsets was described, and a series of statistical tests were used to identify the new categories that could be justified. The effect on the approach capacity was illustrated.

Each of the factors should be considered in greater depth as the capacity procedures for stop-controlled intersections are modified and improved. The results described here may provide some guidance on some of the specific changes that should be considered.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. Brilon, W., M. Grossman, and B. Stuwe. Toward a New German Guideline for Capacity of Unsignalized Intersections. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991.
3. *Transportation Research Circular 373: Interim Materials on Capacity of Unsignalized Intersections.* TRB, National Research Council, Washington, D.C., 1991.
4. Kittelson, W., and M. Vandehey. The Effect of Delay on Gap Acceptance. Presented at 71st Annual Meeting of the Transportation Research Board, Washington, D.C., 1992.
5. Brilon, W., R. Troutbeck, and M. Tracz. *Review of International Practices Used To Evaluate Unsignalized Intersections.* TRB, National Research Council, Washington, D.C. (in preparation).
6. Boesen, A., B. Rindlisbacher, and M. Kyte. *Traffic Data Input Program*, Version 3.0. University of Idaho, Moscow, 1991.

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

Case Study Investigation of Traffic Circle Capacity

GEORGE LIST, SIEW LEONG, YUSRI EMBONG, AZIZAN NAIM, AND JENNIFER CONLEY

A capacity analysis of Latham Circle, a traffic circle in New York State, is presented. From videotapes, 1-min observations of entry flow are correlated with simultaneous observations of circulating flow. Values for the minimum acceptable gap, minimum circulating headway, move-up time, and so forth are calculated so that the predictions of various capacity equations, established abroad, can be compared and contrasted with the traffic circle's performance. It is found not only that the observed parameter values closely match those from abroad, but also that several of the equations appear to provide reasonable estimates of capacity. Because of this, these relationships may be adaptable to U.S. conditions without significant recalibration or reformulation.

Internationally, there has been a resurgence of interest in traffic circles. Germany (*1*), Switzerland (*2,3*), France (*4,5*), Australia (*6*), Norway (*7*), and Israel (*8*) are among the countries experimenting with their use. The United States, however, has had few recent instances in which these traffic control devices have been installed (*9*).

The idea of the traffic circle dates to about 1903, when Henard suggested the concept as a form of traffic control at busy junctions (*10*). The first traffic circle was Columbus Circle, constructed in New York in 1905 (*11*). Two years later two more were built in Paris at the Place de l'Etoile and Place de la Nation (*11*). The operating principle of these facilities is that conflicting vehicles merge, weave, and diverge, at a relatively uniform speed, as they circulate about a central island.

DEFINING TRAFFIC CIRCLE CAPACITY

When traffic circles were introduced, their operation was governed by the on-side priority rule, wherein circulating vehicles gave way to those entering. This practice was an extension of the prevalent operating rule, still in use today for uncontrolled intersections, wherein motorists on the left yield to those on their right (*12*).

The on-side priority rule produced an operating discipline similar to that for weaving sections. Wardrop, and others, found that capacity equations based on weaving principles could be derived (*13*). Generally, as volumes increased, wider and longer weaving sections were required, as was a bigger central island.

Eventually, however, the on-side priority rule led to traffic lockups. Vehicles in the traffic circle came to a standstill because they were blocked by downstream entryway flows. To prevent these conditions, the off-side priority rule was introduced. Vehicles on the entry legs were required to yield to those already in the circulating flow. This produced queues on the entry legs but kept the facility's operation from reaching a standstill.

Shifting to the off-side priority rule also led to changes in traffic circle operation, as one might expect. Entryway junctions behaved more like T-intersections, with the entering vehicles searching for and accepting gaps in the circulating flow. Analysts found it necessary to redefine capacity as the maximum entryway flow rate achievable for a given level of circulatory flow, as shown in Figure 1. From the work by Tanner and by Kimber (*14, 15*), capacity equations were developed that postulated an exponential relationship between the two flows. These relationships, and the linear geometric formulas developed by Kimber (*16*), are used here to analyze the performance of Latham Circle, a traffic circle.

As shown in Figure 2, if one defines the term "throughput" as the pairwise combination of circulatory and entryway flows for a given time period, these capacity equations predict the maximum combinations that are possible, establishing a trade-off surface between circulating and entryway flows. Hence, the issue for Latham Circle is as follows: do any of the capacity equations developed abroad provide a plausible upper bound on the throughput values observed?

SITE DESCRIPTION

Latham Circle is at the junction of New York State Route 2 [annual average daily traffic (AADT) = 19,430] and U.S. Highway 9 (AADT = 21,140) in Latham, New York. Built in 1949, it has been studied heavily before, in 1952 (*17*). Lying northeast of Albany, next to Latham Circle Mall, it is in the midst of a large commercial area that parallels Interstate 87. The facility serves both commuter and shopping trips.

Except for its diameter, the circle is designed much like a modern-day traffic circle (*1,6,18*). It has a two-lane circulating roadway and two-lane entrances and exits. There are four legs and an underpass, built in 1957, for through traffic in the north-south direction (on US-9). The central island has an inscribed diameter of 83 m, all the lanes are nominally 3.65 m wide, and the angle of deflection on entry is approximately 35 degrees. The traffic circle operates under the off-side priority rule since there are stop signs on the US-9 entrances and yield signs on the NY-2 entryways. A wide island separates the entrances and exits on US-9, and splitter islands are present on NY-2. Limited flaring exists on all entryways.

DATA COLLECTION

Data collection for the capacity analysis was two-pronged. First, general information about the site was obtained from New York

Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, Troy, N.Y. 12180.

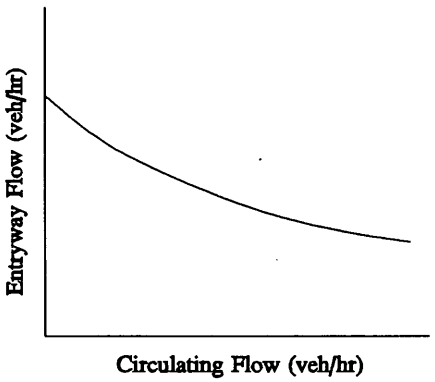


FIGURE 1 Capacity relationships for traffic circles.

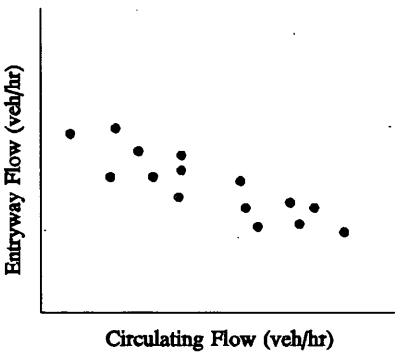


FIGURE 2 Facility throughput: combinations of entryway flow and circulating flow.

State Department of Transportation's Region 1 offices (e.g., plans for the circle, accident statistics, and AADTs). Then the circle's traffic flows were videotaped under peak-hour conditions, the first phase of recording occurring on March 7 and the second phase on April 23, 1992.

The three busiest entryways—eastbound, northbound, and southbound—were taped on March 7 for approximately 20 min each; the busiest of these (the northbound entrance) was taped on April 23 for another 1 hr. In the text that follows, these are called the Group 1 through 4 data sets: Group 1 is the data for the eastbound entrance on March 7; Group 2, the northbound entrance that same day; Group 3, the southbound entrance that same day; and Group 4, the northbound entrance on April 23. Even though the data for Groups 2 and 4 pertain to the same approach, they have been kept separate to enable informal testing for consistency in the observed behavior of a given site.

From the site plans, geometric information was extracted for later use in the capacity formulas. From the videotapes, 1-min observations of traffic flows were developed, as well as estimates of the average minimum critical gap, the minimum circulating headway, and the follow-up time. (There will be more discussion about these efforts later.)

FINDINGS

Of greatest interest is the capacity of the traffic circle. Figure 3 shows a plot of the 1-min entryway flow rates against their corresponding circulating flow rates for the four groups of data, creating a picture of the circle's throughput characteristics. The upper range of these throughput combinations reflects the capacity characteristics of the facility. Although an obvious trend is not apparent, it is clear that the peak-hour circulating flow ranges between 500 and 1,500 vehicles per hour (vph) and the entryway flow ranges between 400 and 1,000 vph. The maximum entering flow rate tends to decrease as the circulating flow rate increases, as would be predicted by the capacity equations developed abroad.

Differences among the approaches are also apparent. The northbound approach (Groups 2 and 4) has its observations clustered in the upper right-hand portion of the graph, which makes sense because that approach is the busiest. The data points for the eastbound approach (Group 1) primarily sit in the upper left-hand portion of the graph, reflecting the predominance of heavy entering traffic. Finally, most of the data points for the southbound entrance (Group 3) are in the bottom left-hand corner of the graph, which is logical since that location has the smallest flows on both the entryway and the circulating roadway.

As for gap parameters, the time headway between circulating vehicles, τ, appears to lie in the range of 0.7 to 1.2 sec, as shown in Table 1. To determine this, the videotapes were processed to identify uninterrupted sequences of circulating vehicles passing by each of the entryways. Of greatest interest were the more tightly grouped of these sequences, which represent "bunches" of traffic. [Troutbeck (*6*) further discusses bunching.)

To estimate τ, these data were sorted and summarized, as shown in Figure 4, producing cumulative density functions (CDFs) by group. It is clear that the CDFs for Groups 2 and 4 are nearly identical, as should be the case since they are for the same entryway. The CDF for Group 3 is more sloped and skewed to the right, which makes sense because the circulating traffic of the southbound approach is more dispersed with fewer bunches (*18*). The CDF for Group 1 is also skewed to the right, which should be the case since the eastbound entrance (Group 1) sees nearly the same dispersed circulating traffic flow as does the southbound approach (Group 3).

Figure 4 also reinforces the validity of using 0.7 to 1.2 sec for τ. This range encompasses the spread of 5th- and 10th-percentile values observed. It also matches with values observed abroad, as in the 1- to 2-sec value cited by Austroads (*18*), the 1.17 sec found by Armitage and McDonald (*19*), and the nominal 2-sec value (per lane) prescribed by Bennett (*20*).

The proportion of free vehicles in the circulating traffic stream, α ($0 \leq \alpha \leq 1$), appears to lie in the range of 0.4 to 0.7. As Table 2 indicates, the specific values by group are 0.62, 0.51, 0.73, and 0.38, respectively, based on the number of vehicles that are not in platoons with average headways of 2 sec or less. The table also gives comparable values for a 3-sec criterion and values derived from the tables provided by Austroads (*18*).

The average minimum acceptable gap or critical gap, t_a, appears to lie between 2.8 and 4.0 sec, as presented in Table 3. Similarly, the follow-up time, t_f, is between 1.8 and 3.7 sec. (These values are for the two entry lanes combined—90 percent or more of the traffic uses just the right-hand-most lane.) The methods of Armitage and McDonald (*19*) and of Siegloch (*21*) were used to estimate these parameters. Close matches exist not only among the values for the different groups but also between the values developed by the two methodologies. This suggests consistency in driver behavior, regardless of which approach is being used, and consistency between the two estimation methodologies.

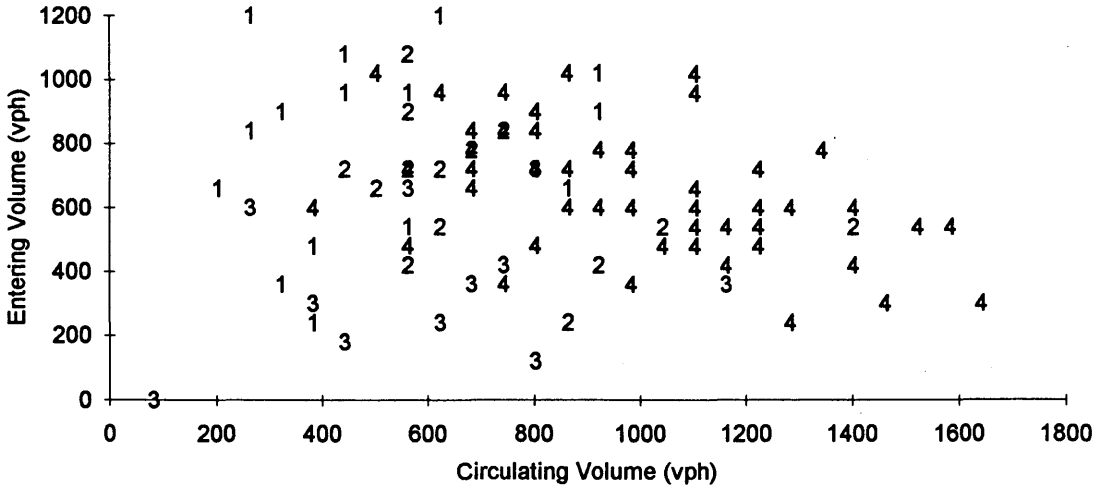


FIGURE 3 Throughput observations: peak-hour conditions, all groups.

TABLE 1 Gap Data for Bunches of Circulating Vehicles

Group	Number of Observations		Observed Gaps for Following Vehicles				
	Bunches	Following Vehicles	Minimum	5th Percentile	10th Percentile	Median	Mean
1	26	55	0.38	0.72	1.10	2.09	2.89
2	41	103	0.99	1.33	1.38	1.95	2.04
3	26	63	0.83	1.21	1.37	2.48	2.87
4	172	605	0.61	1.09	1.18	1.68	1.77

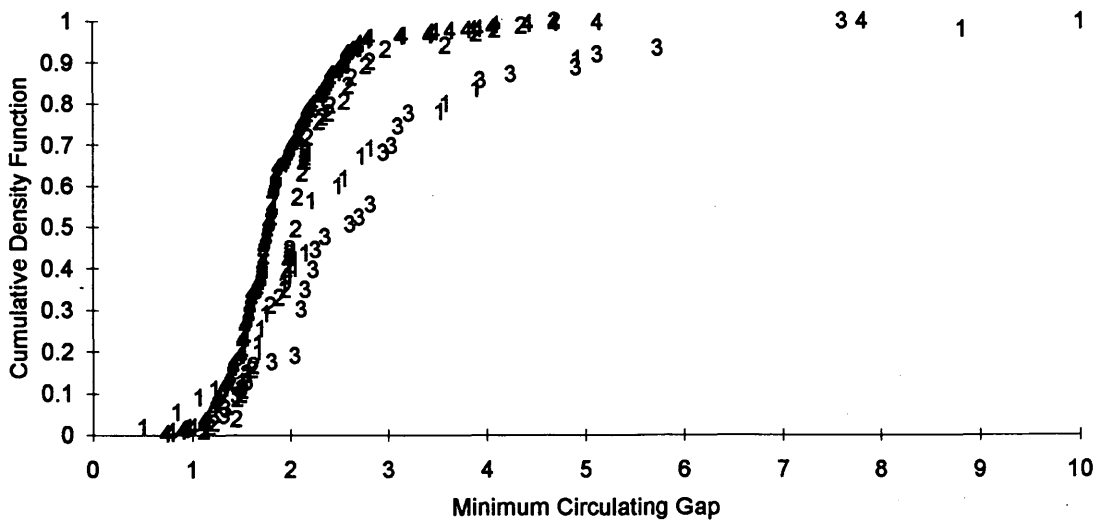


FIGURE 4 Circulating gap cumulative density functions.

TABLE 2 Percentage of Free Vehicles

Group	Observed Volumes		Estimated Values[1]		Suggested Values from (4) [2]	
	Q_c	Q_e	2-sec	3-sec	$N_c = 2$	$N_c = 1$
1	495	795	.72	.54	.64	.55
2	700	664	.51	.25	.58	.46
3	573	360	.73	.37	.63	.53
4	1000	634	.38	.20	.53	.34

Notes

[1] Based on the proportion of vehicles not found in platoons with an average headway of less than the value shown. For example, for Group 1, 72% of the vehicles were not in platoons with an average headway of 2 seconds or less.

[2] N_c: Number of circulating lanes (1 or 2)

Other, qualitative observations, critical in understanding the capacity characteristics of the traffic circle, are worth noting. First, even though the entryways are two lanes wide, nearly all (at least 90 percent) of the entering traffic uses the far right-hand lane. Second, despite the fact that the traffic circle is two lanes wide, vehicles tend to follow single file around the circulating roadway. No more than 1 car in every 20 occupies the inside lane alongside another vehicle. This is not to say that the inside lane goes unused, but that drivers treat it as a passing lane, to be used to advance one's own vehicle around another that is exiting when the probability of being cut off from one's own exit objective is small. Third, vehicles on the entryways tend to take into account all of the circulating flow, not just that in the outside lane, when checking for gaps to accept. Finally, the extent to which both lanes are used, either in the traffic circle or on the entryways, is very limited, a finding that is in keeping with experience abroad (*1,6*).

Before turning to the capacity analysis, two final notes seem appropriate. First, the data in Table 3 for the southbound approach (Group 3) should probably be omitted when estimating default values for t_a and t_f for other locations; that junction rarely operates at capacity. If this is done, the bounds on t_a tighten to 2.9 to 3.7 sec, and on t_f, 1.7 to 2.2 sec. The other note is that these values are for all vehicles on a given approach, and since the approaches are all nominally two lanes wide, a lane-by-lane analysis would most likely generate larger values.

TABLE 3 Minimum Acceptable Gap t_a and Follow-Up Gap t_f

Group	t_a	t_f	R^2	Total Absolute Difference[1]	Maximum Absolute Difference[2]
Based on Siegloch (22)					
1	3.45	1.81	.912	15	2
2	3.65	1.79	.891	7	1
3	4.06	3.29	.998	4	1
4	2.93	2.05	.919	39	2
Based on Armitage and McDonald (3)					
1	2.89	2.18	n/a	72	6
2	3.60	1.71	n/a	54	3
3	3.87	3.68	n/a	28	2
4	3.41	1.84	n/a	247	7

[1] Σ |projected number that could use gap - vehicles using gap|
[2] Maximum (|projected number that could use gap - vehicles using gap|)

CAPACITY ANALYSIS

To conduct the capacity analysis, the data plotted in Figure 3 were compared and contrasted with the predictions of various capacity equations developed abroad. By using the gap parameters described earlier, it was hoped that one or more of these equations might produce reasonable upper bounds for the throughput values depicted in the figure.

Eight equations were examined: three developed for prioritized junctions in general—Siegloch (*21*), Harders (*22*), and Jacobs (*23*)—and five that have been used in traffic circle situations—Troutbeck (*6*), Bennett (*20*), Stuwe (*24*), Brilon and Stuwe (*1*), and Kimber (*16*).

Of the first three, Harders' equation (*22*) is as follows:

$$C_e = Q_c \frac{e^{-q_c t_a}}{1 - e^{-q_c t_f}} \tag{1}$$

where

Q_c = flow rate of circulating stream (vph),
q_c = flow rate of circulating stream (vph/sec), and
C_e = capacity (maximum possible flow rate) for entering stream (vph) given Q_c,
t_a = critical gap for entering drivers (sec), and
t_f = follow-on time for entering drivers (sec).

Siegloch's equation (*21*) is as follows:

$$C_e = \frac{3{,}600}{t_f} e^{-q_c t_o} \tag{2}$$

where t_o is given by

$$t_o = t_a - \frac{t_f}{2} \tag{3}$$

Finally, Jacob's equation (*23*) is

$$C_e = \frac{\alpha Q_c e^{-\lambda(t_a - \tau)}}{\lambda t_f} \tag{4}$$

where α is the proportion of free vehicles in the circulating traffic stream and λ is defined as

$$\lambda = \frac{\alpha q_c}{1 - \tau q_c} \tag{5}$$

From the equations that have been applied to traffic circles, five were explored. The first is Troutbeck's (*6*):

$$C_e = \frac{\alpha Q_c e^{-\lambda(t_a - \tau)}}{1 - e^{-\lambda t_f}} \tag{6}$$

and the second is Bennett's (*20*):

$$C_e = \frac{\alpha Q_c e^{-\lambda(t_a - \tau)}}{1 - e^{-q_c t_f}} \tag{7}$$

The third is the regression equation developed by Stuwe (*24,1*):

$$C_e = A e^{-BQ_c/10{,}000} \tag{8}$$

where $A = 1{,}577$ and $B = 6.61$ on the basis of observations at 4,574 traffic circles with two entry lanes and two circulating lanes in Germany. The fourth model is an alternative regression equation developed by Brilon et al. (*25*):

$$C_e = A e^{-BQ_c/10{,}000} + DN_c + EN_e \tag{9}$$

where

$A = 1{,}549$,
$B = 8.4$,
$D = 208.4$, and
$E = 48.02$.

The fifth and final model in this group is Kimber's (*16*), which predicts traffic circle capacity on the basis of geometric parameters:

$$C_e = K(F - f_c Q_c) \tag{10}$$

where

$$K = 1 - 0.00347(\Psi - 30) - 0.978\left(\frac{1}{r} - 0.05\right) \tag{11}$$

$$F = 303 x_2 \tag{12}$$

$$f_c = 0.21 t_d (1 + 0.2 x_2) \tag{13}$$

$$t_d = 1 + \frac{0.5}{1 + e^{\frac{D-60}{10}}} \tag{14}$$

$$x_2 = v + \frac{e - v}{1 + 2s} \tag{15}$$

and

$$s = \frac{e - v}{l} \tag{16}$$

where

e = entry width (m),
v = approach half-width (m), and
l = average effective length over which flare is developed (m).

These parameters are also defined in Figure 5; D and r must be in meters, and Ψ must be in degrees. (Note that the e used in Equations 15 and 16 is a parameter whereas the e used in Equation 14 is the natural logarithm base.) For Latham Circle, the values that pertain are $\Psi = 35$ degrees, $D = 82.9$ m, $v = 7.32$ m, $r = 18.59$ m, $e = 8.36$ m, and $l = 15.68$ m. This yields constants of $K = 0.9789$, $t_d = 1.046$, $s = 0.0663$, $x_2 = 8.238$, $f_c = 0.5816$, and $F_c = 2496.2$.

Figure 6 shows a plot of the first three models (Harder, Siegloch, and Jacob) against the throughput values for the northbound approach, the busiest of the three (Group 4). (The values used for α, τ, t_a, and t_f are from Tables 1 through 3. For α, the 10th-percentile values were used; for τ, the 2-sec values; and for t_a and t_f, the Armitage- and McDonald-based values.) Although these models might provide plausible upper bounds for the junction's potential throughput, since the observations displayed are all 1-min values, it is not likely that the values purported by the models can be achieved on a sustained basis. One must remember that these equations should reflect the sustained 1-hr capacity of the facility, not its 1-hr capacity. Hence, equations that predict the true 1-hr capacity should lie in the upper reaches of but not above the 1-min throughput values observed. [Hakkert et al. discuss this point (*8*).]

Figure 7 shows the capacity relationships estimated by four of the traffic circle–based capacity models. Whereas Bennett's model

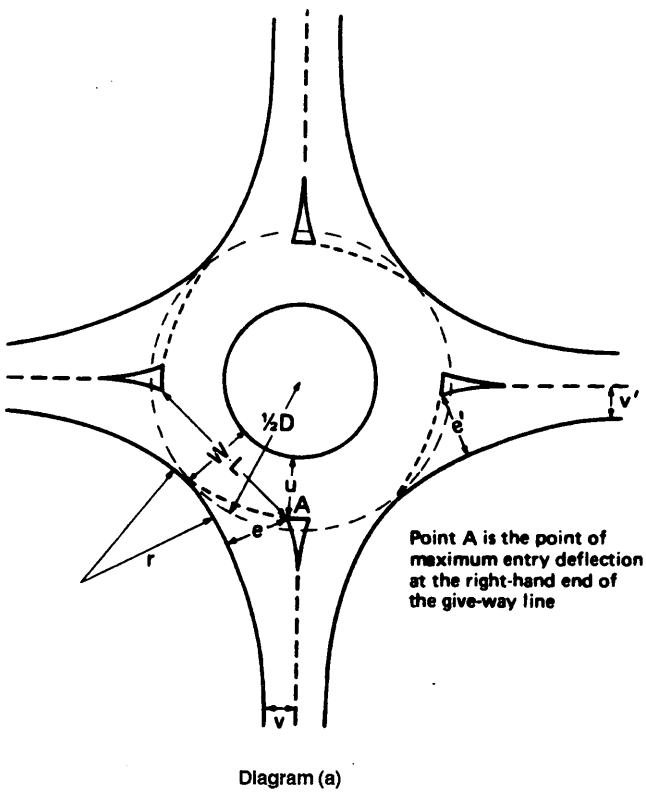


Diagram (a)

(i) The *entry width*, e, is measured from the point A along the normal to the nearside kerb, see Diagram (a).

(ii) The *approach half-width*, v, is measured at a point in the approach upstream from any entry flare, from the median line to the nearside kerb, along a normal, see Diagram (a).

(iii) The *entry width*, e', and *approach half-width*, v', for the *previous entry* are measured in the same way as e and v, see Diagram (a).

(iv) The *circulation width*, u, is measured as the shortest distance between point A and the central island, see Diagram (a).

(v) Two alternative constructions can be used to obtain the *average effective length over which the flare is developed.* The first (ℓ) is as used previously (see reference 3), and is shown in Diagram (b).

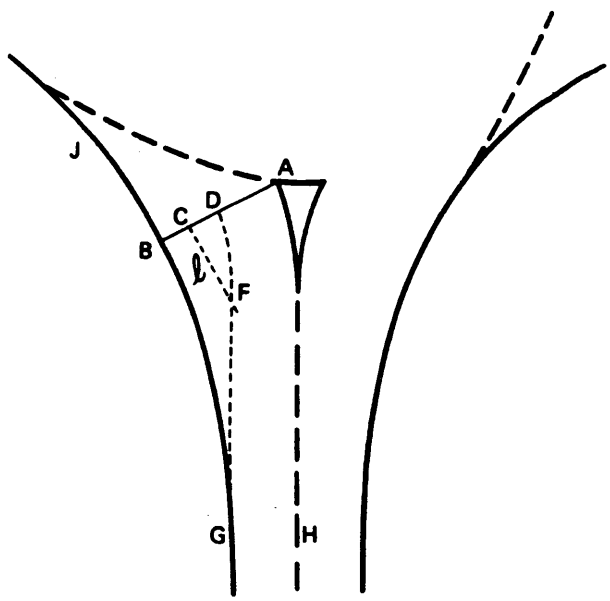


Diagram (b)

FIGURE 5 Kimber's model and parameter definitions.

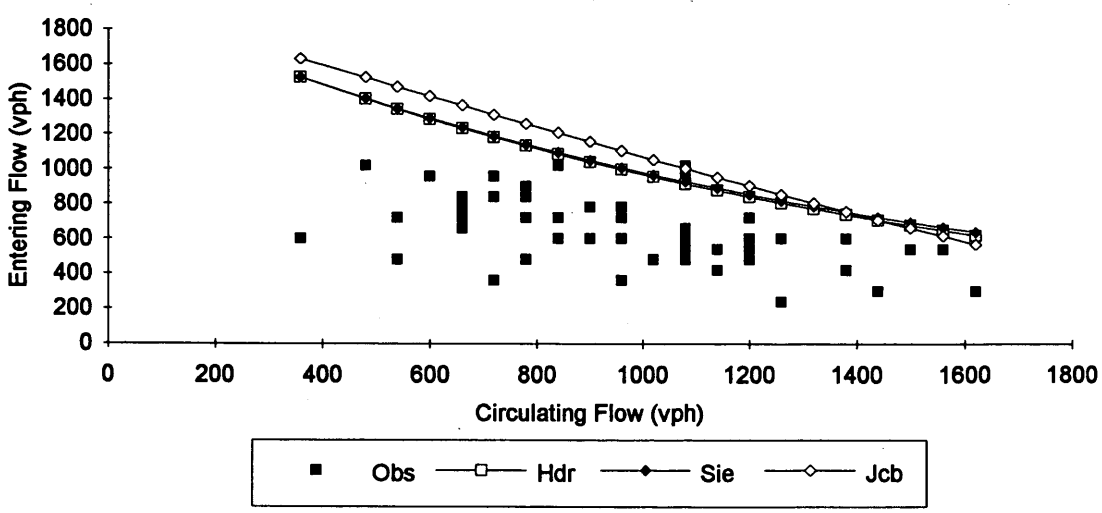


FIGURE 6 Three priority junction models: Group 4, northbound approach.

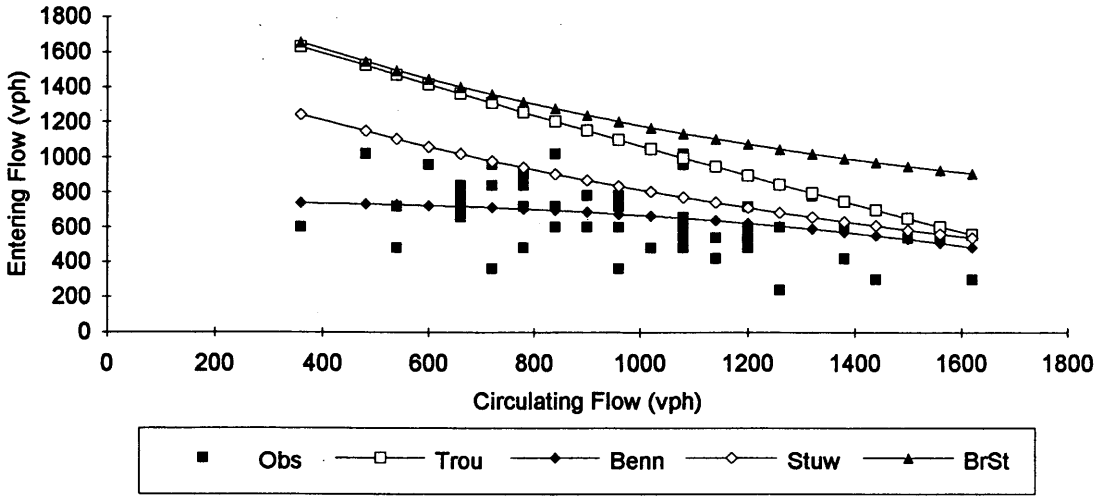


FIGURE 7 Four traffic circle models: Group 4, northbound approach.

appears to underestimate the potential maximum throughput, the models of Troutbeck and of Brilon and Stuwe probably overestimate it. Stuwe's model appears the most likely to provide reasonable values for this particular site given the input parameters employed.

Looking at all of the approaches simultaneously, Figure 8 provides a plot of the predictions of Troutbeck's model for all four data groups. One can see the effects of changes in the parameter values for the capacity predictions. In contrast, if the predictions of Stuwe's model were displayed, there would be no differences in capacity prediction since that model depends on only one parameter: the circulating volume.

Kimber's model (*15*) generates capacity estimates significantly in excess of the throughput values observed. This suggests that the familiarity of British motorists with traffic circles dramatically increases the achievable throughput.

Table 4 gives the predicted values of C_e for each of the eight models for Data Groups 1 through 4. It is clear that all of the models estimate Q_e values higher than those observed. But that is to be expected since, as noted earlier, none of these approaches, except the northbound one (Groups 2 and 4), appeared to be at or near capacity in the field.

ACCIDENT TRENDS

An examination of the accident trends of Latham Circle is important because traffic circles are generally considered by U.S. motorists to be hazardous locations. At Latham Circle, between 1989 and 1991 there were 169 accidents, broken down as given in Table 5. None of the accidents involved fatalities, 37 involved injuries, 54 had property damage only, and the remaining 88 had consequences lower than $600, the minimum reportable threshold. Given the AADTs for US-9 (22,900 south of the traffic circle and 19,380 north) and NY-2 (20,960 west of the traffic circle and 17,900 east), the average annual daily number of vehicles entering the traf-

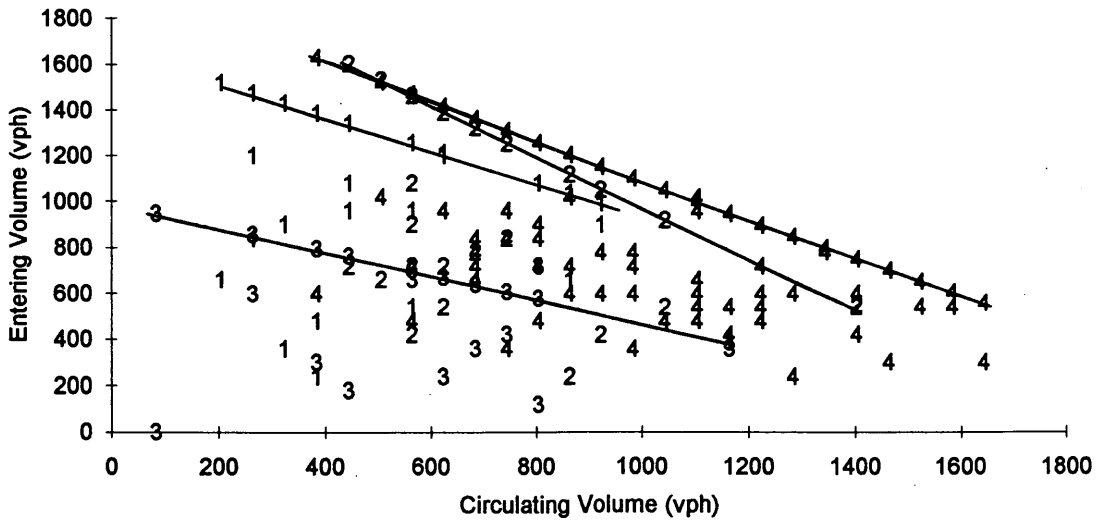


FIGURE 8 Troutbeck model for all four groups.

TABLE 4 Comparison of Capacity Predictions

Group	Observed		Predicted Maximum Entryway Volume							
	Q_e	C_e	Harder	Sieg-loch	Jacob	Trout-beck	Bennett	Stuwe	Brilon-Stuwe	Kimber
1	495	795	1284	1289	1292	1288	1117	1137	1535	2162
2	700	664	1229	1235	1280	1277	934	993	1373	2045
3	573	360	698	708	694	685	651	1080	1470	2117
4	1000	634	969	979	1070	1067	669	814	1181	1874

fic circle is approximately 40,570. This implies an accident rate of 3.8 accidents per million vehicles entering. Although this value exceeds the statewide overall average of 0.88 accidents per million vehicles entering, it is within the range of values reported by Brilon et al. (*25*) for German conditions: 6.58 for medium and large traffic circles, 1.24 for smaller traffic circles, 3.35 for junctions with traffic signals, and 1.00 for junctions without traffic signals.

Consideration of the accidents by type shows that none was fatal and only 21.9 percent involved injuries. By comparison, statewide, 0.2 percent of all accidents are fatal and 37.6 percent involve injuries. The difference from the statewide average in the percentage of fatal accidents may not be significant, but the injurious accident percentage difference might be. Many researchers abroad have found that modern traffic circles tend to produce significantly fewer injury accidents than the average intersection (*4,15,17,25–28*).

Of the accidents, 24 percent were rear-end collisions, predominantly on the ramps, and 11 percent were right-angle collisions, half on the traffic circle and half on the ramps. Nine percent were overtakes and the rest were sideswipes (3 percent), collisions with fixed objects, especially guardrails and curbs (2 percent), and nonreportable accidents (51 percent) for which no event description is provided.

DISCUSSION OF RESULTS

Unfortunately, nowhere in the preceding analysis does the subject of level of service appear. No one abroad has explicitly considered the kind of delay-based performance assessment that has become so much a part of U.S. capacity analyses (*29*). It is uncommon to find a paper, such as that by Troutbeck (*6*), in which the relationship between average delay per vehicle and degree of saturation is addressed. It is important to know how average delay varies with volume-capacity ratios—for example, to understand where traffic circles might be applicable, one should know, among other features, where delay performance is the best among all control devices available.

Other attributes are also important. One should know where the breakpoint is between these facilities and others, such as signalized and unsignalized intersections. One should also learn the characteristics of the delay curves of a traffic circle. Certainly, the capacity equations discussed earlier imply that the delay per vehicle at any given value of entering volume increases as the circulating volume increases. Logic suggests that this is probably true, but there are no empirical data to support this conclusion.

SUMMARY AND CONCLUSIONS

This paper has presented a capacity analysis of Latham Circle, located in Latham, New York, outside of Albany. From videotapes, 1-min observations of entry flow were correlated with simultaneous observations of circulating flow. Values for the average minimum acceptable gap, minimum circulating headway, move-up time, and geometric parameters were calculated so that various capacity equations, established abroad, could be compared and contrasted with the circle's performance. Prioritized junction capacity equations proposed by Harders (*22*), Siegloch (*21*), and Jacob (see *23*) and traffic circle capacity equations proposed by Bennett (*20*), Troutbeck (*6*), Stuwe (*24*), and Brilon et al. (*25*) were all found to provide plausible upper bounds on the circle's observed throughput. Moreover, all but the last of these appear to provide capacity esti-

TABLE 5 Accident Statistics for Latham Circle

Year	Accidents by Category				TOT	AR
	FTL	INJ	PDO	N/R		
1989	0	9	18	47	74	5.00
1990	0	15	19	23	57	3.85
1991	0	9	12	17	38	2.57
TOTAL	0	33	49	87	169	3.80

Facility.....AADT

SR-9 North of Circle.....19,380
SR-9 South of Circle.....22,900
SR-2 West of Circle.....20,960
SR-2 East of Circle.....17,900

Key

FTL: Fatal Accident
INJ: Injurious Accident
PDO: Property Damage Only Accident
N/R: Non-reportable accident (<$600 damage)
TOT: Total accidents
AADT: Average Annual Daily Traffic
SR-9: State Route 9
SR-2: State Route 2

mates that match quite closely the greatest of the throughput values observed. Therefore, it appears plausible that one or more of these capacity relationships may be adaptable to U.S. conditions without significant recalibration.

The capacity models developed for British and Australian conditions, however, appear to overestimate the maximum throughput levels achievable. This indicates either that maximum throughput conditions were not observed, which seems unlikely, or that the experience of British and Australian drivers with traffic circles allows them to achieve higher throughput values than those currently possible here, at least for this particular circle. Hence, these equations might have to be recalibrated before being used in the United States.

Finally, a close correspondence was found between the observed gap parameters and those found to be typical abroad for these equations—for example, values for t_a, t_m, and t_f suggested by Bennett (*20*), Austroads (*18*), and Armitage and McDonald (*19*). This lends further evidence to the fact that similarities in driver behavior may exist.

The conclusion drawn from this analysis is that it appears possible to transfer capacity equations from abroad to the United States. Not only are calibration constants similar in some instances, but maximum levels of throughput also seem to correspond with the capacity predictions of models developed abroad.

It is important to note, however, that the study of traffic circles abroad has not included a focus on their delay characteristics, especially the way in which delay varies with volume-capacity ratios. The perspective of level of service has not been employed.

Hence, the major task ahead appears to be one of developing capacity relationships for domestic conditions that build on the equations already developed overseas. In fact, such an initiative should be seen as an opportunity to increase the international commonality in the treatment of capacity and level-of-service issues pertaining to such facilities.

REFERENCES

1. Brilon, W., and B. Stuwe. Capacity and Design of Roundabouts in Germany. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993, pp. 61–67.
2. Simon, M. J. Roundabouts in Switzerland. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 41–52.
3. Tan, J. A Microscopic Simulation Model of Roundabout Entry Operations. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 159–176.
4. Alphand, F., U. Noelle, and B. Guichet. Roundabouts and Road Safety: State of the Art in France. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 107–125.
5. Alphand, F., U. Noelle, and B. Guichet. Evolution of Design Rules for Urban Roundabouts in France. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 126–140.
6. Troutbeck, R. J. The Capacity and Design of Roundabouts in Australia. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993, pp. 68–74.
7. Seim, K. Use, Design and Safety of Small Roundabouts in Norway. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 270–296.
8. Hakkert, A. S., D. Mahalel, and S. A. Asante. A Comparative Study of Roundabout Capacity Procedures. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 93–106.
9. Ourston, L. British Interchanges, Intersections, and Traffic Control Devices. *Western ITE*, Vol. 35, No. 5, 1992, pp. 1–7.
10. Henard, E. *Etudes sur les Transformations de Paris, Fascicule 7: Carrefours a giration* (in French). Libraries-Imprimeries Reunis, Paris, France, 1906, pp. 283–302.
11. Todd, K. A History of Roundabouts in the United States and France. *Transportation Quarterly*, Vol. 42, No. 4, 1988, pp. 599–623.
12. *Driver's Manual*. New York State Department of Motor Vehicles, Albany, 1991.
13. Wardrop, J. G. The Capacity of Weaving Sections of Roundabouts. *Proc., 1st International Conference on Operations Research*, 1957.
14. Tanner, J. C. A Theoretical Analysis of Delay at an Uncontrolled Intersection. *Biometrica*, Vol. 49, 1962, pp. 163–170.
15. Kimber, R. M. Gap-Acceptance and Empiricism in Capacity Prediction. *Transportation Science*, Vol. 23, No. 2, 1989, pp. 100–111.
16. Kimber, R. M. *The Capacity of Roundabouts*. Laboratory Report 942. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1980.
17. Shrope, E. B. Testing a Traffic Circle for Possible Capacity. *Proc., Highway Research Board*, HRB, National Research Council, Washington, D.C., 1952, pp. 415–424.
18. *Guide to Traffic Engineering Practice, Part 6: Roundabouts*. Austroads, Sydney, Australia, 1993.
19. Armitage, D. J., and M. McDonald. Roundabout Capacity. *Traffic Engineering and Control*, Vol. 15, No. 18, 1974, pp. 812–815.
20. Bennett, R. F. The Design of Roundabouts Since the Priority Rule. *Journal of the Institute of Highway Engineers*, Vol. 18, No. 9, 1971, pp. 13–23.
21. Siegloch, W. Capacity Calculations for Unsignalized Intersections. *Schriftenreihe Strassenbau und Strassenverkehrstechnik 154*, 1973.
22. Harders, J. The Capacity of Unsignalized Urban Intersections. *Strassenbau und Strassenverkehrstechnik 76*, 1968.
23. Brilon, W. Recent Developments in Calculation Methods for Unsignalized Intersections in West Germany. In *Intersections Without Traffic Signals* (W. Brilon, ed.), Springer Publications, Berlin, Germany, 1988.
24. Stuwe, B. *Untersuchung der Leistungsfahigkeit und Verkehrssicherheit an deustschen Kreisverkehrsplatzen* (Investigation of Capacity and Safety at German Roundabouts). Ph.D. dissertation. Lehrstuhl fur Verkehrswesen, Ruhr-Universitat Bochum, Germany, 1992.
25. Brilon, W., B. Stuwe, et al. *Einsatzmoglichkeiten von Kreisverkehrsplatzen und aufgeweiteten Knotenpunkten unter besdonderer Berucksichtigung auslandischer Erfahrungen. Teil A: Kreisverkehrsplatze* (Possibilities for Using Roundabouts and Widened Intersections with Special Consideration of Foreign Experiences). Forschungsbericht FE 77198, 87 BMV, No. 7, 1990.
26. Smith, M. J. *Improved Signing for Traffic Circles: Final Report*. Report FHWA/NJ-91-003. FHWA, U.S. Department of Transportation, 1990.
27. Hoglund, P. G. Case Study: Performance Effects of Changing a Traffic Signal Intersection to Roundabout. In *Intersections Without Traffic Signals II* (W. Brilon, ed.), Springer-Verlag, New York, 1991, pp. 141–158.
28. Lalani, N. The Impact on Accidents of the Introduction of Mini, Small and Large Roundabouts at Major/Minor Priority Junctions. *Traffic Engineering and Control*, Vol. 16, No. 1, 1975, pp. 560–561.
29. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.

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

Estimating Freeway Origin-Destination Patterns Using Automation Traffic Counts

PING YU AND GARY A. DAVIS

To enable the efficient use of existing roadway capacity, researchers and practitioners are developing advanced traffic management systems (ATMS), which has led to an increased interest in problems connected to the estimation of origin-destination (O-D) flows using information provided by freeway surveillance and control systems. A number of methods based on a linear traffic assignment model have been applied successfully to single intersections, and some of these estimators were extended to a section of freeway. The results from Monte Carlo simulation suggest that ordinary least squares (OLS) and expectation-maximization approaches were either biased or inefficient. A nonlinear least squares (NLS) estimator that eliminated model specification error was introduced, and it performed better in terms of statistical efficiency and lack of bias. This implies that accurate O-D estimation may require an accurate traffic flow model and that actual implementation may require joint estimation of O-D patterns and traffic flow model parameters. On the other hand, a constrained approximate maximum likelihood estimator performed better than OLS but somewhat worse than NLS, showing some potential for providing a simple and yet plausibly accurate approach.

Traffic congestion is an increasingly serious problem for many of the world's urban areas, but fiscal, social, and environmental constraints prohibit large increases of highway capacity. Thus the advanced traffic management systems (ATMS) and advanced driver information systems (ADIS) initiatives in the United States, and similar programs in other nations, have as one of their major objectives the efficient use of existing highway capacity. This is to be achieved by an increased availability of high-quality real-time information about traffic conditions, along with a more intimate linking of traffic control with travel demand management tactics. The success of such an approach will depend heavily on the availability of practical models describing the interaction between travel demand and traffic flow phenomena, models that can give real-time predictions of the effects of proposed traffic management actions. Most traffic models use some form of an origin-destination (O-D) matrix as the basic description of the demand for travel, which has led to an interest in using the data collected by traffic surveillance systems, especially traffic counts, to generate real-time estimates of O-D matrices.

In particular, it is hoped that the availability of time-series data of traffic counts will permit development of O-D estimators that have desirable statistical properties, such as consistency, efficiency, and lack of bias, and that will be able to track changes in the O-D patterns. For general networks, constructing such O-D estimators can be a difficult task, because of the possibility that many routes may connect any given O-D pair [Davis (*1*)], but the problem is simplified somewhat when one considers simple "linear" networks, such as single intersections, transit routes, and freeway segments, where each origin and destination are connected by at most one route (2).

P. Yu, Bucher, Willis and Ratliff, 7920 Ward Parkway, Suite 100, Kansas City, Mo. 64114. G. A. Davis, Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive, S.E., Minneapolis, Minn. 55455.

Since urban freeways carry a large fraction of total urban travel, it is not surprising that estimation of freeway O-D patterns has been receiving increased attention (*3–6*); one has available time-series data of on-ramp, off-ramp, and mainline traffic counts, and can infer the O-D pattern that generated them. The freeway O-D problem is similar to that of estimating turning movement volumes from entering and exiting counts at single intersections, a problem that has been treated extensively during the past decade (*7–11*). Particularly relevant here is the paper by Nihan and Davis (*10*) that described a Monte Carlo study comparing several variants of ordinary least squares (OLS) estimators of turning movement proportions. Nihan and Davis found that although the OLS-based estimators tended to be consistent and unbiased, data from 50 to 60 time points were needed before the standard error of estimate could be reduced to a usefully low level. This finding indicated a possible bound on the ability of time-varying implementations of OLS to track within-day changes in the O-D pattern, so that, even if a recursive estimator is consistent (i.e., converges eventually to the true values of the unknown parameters), when the rate at which its standard error of estimate goes to zero is slow compared with the time variation of the underlying parameter, the estimated values are not likely to be close to the (unknown) true values. This in turn suggests that a naive embrace of recursive estimation procedures without due consideration of their convergence properties is as likely to inject error into travel demand modeling as it is to inject truth.

When applied to freeway segments, the simple linear traffic assignment model takes the form

$$\hat{y}_j(t) = \Sigma_i\, b_{ij}\, q_i(t) \tag{1}$$

where

$\hat{y}_j(t)$ = predicted traffic count at off-ramp j during time interval t, $j = 1, \ldots, n$;
$q_i(t)$ = actual traffic count at on-ramp i during time interval t, $i = 1, \ldots, m$; and
b_{ij} = probability that a vehicle entering at i is destined for j.

Traffic conservation considerations require that

$$0 \le b_{ij} \le 1 \qquad i = 1, \ldots, m,\ j = 1, \ldots, n \tag{2a}$$

$$\Sigma_j\, b_{ij} = 1.0 \qquad i = 1, \ldots, m \tag{2b}$$

Generally, on-ramp counts $q_i(t)$ and off-ramp counts $y_j(t)$ will be available from a freeway's surveillance system, and unconstrained OLS estimates of the unknown b_{ij} can be computed by minimizing the sum of squares function

$$S = \Sigma_t \Sigma_j\, [y_j(t) - \hat{y}_j(t)]^2 \tag{3}$$

while constrained OLS would minimize Equation 3 subject to Equations 2a and 2b. Unfortunately, even though unconstrained OLS and its variants give plausible estimates when applied to simple intersection counts, they tend to fail when applied to counts obtained from freeway on- and off-ramps. Table 1 presents unconstrained OLS estimates for a short section of Interstate I-35W with four on-ramps and two off-ramps, where by convention the upstream mainline boundary is denoted as On-Ramp 1 while the downstream mainline boundary is denoted as Off-Ramp 2. The estimated proportions given in Table 1 were obtained by minimizing Equation 3 using 5-min on- and off-ramp counts. Since this section was about 1 mi (1.7 km) long, most of the vehicles entering during a 5-min interval will have exited during that same interval, and one would expect that time-varying travel times would not be a factor.

Clearly, these estimates show serious violations of the conservation conditions in Equations 2a and 2b, and although it would still be possible to minimize Equation 3 subject to Equations 2a and 2b, a usefully consistent or unbiased estimator should be able to produce reasonably close estimates without such devices. Thus it appears that when applied to freeway data, OLS estimators can lose the consistency and unbiasedness properties shown when applied to single intersections, and it has been the authors' experience that such results are the norm rather than the exception when using OLS and the linear traffic assignment model to estimate freeway O-D proportions. This situation is unfortunate because from a practical standpoint, recursive versions of OLS are very easy to implement and tend to be computationally fast (*10*).

This discussion has identified two basic statistical issues with regard to freeway O-D estimators. The first concerns whether an estimator is unbiased or consistent, that is, whether on the average or in the long run the estimated O-D parameters will equal the true underlying values. A primary cause of bias or inconsistency is model specification error, in which the model that is assumed to generate the data differs substantially from the process actually generating the data. The linear traffic assignment model just described neglects the fact that the travel time between O-D pairs will differ both as a function of the distance separating the origin from the destination and as a function of the intervening traffic conditions. Such specification error may be responsible for the apparently biased estimates generated by OLS. But even if two estimators are unbiased, they may differ in efficiency, measured by the standard errors of the estimates as functions of sample size. The estimator with the lower standard error of estimate is more likely to generate estimates that are "close" to the true values when finite data sets are used. For example, in a linear regression model with heteroscedastic, normally distributed errors, simple OLS remains an unbiased estimator of the regression coefficients but is no longer efficient, the corresponding maximum likelihood estimator having smaller standard error.

This paper describes a Monte Carlo evaluation of four different approaches to estimating freeway O-D proportions b_{ij}, the objective being to decide which of the methods, under practically useful conditions, tend to be unbiased and to assess their relative statistical efficiency. Attention is restricted to off-line estimates of time-invariant parameters because the algorithms used to track time-varying O-D patterns are, for the most part, simply recursive versions of their off-line counterparts (*12*). For instance, the extended Kalman filter approach described by Chang and Wu (*5*) can be viewed as a recursive implementation of a nonlinear weighted least-squares approach, whereas the Kalman filter method tested by Ashok and Ben-Akiva (*6*) is a recursive implementation of a linear, multilag least-squares approach. A biased or inefficient estimator will not lose these properties when implemented recursively, but a good off-line estimator is a good candidate for recursive implementation. In particular, there is a natural connection between the efficiency of an off-line estimator and the convergence rate of its recursive counterpart, in that the standard error of estimate for the off-line estimator obtained with a sample of size N is a lower bound for the standard error of the recursive estimator after N iterations.

TABLE 1 OLS Estimates for Typical Freeway Data

	Off-Ramp	
On-Ramp	1	2
1	0.0	0.79
2	−0.10	2.22
3	0.34	2.43
4	0.0	1.35

Of the four candidate estimators considered here, three are based on the simple linear assignment model—and hence are subject to specification error—but differ as to the optimization criterion used to compute the estimates. The fourth minimizes the same least-squares criterion used by OLS, but the predicted off-ramp volumes are computed by a nonlinear model that eliminates specification error, which is possible because simulated data are being used. The objective is to determine if the computational simplicity of the linear model can be retained by shifting to a different optimization criterion or whether its inherent specification error is so serious as to make it unusable.

The authors first describe the simulation model used to generate the Monte Carlo sample, then describe the four estimation procedures. Results of the estimators' performance on the simulated data are presented next, and the paper ends with a discussion of these results.

STOCHASTIC FREEWAY TRAFFIC SIMULATION MODEL

As noted earlier, the objective of this study is to assess the statistical properties of several candidate procedures for estimating freeway O-D parameters. The primary method of assessment is Monte Carlo simulation, in which a sample of simulated freeway on-ramp and off-ramp counts is generated, and then each candidate estimation procedure used to compute estimates from each simulated data set. This produces a pseudorandom sample of estimates for each procedure, and these samples are used to determine the presence or absence of desirable statistical properties. To produce simulated data that preserve both the random assignment of vehicles to off-ramps and the general features of traffic flow, the authors developed the STOMAC (stochastic macroscopic) simulation model, which is described in the following.

Before the model is introduced, it is necessary to clarify the following notation and terms. N time intervals (e.g., 5 min each) are assumed during the period of interest; let $t = 1, \ldots, N$ index these intervals. Each of the N intervals is in turn divided into T subintervals, each of duration Δ. Let these subintervals be indexed by $\tau = 1, \ldots, NT$. The intervals represent the level of aggregation at which count data is available, and the subintervals are the basic time unit of the simulation model.

Figure 1 shows a section of freeway with m on-ramps (including the upstream boundary of the section of freeway) and n off-ramps

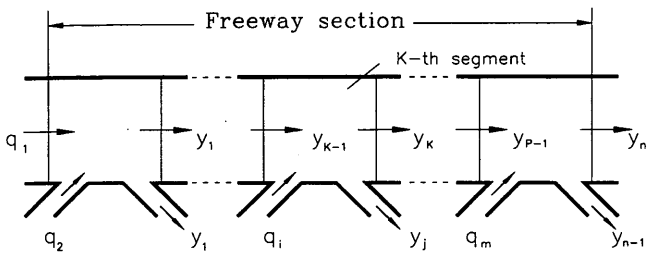


FIGURE 1 Freeway section with *m* on-ramps and *n* off-ramps, divided into *p* segments.

(including the downstream boundary), listed as $i = 1, \ldots, m$ and $j = 1, \ldots, n$, respectively. The upstream boundary will be treated as the first on-ramp and the downstream boundary will be the last off-ramp. The section of freeway has been divided into p segments, indexed by $k = 1, \ldots, p$, such that on-ramps are located only at the upstream boundary of a segment and off-ramps leave only at the downstream boundaries of segments. A further division of segments may be necessary to ensure that geometric features are constant within the segments.

The notation is defined as follows:

$q_i(t)$ = traffic entering at on-ramp i during time interval t
$\mathbf{q}(t)$ = m-dimensional vector whose elements are $q_i(t)$
$\hat{y}_j(t)$ = traffic exiting at off-ramp j during time interval t
$y_j(t)$ = forecast of traffic exiting at off-ramp j, during t
$x_{ij}(t)$ = traffic entering on-ramp i and destined for off-ramp j, during time interval t
b_{ij} = probability that a vehicle entering from on-ramp i is destined for off-ramp j,
$\mathbf{B}$ = $m \times n$ dimensional matrix whose elements are b_{ij}
$\mathbf{B}^T$ = transpose of $\mathbf{B}$
$\mathbf{b}_i$ = m-dimensional vector containing $b_{ij}, j = 1, \ldots, n$
$\mathbf{y}(t)$ = n-dimensional vector containing $y_j(t), j = 1, \ldots, n$
$\mathbf{V}(t)$ = $n \times n$ covariance matrix of $\mathbf{y}(t)$, given $\mathbf{q}(t)$
$P_k(\tau)$ = probability that a vehicle in segment k exits during subinterval τ
$z_{kj}(\tau)$ = number of vehicles in segment k destined for j at beginning of subinterval τ
$z_k(\tau)$ = $\Sigma_j z_{kj}(\tau)$
$q_{ij}(\tau)$ = number of vehicles entering at on-ramp i and destined for j during subinterval τ
$q_i(\tau)$ = $\Sigma_j q_{ij}(\tau)$
$y_{kj}(\tau)$ = number of vehicles exiting from segment k and heading for j during subinterval τ
$y_k(\tau)$ = $\Sigma_j y_{kj}(\tau)$
L_k = length of segment k
M_k = number of lanes in segment k
$r_k(\tau)$ = traffic density in segment $k = z_k(\tau)/(L_k * M_k)$
$\overline{U}_e(r)$ = equilibrium speed and density function

The basic idea was to treat traffic flow on a freeway as the outcome of a type of stochastic process known as a Markov compartment process (*13*). In this model, each segment of the section of the freeway was treated as a Markovian compartment, from which vehicles exit with probability $P_k(\tau)$. Given the size of the compartment population at τ, each vehicle makes its exit independently of the others, so that the number of vehicles exiting is a binomial random variable with parameters $z_k(\tau)$ and $P_k(\tau)$. To derive plausible forms for the exit probabilities $P_k(\tau)$, imagine that vehicle l in segment k at the beginning of τ has a speed u_{kl} and a location s_{kl} that denotes the distance from vehicle l to the end of the downstream boundary of segment k. Also assume that the speeds u_{kl} and s_{kl} are assigned to the vehicles as independent, identically distributed random variables with density functions $f_k(u)$ and $g_k(s)$, respectively. Since vehicle l will exit segment k only if $s_{kl} < u_{kl} * \Delta$, the exiting probability $P_k(\tau)$ is

$$P_k = \text{prob}[s_{kl} < \Delta u_{kl}] = \int_0^\infty \int_0^{\Delta u_{kl}} g_k(s) f_k(u)\, ds\, du \tag{4}$$

and if it is assumed that the locations of vehicles are uniformly distributed, so that $g_k(s) = 1/L_k$, this double integral can be easily evaluated to produce

$$P_k = \frac{\Delta \overline{U}_k(\tau)}{L_k} \tag{5}$$

Here $\overline{U}_k(\tau)$ denotes the space mean speed of vehicles in segment k at the beginning of τ. This connection between the exit probability for a segment and its space-mean speed implies that a stochastic version can be formulated for any traffic flow model that describes space-mean speed. More detailed discussion of the ideas underlying Equation 5 can be found elsewhere (*14,15*).

In STOMAC, the state variables are $z_{kj}(\tau)$, the number of vehicles in segment k destined for off-ramp j at the beginning of the subinterval, and $\overline{U}_k(\tau)$, the space-mean speed of the vehicles in segment k. Assume that the random arrivals at on-ramps follow Poisson distributions and that the random exits from segments follow binomial distributions. That is,

$$y_{kj}(\tau) = \text{binomial}\,[z_{kj}(\tau), \Delta \overline{U}_k(\tau)/L_k] \tag{6}$$

$$q_{ij}(\tau) = \text{Poisson}\,[b_{ij} * q_i(\tau)] \tag{7}$$

In each segment, the number of vehicles satisfies the conservation equation

$$z_{kj}(\tau + 1) = z_{kj}(\tau) + y_{k-1,j}(\tau) - y_{kj}(\tau) + \Sigma_i w_{ik} q_{ij}(\tau) \tag{8}$$

where $w_{ik} = 1$ if on-ramp i joins segment k, and 0 otherwise.

Finally, Payne's discretized momentum equation (*16*) describes the evolution of $\overline{U}_k(\tau)$,

$$\overline{U}_k(\tau + 1) = \overline{U}_k(\tau) + \Delta \overline{U}_k(\tau) \frac{\overline{U}_{k-1}(\tau) - \overline{U}_k(\tau)}{L_k} + \frac{\Delta}{\Gamma}\{\overline{U}_e[r_k(\tau)] - \overline{U}_k(\tau)\} - \nu\Delta \frac{d_k r_{k+1}(\tau) - r_k(\tau)}{L_k \Gamma[r_k(\tau) + \kappa M_k]} \tag{9}$$

where $d_k = M_k/M_{k+1}$ and Γ, κ, ν are momentum equation parameters, which generally must be estimated.

STOMAC can be used to generate a series of simulated on-ramp volumes, distribute these volumes to off-ramps, and then propagate these destination-specific subflows. These simulated data make it possible to investigate the statistical properties of estimators for the O-D parameters, **B**. FORTRAN source listings for STOMAC and other computer programs used in this study can be found elsewhere (*17*).

DESCRIPTION OF ESTIMATION APPROACHES

Ordinary Least Squares

As pointed out earlier, the problem of estimating freeway O-D patterns is analogous to the problem of estimating the turning movement proportions for single intersections, where methods based on OLS can give useful estimates. The basic idea behind this approach is that from the standpoint of the traffic manager, the actual destinations selected by the vehicles arriving at an on-ramp are unknown, and if all that is known are the O-D proportions b_{ij} and the arrival volumes $q_i(t)$, the O-D demands $x_{ij}(t)$ can be viewed as generated by multinomial random outcomes. Ignoring travel time lags, the expected values of the off-ramp volumes are then as given in Equation 1 and the OLS estimates of the b_{ij} are found by minimizing Equation 3. This problem is solved easily using standard linear regression software.

Expectation-Maximization

Under reasonably general conditions, maximum likelihood (ML) estimates tend to be asymptotically efficient, which suggests that ML estimates may be more effective in tracking time-varying O-D patterns. Since, under the linear model, the off-ramp counts are simply sums of independent multinomial outcomes, the likelihood function of the off-ramp counts is theoretically available; in practice, however, since it will have the form of a multidimensional convolution, it will be very difficult to compute. The expectation-maximization (EM) algorithm has been recommended for problems of this type (*18*), and its basic idea is as follows. If one were able to observe the individual O-D-specific traffic flows $x_{ij}(t)$, the ML estimator for the O-D parameters would simply be

$$\hat{b}_{ij} = \frac{\Sigma_t x_{ij}(t)}{\Sigma_t q_i(t)} \qquad i = 1, \ldots, m, j = 1, \ldots, n \tag{10}$$

The practical problem, however, is to estimate b_{ij} when no $x_{ij}(t)$ can be observed directly and only the entering counts $q_i(t)$ and the exiting counts $y_j(t)$ are known from the freeway surveillance and control systems. Note that since

$$y_j(t) = \Sigma_i x_{ij}(t) \qquad j = 1, \ldots, n$$

this is an incomplete data problem, in which the sufficient statistics $\Sigma_i x_{ij}(t)$ are not observed directly. The authors' implementation of the EM algorithm begins with an initial estimate $\mathbf{B}_o$ and then estimates the conditional expectations of the $x_{ij}(t)$ using normal distribution methods:

$$\Sigma_t x_{ij}(t) = E\{\Sigma_t x_{ij}(t) \mid \mathbf{B}, \mathbf{y}(t), t = 1, \ldots, N\} \tag{11}$$

The $\mathbf{B}_o$ is then reestimated by substituting $\Sigma_t x_{ij}(t)$ in Equation 10 for $\Sigma_t x_{ij}(t)$. The process iterates between Equations 10 and 11 until a convergence criterion is satisfied. For single intersections, where the travel time differences between each O-D pair can be ignored, this EM estimator tends to give O-D estimates with considerably less sampling variability than does the OLS estimator. More detailed presentation of the EM formulas can be found elsewhere (*3,10*).

Constrained Approximate Maximum Likelihood

The EM algorithm was based on a multinomial likelihood function but used a normal approximation for the probability distribution of the $x_{ij}(t)$. Alternatively, a normal approximation could be used for the $y_j(t)$ and approximate ML estimates could be computed via the resulting likelihood function. A description of this estimation approach, called constrained approximate ML (CAML), is presented in the following.

Given the on-ramp observations $\mathbf{q}(t)$, the expected value of $\mathbf{y}(t)$ is

$$\hat{\mathbf{y}}(t) = E[\mathbf{y}(t) \mid \mathbf{q}(t)] = \mathbf{B}^T * \mathbf{q}(t) \tag{12}$$

where the covariance matrix of $\hat{\mathbf{y}}(t)$ can be obtained as

$$\mathbf{V}(t) = \text{cov}[\hat{\mathbf{y}}(t) \mid \mathbf{q}(t)] = \Sigma_i q_i(t)[\text{diag}\{\mathbf{b}_i\} - \mathbf{b}_i * \mathbf{b}_i^T] \tag{13}$$

Since $\mathbf{y}(t)$ is the sum of multinomial random vectors, for large values of $q_i(t)$ it will be approximately normally distributed, with approximate likelihood function

$$L = \Pi_t[(2\pi)^{m-1} * \mid \mathbf{V}(t) \mid]^{-0.5} * \exp\{-0.5 * [\mathbf{y}(t) - \mathbf{B}^T * \mathbf{q}(t)]^T * \mathbf{V}^{-1}(t)[\mathbf{y}(t) - \mathbf{B}^T * \mathbf{q}(t)]\} \tag{14}$$

Taking the logarithm of Equation 14 and simplifying results in the final objective function: Minimize

$$LL = \Sigma_t\{\log \mid \mathbf{V}(t) + [\mathbf{y}(t) - \mathbf{B}^T * \mathbf{q}(t)]^T * \mathbf{V}^{-1}(t)[\mathbf{y}(t) - \mathbf{B}^T * \mathbf{q}(t)]\} \tag{15}$$

subject to the constraints of Equation 2.

The $\mathbf{B}$ matrix that solves this problem will be the CAML estimates for the O-D parameters. Both the EM and CAML estimators can be viewed as constrained quasi-ML methods in which the underlying data generation process is approximated by the simple linear assignment model. Thus they can be viewed as attempts to preserve the simplicity of the linear model on the assumption that inefficiency rather than bias is responsible for the poor performance of OLS in Table 1.

Nonlinear Least Squares

One of the major dissimilarities between traffic flow at a single intersection and that on freeways is that the travel times between each freeway O-D pair will vary depending on the intervening traffic conditions. The three estimation procedures described earlier achieve computational simplicity by ignoring this travel time variability. As a benchmark, it was desirable to have an estimator that was not subject to specification error but still optimized with respect to the least-squares criterion. This led to the following nonlinear least-squares (NLS) approach.

Given a current estimate of $\mathbf{B}$, forecasted off-ramp counts were computed by performing the STOMAC recursion with the Poisson and binomial random numbers replaced by their expected values. Forecasted off-ramp counts for the subintervals were aggregated to produce forecasted 5-min counts $\hat{y}_j(t)$, and the sum-of-square function was computed as

$$SS = \Sigma_t \Sigma_j [y_j(t) - \hat{y}_j(t)]^2 \tag{16}$$

A final **B** matrix that minimized Equation 16 was computed iteratively by the nonlinear optimization routine E04JAF, which is contained in the NAG Workstation Library. By comparing the performance of NLS with that of EM and CAML, it should be possible to separate the effects of specification error versus an inefficient optimization criterion on O-D estimator performance. Software implementing STOMAC and the four estimators was written in FORTRAN, and all computations were carried out on a Sun Sparcstation 1+.

EVALUATION AND COMPARISON OF PARAMETER ESTIMATORS

Generation of Simulated Data Sets

So that the statistical properties of these estimators could be evaluated under plausible conditions, it was decided to calibrate STOMAC to an existing section of freeway rather than to construct a hypothetical example. Figure 2 depicts a seven-origin, four-destination section of northbound Interstate I-35W in south Minneapolis, Minnesota. The section is 2.5 mi (4.2 km) long and has a somewhat complicated O-D pattern. A sequence of 36 five-minute counts was obtained from the Minnesota Department of Transportation (MNDOT) for a typical morning weekday peak period from 6:00 to 9:00 a.m.

For STOMAC to be used to simulate traffic flow on this freeway section, two sets of model parameters must be determined. The first set governs the traffic flow properties of the freeway and consists of estimates of capacity and free-flow speed, needed for the equilibrium speed-density relationship, and the momentum equation parameters Γ, κ, ν. The second set of parameters consists of the O-D proportions b_{ij}. For this example, a capacity of 2,000 vehicles per lane per hour and free-flow speed of 65 mph (108 km/hr) were used, and values for the momentum equation parameters were taken from the paper by Cremer and May (*19*). Given these values, the O-D proportions were estimated using the 5-min count data provided by MNDOT via the NLS procedure described earlier; as a rough check of the plausibility of this model, Figure 3 displays the actual 5-min traffic counts for the downstream boundary of this freeway segment, along with the predicted values obtained using the parameterized model. The predicted values track the actual ones reasonably well. These estimates were then used in STOMAC to generate 50 data sets, each consisting of a simulated 3-hr sequence of 5-min on- and off-ramp counts, with the mean value of the Poisson arrivals being set equal to the actual 5-min on-ramp counts.

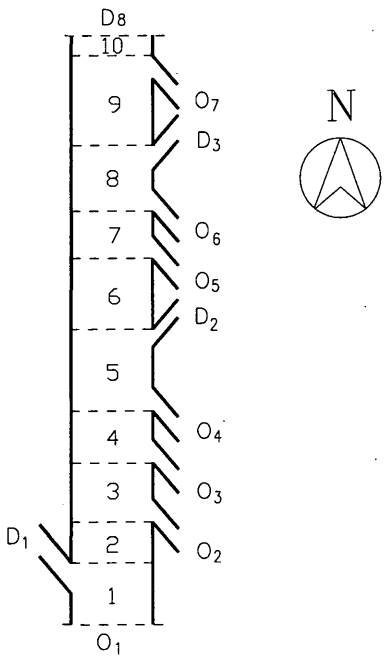


FIGURE 2 Freeway section with seven on-ramps and four off-ramps on I-35W, northbound.

Comparison and Evaluation of Results

By running each estimator mentioned earlier on each of the 50 data sets, the authors obtained samples of the estimators' behavior. From these samples, the sample means and standard deviations were computed in order to evaluate the statistical properties of unbiasedness and efficiency. Efficiency of an estimator is defined here in term of its sampling variance, which estimates the standard error. That is, the standard deviation of estimated parameters from the sample should be small in order for the estimator to be recognized as efficient. An unbiased estimator should be able to produce estimates that on the average equal the "true" parameter values. These results are displayed in Table 2. As an aggregate measure of the joint effect of bias and inefficiency, Table 2 also presents the root mean square (RMS) error between the true value and the estimates, which is computed by

$$RMS_{ij} = \sqrt{(\bar{b}_{ij} - b_{ij})^2 + var(\hat{b}_{ij})}$$

and can be interpreted as the average distance separating an estimate from the true value. The average CPU times needed to compute estimates for one data set are given here:

Estimator	*Average CPU Time (sec)*
OLS	0.18
NLS	2394.4
EM	60.1
CAML	67.0

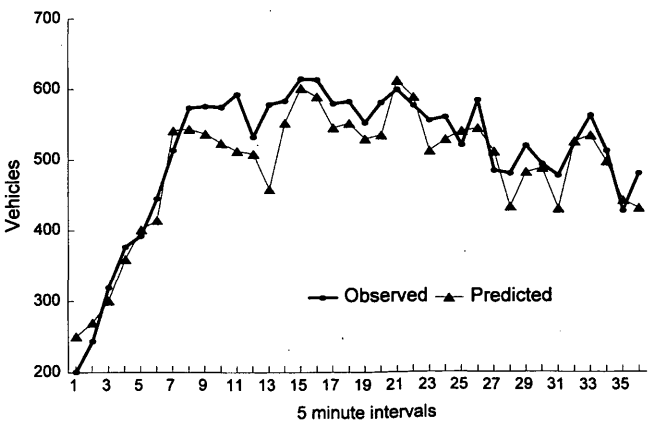


FIGURE 3 Comparison between observed and predicted mainline traffic volumes.

TABLE 2 Performance of Freeway O-D Parameter Estimators

		OLS			NLS		
	True	Mean	S.Dev	RMS	Mean	S.Dev	RMS
b11	.056	.055	.003	.003	.055	.004	.004
b12	.134	.133	.041	.041	.136	.022	.022
b13	.019	.025	.056	.056	.020	.030	.030
b14	.790	.671	.123	.171	.798	.042	.043
b22	.291	.388	.321	.335	.303	.055	.056
b23	.290	.250	.346	.348	.295	.080	.080
b24	.419	.993	.902	1.07	.409	.062	.063
b32	.223	.225	.055	.055	.234	.034	.036
b33	.148	.140	.047	.048	.145	.028	.028
b34	.629	.618	.146	.146	.619	.045	.046
b42	.194	.203	.129	.129	.195	.038	.038
b43	.240	.208	.138	.142	.232	.055	.056
b44	.566	.646	.354	.363	.574	.059	.060
b53	.263	.248	.077	.078	.260	.047	.047
b54	.737	.825	.191	.210	.743	.054	.054
b63	.284	.297	.376	.376	.281	.029	.029
b64	.716	.841	1.07	1.08	.715	.058	.058
		EM			CAML		
b11	.056	.056	.002	.002	.057	.005	.005
b12	.134	.176	.012	.044	.133	.031	.031
b13	.019	.137	.022	.120	.032	.031	.034
b14	.790	.630	.031	.163	.778	.043	.045
b22	.291	.193	.041	.106	.246	.121	.129
b23	.290	.167	.064	.139	.293	.134	.134
b24	.419	.639	.074	.232	.461	.151	.157
b32	.223	.182	.018	.045	.219	.043	.043
b33	.148	.134	.043	.045	.138	.041	.042
b34	.629	.685	.042	.070	.644	.043	.046
b42	.194	.201	.071	.071	.266	.158	.174
b43	.240	.220	.092	.094	.263	.138	.140
b44	.566	.579	.099	.100	.470	.181	.205
b53	.263	.105	.031	.161	.254	.054	.055
b54	.737	.895	.031	.161	.746	.054	.055
b63	.284	.123	.084	.182	.329	.175	.181
b64	.716	.877	.084	.182	.671	.175	.181

Table 2 indicates that the NLS estimator, for which specification error was not present, appears to be unbiased and is more efficient than the other approaches. Although the OLS approach seems to produce unbiased estimates, the high sampling variability makes even substantial differences between the sample average and the true value appear statistically insignificant. For practical purposes, the OLS estimates are essentially useless. For example, the "true" value of b_{44} is 0.566 and the mean and the standard deviation of OLS estimate are 0.646 and 0.354, respectively. The means that, appealing to the approximate normality of the OLS estimates, 95 percent of the estimates for b_{44} would fall in the interval [–0.62, 1.35], an interval that includes the true value; but, practically speaking, 100 percent of the estimates should fall in the interval [0,1].

For the EM estimator, the efficiency is comparable to NLS, but EM tended to give highly biased estimates, a tendency that Nihan and Davis also reported in their intersection study. One interesting finding is that the CAML estimator appears to be a useful compromise between the accurate, but computationally demanding, NLS procedure and the computationally simple, but inaccurate, OLS procedure. For the proportions corresponding to On-Ramps 1, 3, and 5, CAML has an efficiency of the same order of magnitude as NLS. This suggests that, at least for fairly short freeway segments, switching to an approximate ML approach can partly compensate for the effects of specification error.

The results given in Table 2 can also be interpreted as giving bounds on the expected accuracy of recursive, tracking algorithms. For example, after processing 3 hr worth of 5-min observations, one could expect an NLS estimate of b_{22} to be in the interval [0.29 – 0.11, 0.29 + 0.11] = [0.18, 0.40]. If b_{22} had changed during this period, then any estimate of these time-varying values, being based on fewer observations, would be less accurate than this.

CONCLUSIONS

This paper began by pointing out the importance of parameter estimation for ATMS practice, with attention to the fact that uncertainty in model parameter estimates will affect the effectiveness of control

policies and their potential benefits. As with all simulation studies, these results should be considered illustrative rather than definitive. Certainly one can imagine constructing examples for which the outcome might be different. However, the simulation example used here was based on an existing section of freeway and on a traffic flow model that most would regard as plausible, if not conclusive, so there is good reason to expect that these results are more likely to be typical rather than anomalous.

Probably the most challenging aspect of these results is that no matter what O-D estimation procedure is used, a nontrivial amount of uncertainty concerning the actual parameter value will remain after processing 3 hr of data, and if the O-D parameters are in fact time-varying, this residual uncertainty will only increase. This calls into question the common practice of "certainty equivalent" prediction and control, in which parameter estimates are treated as known constants rather than as the uncertain quantities that they are and suggests that forecasting and control models that explicitly treat parameter uncertainty may improve on current practice.

ACKNOWLEDGMENTS

This research was supported in part by the Minnesota Department of Transportation and Center for Transportation Studies and in part by the National Science Foundation.

REFERENCES

1. Davis, G. A Statistical Theory for Estimation of Origin-Destination Parameters from Time Series of Traffic Counts. In *Proc., 12th International Symposium on Transportation and Traffic Theory* (C. Daganzo, ed.), Elsevier Science, Amsterdam, The Netherlands, 1993, pp. 441–463.
2. Nguyen, S. Estimating Origin-Destination Matrices from Observed Flows. In *Transportation Planning Models* (M. Florian, ed.), North Holland, Amsterdam, The Netherlands, 1984, pp. 363–380.
3. Davis, G. Estimating Freeway Demand Patterns and Impact of Uncertainty on Ramp Controls. *Journal of Transportation Engineering*, ASCE, Vol. 119, 1993, pp. 489–503.
4. Nihan, N., and M. Hamed. Fixed Point Approach to Estimating Freeway Origin-Destination Matrices and the Effect of Erroneous Data on Estimate Precision. In *Transportation Research Record 1357*, TRB, National Research Council, Washington, D.C., 1991, pp. 18–28.
5. Chang, G.-L., and J. Wu. Recursive Estimation of Time-Varying O-D Flows from Traffic Counts in Freeway Corridors. Presented at 72nd Annual Meeting of the Transportation Research Board, Washington, D.C., 1993.
6. Ashok, K., and M. E. Ben-Akiva. Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems. In *Proc., 12th International Symposium on Transportation and Traffic Theory* (C. Daganzo, ed.), Elsevier Science, Amsterdam, The Netherlands, 1993, pp. 465–484.
7. Cremer, M., and H. Keller. Dynamic Identification of Flows from Traffic Counts at Complex Intersections. In *Proc., 8th International Symposium on Transportation and Traffic Theory* (V. Hurdle, ed.), University of Toronto Press, Ontario, Canada, 1983, pp. 121–142.
8. Cremer, M., and H. Keller. A New Class of Dynamic Methods for Dynamic Identification of Origin-Destination Flows. *Transportation Research*, Vol. 21B, 1987, pp. 117–132.
9. Nihan, N., and G. Davis. Recursive Estimation of Origin-Destination Matrices from Input/Output Counts. *Transportation Research*, Vol. 21B, 1987, pp. 149–163.
10. Nihan, N., and G. Davis. Application of Prediction-Error Minimization and Maximum Likelihood To Estimate Intersection O-D Matrices from Traffic Counts. *Transportation Science*, Vol. 23, 1989, pp. 77–90.
11. Bell, M. The Real-Time Estimation of Origin-Destination Flows in the Presence of Platoon Dispersion. *Transportation Research*, Vol. 25B, 1991, pp. 115–125.
12. Ljung, L., and T. Soderstrom. *Theory and Practice of Recursive Identification*. MIT Press, Cambridge, Mass., 1983.
13. Lehoczky, J. Approximations for Interactive Markov Chains in Discrete and Continuous Time. *Journal of the Mathematics Society*, Vol. 7, 1980, pp. 139–157.
14. Weiss, G., and R. Herman. Statistical Properties of Low-Density Traffic. *Quarterly of Applied Mathematics*, Vol. 20, 1962, pp. 121–130.
15. Solomon, H., and P. Wang. Nonhomogeneous Poisson Fields of Random Lines with Applications to Traffic Flow. In *Proc., 6th Berkeley Symposium on Mathematical Statistics and Probability* (L. Lecam, ed.), Vol. 3, University of California Press, Berkeley, 1970, pp. 383–400.
16. Payne, H. FREFLO: A Macroscopic Simulation Model of Freeway Traffic. *Transportation Research*, Vol. 722, 1979, pp. 68–77.
17. Yu, P. *Estimating Freeway Origin-Destination Patterns Using Automatic Traffic Counts*. Master's thesis. Department of Civil and Mineral Engineering, University of Minnesota, 1992.
18. Dempster, A., N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Vol. B39, 1977, pp. 1–38.
19. Cremer, M., and A. May. *An Extended Model of Freeway Traffic*. Report UCB-ITS-RR-85-7. Institute of Transportation Studies, University of California, Berkeley, 1985.

All opinions and conclusions expressed here are solely the responsibilities of the authors.

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

Using Neural Networks To Synthesize Origin-Destination Flows in a Traffic Circle

SHIH-MIAO CHIN, HO-LING HWANG, AND TZUSHENG PEI

The traffic circle is a classic transportation problem for traffic engineers. Although it is easy to determine the volume of vehicles entering and exiting the circle at all points, it is difficult to determine the actual flow pattern of these vehicles. In other words, although it is easy to determine how many vehicles enter the circle from a given street, it is difficult to determine how many of those vehicles will leave the circle at each possible exit point. Currently, the only method of accurately determining this traffic flow is to visually track each vehicle as it enters and exits the circle, a laborious method of collecting data. However, emerging neural network technologies give researchers another approach. The capability of neural networks to handle subtle or contradictory information by organizing and capturing complex relationships, optimizing and generating analytical models, and learning and adapting the model when new data become available has made them increasingly popular in transportation and traffic flow models. The objective is to describe the development of a neural network model for generating origin-destination (O-D) information for traffic circles based on observed flow volumes on approaching and exiting legs. The quality of the model is evaluated with respect to the different methods used to train the model. Observations about the synthesized O-D matrices and the corresponding errors generated by the neural network model are also described.

The traffic circle is a classic transportation problem for traffic engineers. It may be easy to determine the volume of vehicles entering and exiting the circle at all points, but it is difficult to determine the actual flow pattern of the vehicles entering and exiting the circle. In other words, although it is easy to determine how many vehicles enter the circle from a given street, it is difficult to determine how many of those vehicles will leave the circle at each possible exit point. Currently, the only method of accurately determining this traffic flow is to visually track each vehicle as it enters and exits the circle, a laborious method of collecting data. However, emerging neural network technologies give researchers an alternative approach. The capability of neural networks to handle subtle or contradictory information by organizing and capturing complex relationships, optimizing and generating analytical models, and learning and adapting the model when new data become available has made them increasingly popular in transportation and traffic flow models.

The objective of this paper is to describe the development of a neural network model for generating origin-destination (O-D) information for traffic circles based on observed flow volumes on approaching and exiting legs. The major emphasis of this paper is to evaluate the quality of the model with respect to the different methods used to train the model. Observations regarding the synthesized O-D matrices and the corresponding errors generated by the neural network model are also described.

S.-M. Chin and H.-L. Hwang, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tenn. 37831. T. Pei, University of Tennessee Transportation Center, Knoxville, Tenn. 37932.

NEURAL NETWORKS

Neural network models are algorithms for cognitive tasks, such as learning and optimization, that are loosely based on concepts derived from research into the nature of the brain. This paper adapts a formal definition of the neural network model as given by Müller and Reinhardt (*1*). In mathematical terms, a neural network model is defined as a directed graph with the following properties:

- A state variable s_i is associated with each node (neuron) i;
- A real-value weight w_{ij} (also known as coupling strength, synaptic strength, or synaptic efficacy) is associated with each link (synapse) (i,j) between two nodes i and j;
- A real-value bias (activation threshold) Φ_i is associated with each node (neuron) i; and
- A transfer function $f_i[s_i, w_{ij}, \Phi_i, (j \neq i)]$ or $f(\Sigma_j w_{ij} s_i - \Phi_i)$ is defined for node i that determines the state of the node as a function of its bias, the weights of its incoming links, and the states of the nodes connected to it by these links. The transfer function is often either a discontinuous step function or a smoothly increasing generalization known as a sigmoidal function.

Nodes without links leading toward them are called input nodes, and nodes without links leading away from them are called output nodes. A feed-forward network is a neural network that admits no closed path. A simple, multilayered, feed-forward neural network model is presented in Figure 1.

BACKPROPAGATION

Multilayered, feed-forward neural network models have recently been applied to many fields because of the development of an efficient method for determining the synaptic coupling strengths of such models. This method, called error backpropagation, is a supervised learning algorithm that iteratively adjusts the synaptic strengths w_{ij} so that the output signal differs as little as possible from the desired target. This is achieved by applying the gradient method, which yields the required modification Δw_{ij}. Since the operation of the network corresponds to a highly nonlinear mapping between the input and output—the transfer function is nonlinear—the method must be applied many times until convergence is reached.

Before training, initial synaptic strengths are applied to all node connections, and activation thresholds, which change over the training process, are set for each node. A global activation function calculates the output value of each node as the sum of the synaptic strengths multiplied by the corresponding values of the previous layer's nodes. An error function is defined that is the sum of the squares of the difference between the desired output and the actual

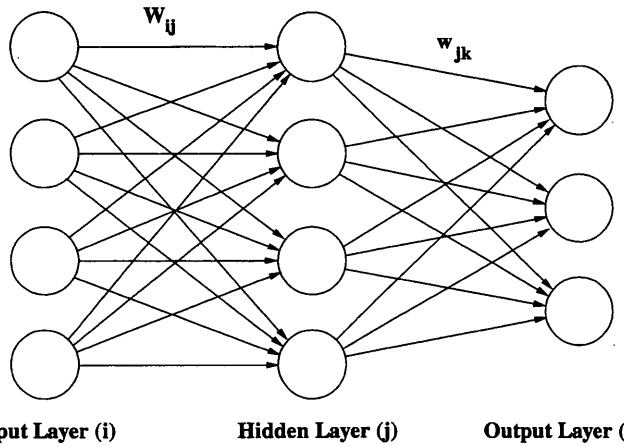


FIGURE 1 Example of multilayer, feed-forward neural network model.

output from the network. The backpropagation algorithm prescribes a method to minimize this error function by taking gradient with respect to the synaptic strengths and decides on the amount of incremental strength adjustment. Training is a series of running input-output pairs over the network and making incremental adjustments to the synaptic strength values. The backpropagation mechanism is described in more detail by Hertz et al. (*2*) and Müller and Reinhardt (*1*).

PREVIOUS RESEARCH

Yang et al. (*3*) have adopted a feed-forward neural network model for synthesizing O-D flows for both a four-way intersection and a short freeway segment. A two-layered, feed-forward neural network was built to model a four-way intersection. This network has four nodes in the input layer for modeling the entrances and four nodes in the output layer for modeling exits. A sigmoidal function serves as the transfer function. The optimization on a squared-error function was based on the error backpropagation method. After the training is completed, the weights of the connections from the input to the output layers are interpreted as the turning movement ratios. On the basis of the training data, the trained weights essentially summarize the traffic coming into and going out from the intersection in terms of ratios. From simulation results, the Yang et al. model has shown that a method based on backpropagation can estimate turning movement ratios with high tracking ability and stability.

NEURAL NETWORK TRAFFIC CIRCLE O-D MODEL

Physical Network

The network modeled in this study, Church Circle, is a traffic circle in the historic district of Annapolis, Maryland. Church Circle is the primary focal point for traffic entering and leaving downtown Annapolis, so large volumes of traffic flow through it. The circle connects to eight streets: College Avenue, School Street, Main Street, Duke of Gloucester Street, South Street, Franklin Street, West Street, and Northwest Street (Figure 2). Main, Franklin, West, School, and Northwest Streets and College Avenue contain lanes that enter the circle; Duke of Gloucester, South, Franklin, West, and School Streets and College Avenue contain lanes that exit the circle. It should also be noted that Main, Northwest, South, and Duke of Gloucester are one-way streets.

Traffic Flow Data

The data used for this model were extracted from an O-D license plate survey conducted during morning (7:00 to 9:00 a.m.), noon (11:30 a.m. to 2:30 p.m.), and afternoon (3:30 to 6:00 p.m.) peak periods to provide actual O-D volumes. These O-D traffic volumes were collected for 15-min intervals. Therefore, the data are comprised of 8 sets of O-D matrices for the morning peak period, 10 sets for the noon peak period, and 10 sets for the afternoon peak period.

On the basis of the data collected from the Church Circle site, the traffic patterns were significantly different for the morning, noon, and afternoon periods (Figure 3). In the morning, the O-D volume from Main to College was the highest, and the O-D traffic from Main to West was the second highest. Traffic volumes from all origins to Duke of Gloucester were significantly higher than those to South, Franklin, and School. During the noon period, the two highest O-D volumes were from School to College and from West to Duke of Gloucester. This is significantly different from the morning traffic pattern. On the other hand, traffic volumes from all origins to Duke of Gloucester are higher than during the morning period, but they are still significantly higher than those from all origins to South, Franklin, and School. During the afternoon period, the two highest O-D volumes were from School to College and from School to West. However, the overall traffic pattern is similar to that of the noon period except that Duke of Gloucester ceases to be a significant destination in the afternoon.

The standard deviations for the actual traffic flow during each of the three periods are shown in Figure 4. The standard deviations were calculated so that their effects with regard to the synthesized results could be determined. These standard deviations were calculated as follows:

$$Std\ dev_{odp} = \sqrt{\frac{\Sigma_{l=1,Lp}\,(t_{odl} - \bar{t}_{od})^2}{L_p}} \qquad (1)$$

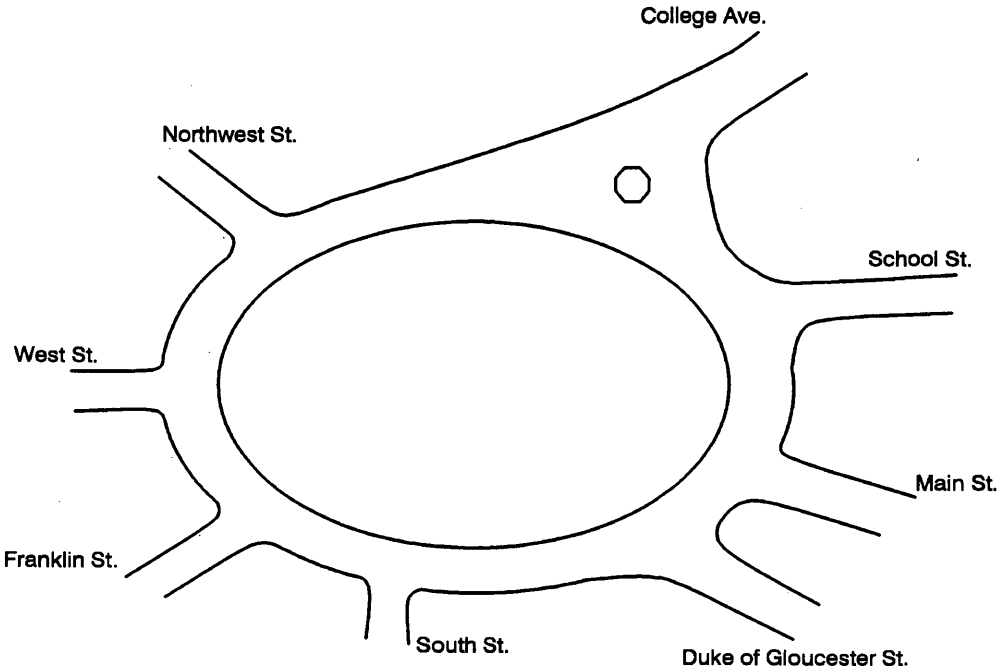


FIGURE 2 Church Circle in Annapolis, Md.

where

o = origin street,
d = destination street,
p = period (morning, noon, afternoon),
l = 15-min interval (1–8 for morning, 1–10 for noon and afternoon),
t = traffic volume, and
L_p = number of 15-min intervals in period p.

The pooled total standard deviation for a period p is defined as follows:

$$Pooled\ total\ std\ dev_p = \sqrt{\frac{\Sigma_{o=1,6}\ \Sigma_{d=1,6}\ std\ dev^2_{odp}}{6 \times 6}} \quad (2)$$

The pooled total standard deviations based on actual O-D traffic flows were 10.64, 13.88, and 9.20 for morning, noon, and afternoon periods, respectively.

Model Formulation

A multilayer, feed-forward neural network model was formulated to synthesize the O-D matrix based on the traffic entering and exiting the circle (Figure 5). Prior experiences with neural networks and other documented sources have indicated that feed-forward neural networks with multiple hidden layers and many nodes do not necessarily produce better results. Models with one, two, and three hidden layers with 5 nodes were tested, as was a model with one hidden layer with 20 nodes. There was no significant difference in the results produced by these models. Thus, for the sake of simplicity, a basic three-layered model with a single five-node hidden layer was used. The primary reason for using the model presented in Figure 5 is that such a model has the potential to "learn" from rigorous "training" and to synthesize O-D traffic flows for any traffic circle intelligently.

There are two significant differences between the model presented in this paper and the model of Yang et al. (*3*). The first is the physical network to which each model is applied. The Yang et al. model synthesizes traffic turning movements for an intersection, whereas the model presented in this paper attempts to synthesize O-D traffic flows for a traffic circle. The traffic flow for the traffic circle is much more complex than the traffic flows for a typical four-way intersection, and the effort required to manually determine these movements is much more involved. The second difference is in the configuration of the neural network model itself. The Yang et al. model uses a two-layer network with no hidden layer. The entering counts are fed into the input layer, and the exiting counts are fed into the output layer. The resulting synaptic strengths between these layers are the turning movement ratios. Thus, Yang et al. merely use the neural network framework to model the intersection traffic turning movements. The neural network model based on field data is applicable only for that particular set of data, and the backpropagation algorithm would have to be used to reestimate the synaptic strengths every time a new set of traffic volumes was collected.

The model presented in this paper uses the neural network in the "conventional" manner. It has a hidden layer. The entering and exiting traffic volumes are fed into the input layer and the desired, or target, O-D matrix is fed into the output layer to adjust the synaptic strengths (i.e., train the model). There is no constraint on the synaptic strengths associated with each connection, and all synaptic strengths estimated by the backpropagation algorithm are retained as an integral part of the neural network model. After training, entering and exiting volumes can be input into the model to generate an O-D matrix. Using this modeling approach, the backpropagation algorithm does not have to be reapplied to estimate the synaptic strengths for the model every time a new set of traffic volumes is collected.

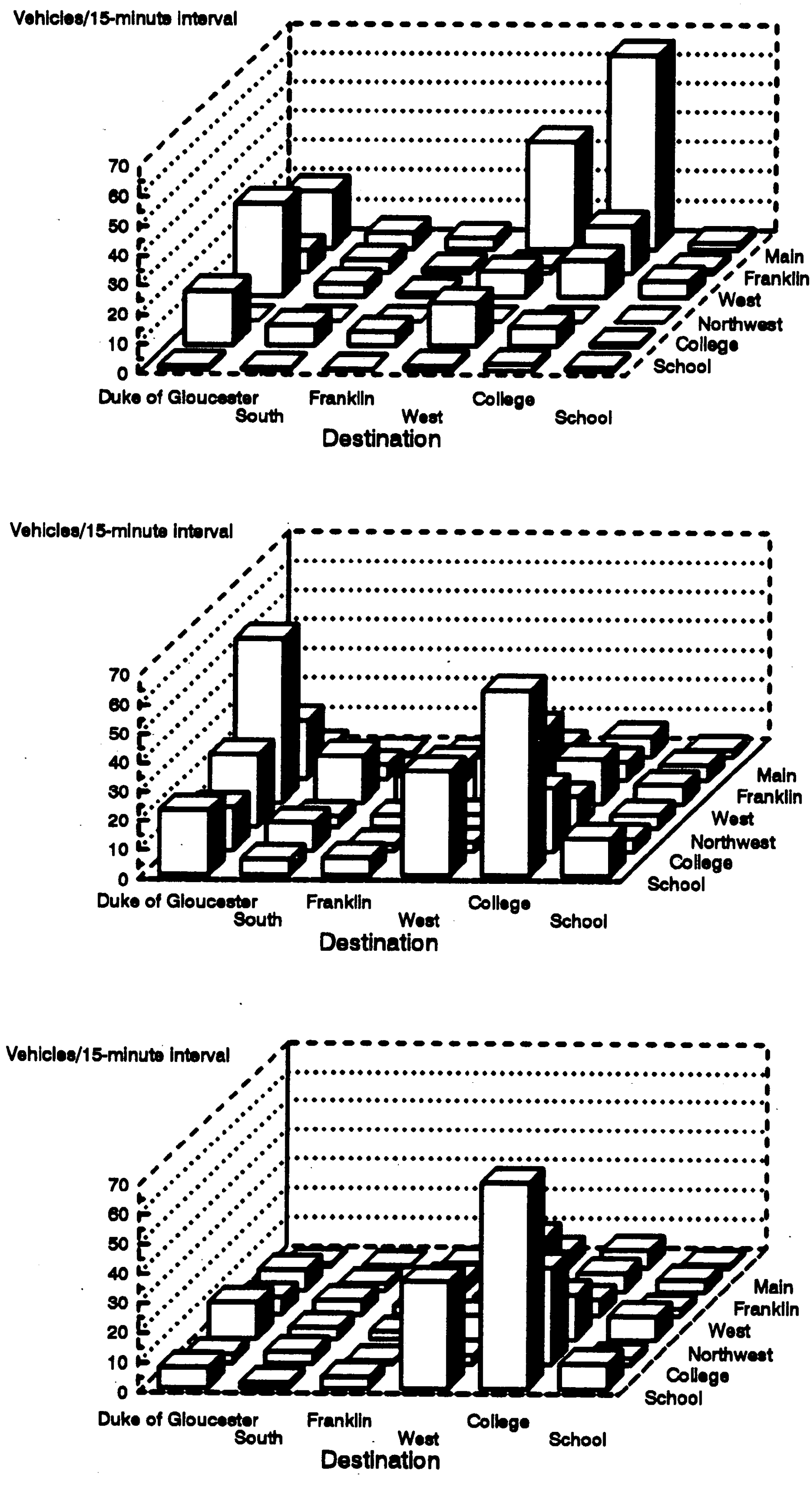


FIGURE 3 Average O-D volumes for morning (*top*), noon (*middle*), and afternoon (*bottom*) peak periods.

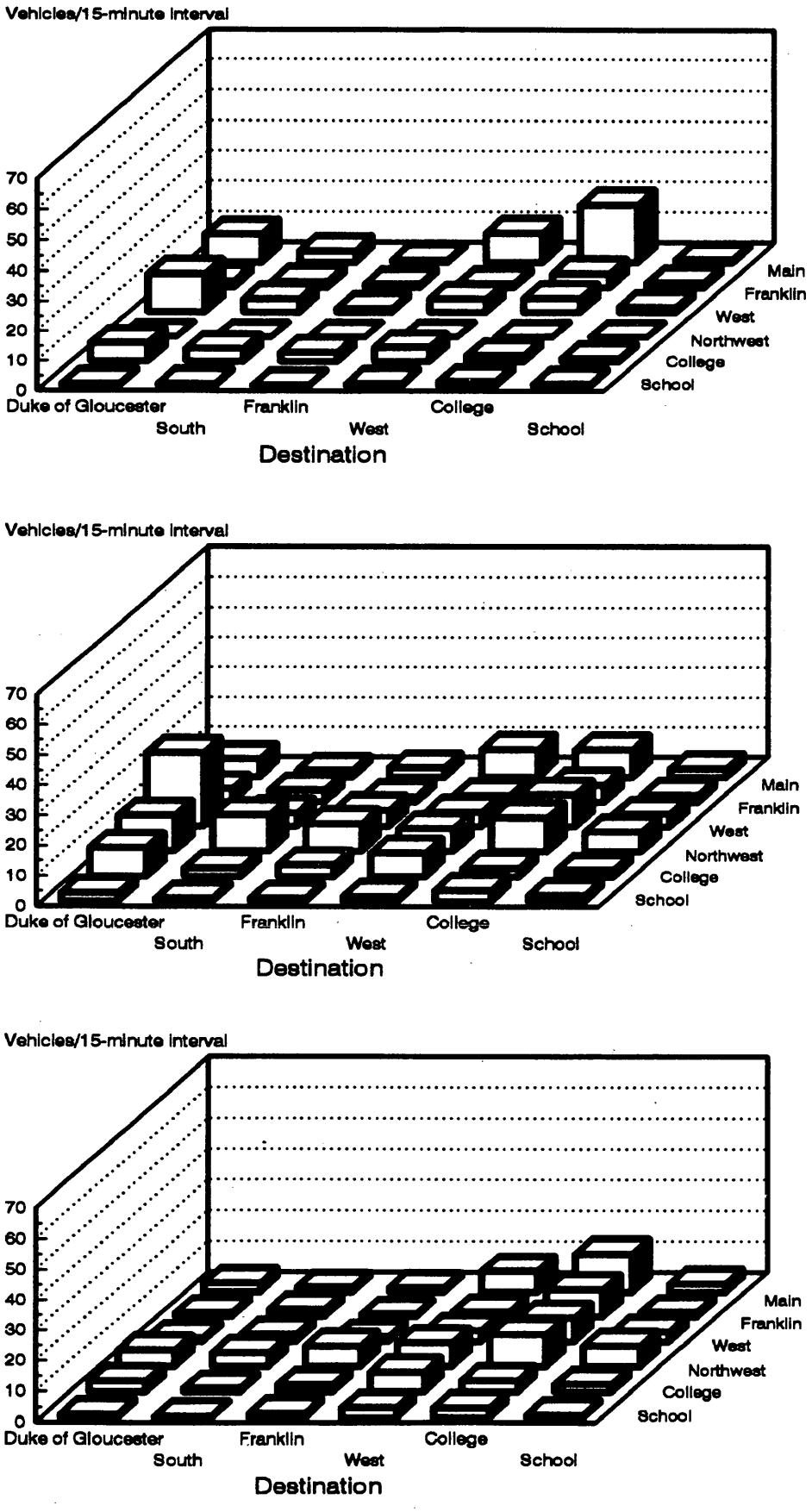


FIGURE 4 Standard deviations for morning (*top*), noon (*middle*), and afternoon (*bottom*) volumes.

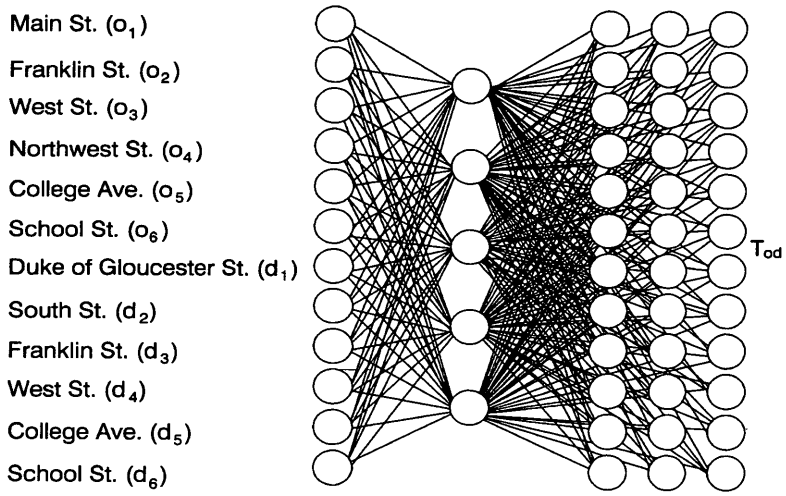


FIGURE 5 Neural network O-D flow model for Church Circle.

EXPERIMENTAL PROCEDURE

The experimental procedures for all of the trials described in this paper consist of the following three steps:

1. *Training the model.* The training set was extracted from the O-D license plate survey. This was accomplished by using the traffic volumes from the approaching and exiting streets as the input and the O-D volumes as the target output for each corresponding time interval. A time-interval parameter was included so that the model could develop a relationship between time periods and traffic patterns.
2. *Synthesizing O-D matrices.* The traffic volumes for approaching and exiting streets were used as input for the model to synthesize an O-D matrix for each corresponding time interval. As in training, the time interval for each set of volumes was specified so that it could be considered by the model.
3. *Evaluating the synthesized O-D matrices.* The synthesized O-D matrices produced by the model were evaluated by comparing them to the actual O-D matrices determined by the survey.

However, rather than comparing the synthesized and actual matrices for each time interval, the mean value for each O-D pair was calculated across all time periods. Mathematically, the mean of each O-D matrix was calculated using the following equation:

$$\bar{t}_{odp} = \Sigma_{l=1,L_p} \frac{t_{odl}}{L_p} \tag{3}$$

The averaged actual O-D matrices for the morning, noon, and afternoon periods were also derived using similar formulas. Then, the mean synthesized O-D flows were plotted against the averaged actual O-D flows for each corresponding cell with the matrix for morning, noon, and afternoon periods.

The results of the different trials are presented in Figures 6 through 9. These diagrams depict goodness-of-fit measures used to evaluate the neural network model's capability to synthesize O-D matrices based on the traffic flows from entering and exiting streets. If the neural network model can "learn" the O-D travel pattern on the basis of such a small data set, then the synthesized O-D flows should be close to the actual O-D flows. If the synthesized O-D flows are close to the actual O-D flows, then the points on these diagrams should be very close to the diagonal line. Thus, a diagram depicting the results of a good neural network model will have all or most points on or close to the diagonal line.

Two other goodness-of-fit measures are used to evaluate the synthesized O-D matrices generated by the model: mean absolute error and mean absolute average error. The formulations for these measures are described here:

$$\textit{Mean absolute error}_p = \frac{\Sigma_{o=1,6} \Sigma_{d=1,6} \Sigma_{l=1,L_p} |\hat{t}_{odl} - t_{odl}|}{36 \times L_p} \tag{4}$$

$$\textit{Mean absolute average error}_p = \frac{\Sigma_{o=1,6} \Sigma_{d=1,6} \left| \frac{\Sigma_{l=1,L_p} \hat{t}_{odl}}{L_p} - \frac{\Sigma_{l=1,L_p} t_{odl}}{L_p} \right|}{36} \tag{5}$$

where $\hat{t}_{odl}$ is the synthesized traffic from origin street o to destination street d during 15-min interval l, and t_{odl} is the actual traffic from origin street o to destination street d during 15-min interval l.

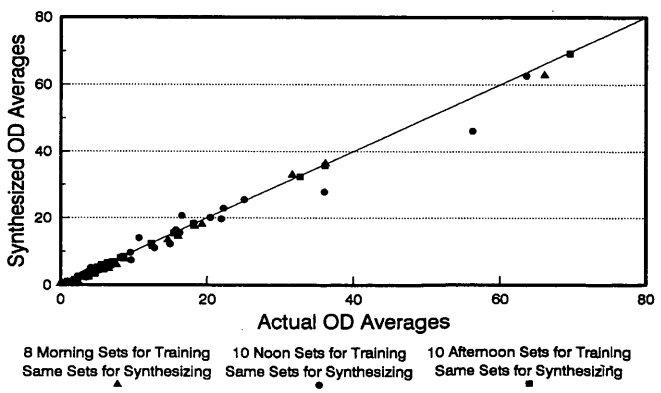


FIGURE 6 Comparison of O-D volumes synthesized by Method 1 and actual O-D volumes.

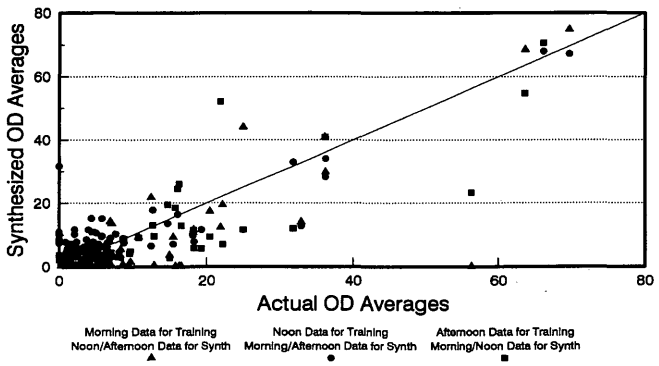


FIGURE 7 Comparison of O-D volumes synthesized by Method 2 and actual O-D volumes.

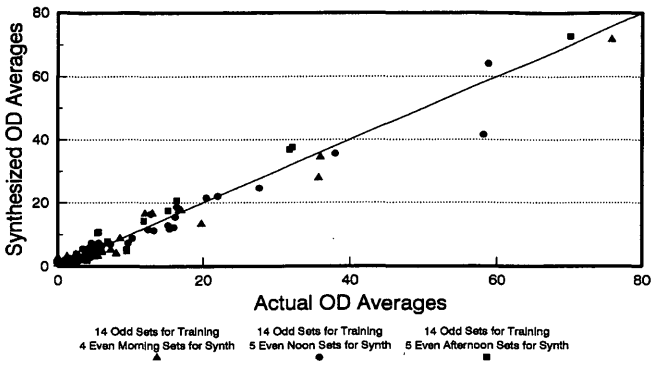


FIGURE 9 Comparison of O-D volumes synthesized by Method 4 and actual O-D volumes.

Training Model and Synthesizing O-D Matrices

Using the backpropagation method, various training methods were tested. O-D matrices were generated for each training method used. The various training schemes are described in the following sections.

Method 1

In Method 1, all traffic volume sets and all O-D matrices for each time period were used to train the model, and the same volume sets were used to synthesize O-D matrices.

In Trial 1, eight sets of morning traffic volume data (one set for each 15-min interval) from six entering streets and six exiting streets were used as input, and 36 O-D volumes were used as the output to train the model. The same eight sets of morning traffic volume data were input into the model to generate eight 36-element O-D matrices. These eight synthesized O-D matrices were then compared with the actual origin and destination traffic flows. This training scheme was used merely to generate some goodness-of-fit measures for the developed neural network model. Since the data used as input to synthesize O-D matrices were identical to those used to train the model, it is possible that the model "memorized" the training data and provided biased results.

Trial 2 was performed similarly to Trial 1, except that the noon traffic volume data were used to train the model and synthesize O-D matrices. The same procedure was followed in Trial 3 using afternoon training data and input.

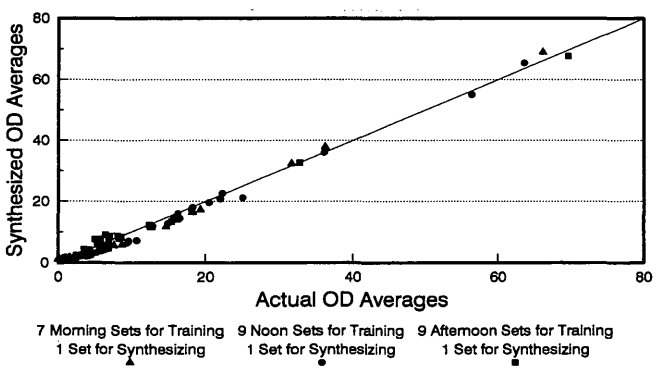


FIGURE 8 Comparison of O-D volumes synthesized by Method 3 and actual O-D volumes.

Method 2

In Method 2, traffic volume and O-D volume data from one time period were used to train the model, and traffic volume data from the other two time periods were used to synthesize O-D matrices for those periods.

For Trial 1, eight sets of morning traffic volume data (one set for each 15-min interval) from six entering streets and six exiting streets were used as input, and eight sets of 36-element O-D matrices were used as the output to train the model. The 10 sets of noon and afternoon traffic volumes (20 sets in all) were used as input for the model trained with morning traffic information. This input was used to synthesize ten 36-element O-D matrices for the noon peak period and ten for the afternoon period.

Trial 2 was conducted similarly to Trial 1, except that noon traffic volumes and O-D matrices were used to train the model, and morning and afternoon volumes were used to synthesize O-D matrices. In Trial 3, afternoon traffic volumes and O-D matrices were used to train the model, and morning and noon volumes were used to synthesize O-D matrices.

Method 3

In Method 3, all but one set of traffic volumes and flow matrices from one peak period were used to train the model, and the remaining set of traffic volumes were used to synthesize O-D matrices for that period.

For Trial 1, seven of the eight sets of morning volumes were used as input and the corresponding seven O-D matrices were used as output to train the model. The remaining set of morning volumes was used as input for the model to synthesize a 36-element O-D matrix.

Trial 2 was conducted similarly to Trial 1, except that nine sets of noon volumes and their corresponding matrices were used to train the model, and the remaining set of noon volumes was used as input for the model to synthesize a 36-element O-D matrix. In Trial 3, nine sets of afternoon volumes and their corresponding matrices were used to train the model, and the remaining set of afternoon volumes was used as input for the model to synthesize a 36-element O-D matrix.

This training scheme allows the model output to be compared with actual data not used to train the model. Thus, goodness-of-fit measures can be developed, ensuring that the output was not the

result of the network memorizing the training set. This procedure was repeated using a different combination of sets to train the model and synthesize an O-D matrix each time. Since the developed neural network model used different information in estimating the model parameters, there is little chance that the model "memorized" the training data; thus, the model results should be unbiased.

Method 4

In Method 4, the odd sets of morning, noon, and afternoon traffic volumes and flow matrices were used to train the model, and the even sets of morning, noon, and afternoon traffic volumes were used to synthesize morning, noon, and afternoon flow matrices.

To really evaluate the ability of the neural network model to synthesize O-D traffic flows based on general traffic conditions, a training scheme was designed that would not isolate data from each time period. Under this scheme, all 28 sets of traffic volumes were grouped together and numbered from 1 to 28 depending on the time that the data were collected. Of these 28 sets, the 14 odd-numbered sets were used as input and the 36 O-D traffic volumes were used as target output to train the model. The remaining fourteen sets were used as input for the model to synthesize fourteen 36-element O-D matrices. These 14 synthesized matrices were then compared with the actual O-D traffic flows for morning, noon, and afternoon separately. This training scheme allows the model output to be compared with actual data not used to train the model. Thus, goodness-of-fit measures can be developed, ensuring that the output is not the result of the network memorizing the training set. Since the developed neural network model used different information for training the model, there is little chance that the model "memorized" the training data; thus, the model results should be unbiased.

PRELIMINARY RESULTS AND CONCLUSIONS

On the basis of the results presented in Table 1 and Figures 6 through 9, the following preliminary conclusions have been reached:

- The backpropagation technique was used to estimate the weights w_{ij} used in all transfer functions $f_i[s_i, w_{ij}, \Phi_i, (j \neq i)]$ or $f(\Sigma_j w_{ij} n_j - \Phi_i)$. The synthesized O-D traffic volumes were compared with the actual O-D traffic volumes. As shown in Figure 6 and Table 1, the synthesized O-D volume averages are very close to the actual averages (all points lie very close to the diagonal line). This indicates that the backpropagation technique did find a set of weights w_{ij} used in the applied transfer functions such that the discrepancies between the synthesized data and actual data have been minimized to a satisfactory level. From the results presented in Figure 6 and Table 1, the backpropagation technique generated good neural network models for morning, noon, and afternoon periods for average origin and destination traffic flows.
- As Figure 7 and Table 1 indicate, the errors induced by the second method are much larger than those generated by the other three methods. However, these errors were expected since the traffic patterns are quite different among the three periods. As these results demonstrate, a neural network model cannot produce acceptable results for situations—in this case, variations in traffic patterns—for which it has not been trained. Consequently, a model trained with traffic data from one period cannot be used to synthesize O-D volumes for other periods. This is also true for modeling other traffic circles. If the traffic circle has flow or network configuration characteristics that the model has not "seen," it theoretically cannot produce acceptable results.
- The errors generated by Method 3 are only slightly larger than those generated by the first method. From the results presented in Figures 6 and 8, it can be concluded that the neural network models do not particularly "memorize" patterns in the training data sets. In Method 3, 28 similar neural network models were developed. However, the synthesized O-D traffic volumes from these 28 neural network models were grouped and averaged for morning, noon, and afternoon periods. The average synthesized O-D volumes were very close to the actual O-D traffic volume averages.
- Figure 9 and Table 1 contain results generated by Method 4. The O-D volumes synthesized by the fourth method were grouped and averaged according to time period and were compared with the corresponding actual average O-D volumes. The errors generated by Method 4 are greater than those generated by the first and third

TABLE 1 Mean Absolute Error and Mean Absolute Average Error for Four Training Methods

Training Method	Time Period Used		Error	
	For Training	For Synthesizing Matrices	Mean Absolute	Mean Absolute Average
Method 1	Morning	Morning	2.94	0.93
	Noon	Noon	5.09	1.46
	Afternoon	Afternoon	2.62	0.19
Method 2	Morning	Noon	10.99	7.32
		Afternoon	5.18	4.03
	Noon	Morning	7.32	5.45
		Afternoon	5.32	3.72
	Afternoon	Morning	4.72	3.39
		Noon	9.24	5.79
Method 3	Morning	Morning	3.81	1.25
	Noon	Noon	6.64	1.54
	Afternoon	Afternoon	4.10	0.75
Method 4	Morning	Morning	3.41	1.92
	Noon	Noon	7.33	2.04
	Afternoon	Afternoon	3.37	1.70

methods, but these results are still quite acceptable. This indicates that a neural network model having the general form presented in Figure 5 is able to recognize variations among the morning, noon, and afternoon O-D traffic flows.

- According to the data in Table 1, the neural network models presented in this paper generated larger errors during the noon period. In other words, the neural network models were consistently less accurate in synthesizing O-D volumes for the noon period than for the morning and afternoon periods. This is probably due to the fact that the O-D traffic flow patterns have more variations during the noon period (as demonstrated earlier in this paper) than during the morning and afternoon periods, which consist mostly of work-related traffic.
- Figure 9 clearly indicates that a neural network model having the general form presented in Figure 5 can recognize traffic pattern variations among the morning, noon, and afternoon peak periods. Thus, the final conclusion is that such a neural network model can synthesize adequate O-D traffic volumes as long as the model has been trained with O-D traffic volumes that cover all the anticipated traffic patterns. In other words, as long as the model has "seen" similar data sets, it can recognize variations in traffic patterns and synthesize reasonable O-D traffic volumes.

SUMMARY

The neural network model developed using the backpropagation method can be used to synthesize O-D traffic flows for traffic circles. However, this conclusion has been reached by analyzing results of a few neural network models based on traffic data collected from one traffic circle for 7 hr. More data from this and other traffic circles are needed to verify further the findings of this study.

It should be noted that although this study concentrated on synthesizing O-D flows for traffic circles, the proposed model formulation easily could apply to synthesizing O-D information on linear freeway sections where on-ramp, off-ramp, and link traffic volumes are readily available. The model formulation could also apply to synthesizing O-D matrices for an urban street network, on the basis of the observed link traffic volumes. The effectiveness of neural network models in synthesizing O-D matrices based on the observed link traffic volume, however, might depend on the geometry of the network and the availability of link traffic volumes.

Traffic engineers have two ways to apply similar neural network models to synthesize O-D flows based on traffic volumes from entering and exiting streets. One is the procedure presented in this summary. Traffic engineers could collect actual O-D traffic flow information for morning, noon, and afternoon periods. Then a single model for all periods or multiple neural network models for individual periods could be generated. Next, any future O-D traffic flow data could be synthesized on the basis of traffic flow information collected at the entrances and exits of a traffic circle.

Neural network models can be used to synthesize origin and destination information for traffic circles in a second way. As discussed, neural network models are unable to synthesize traffic flow data for patterns for which they have not been trained. Also, one combined model can adequately synthesize O-D traffic flow as long as the model has "seen" the given pattern before. Thus, it is conceivable that one can devise a procedure that simulates a set of O-D traffic flow conditions that will cover all actual traffic patterns. The neural network model based on training using such simulated data should be able to synthesize O-D traffic flows as long as the simulated training data sets cover actual traffic patterns. Thus, the proposed neural network model should be able to "learn" traffic circle traffic patterns based on the simulated information and reduce the data collection task. If the model could "learn" the "rules" from simulated data and make inferences about actual information collected from streets, the neural network model presented in this paper would be an actual intelligent model.

REFERENCES

1. Müller, B., and J. Reinhardt. *Neural Networks—An Introduction.* Springer-Verlag, Berlin, Germany, 1990.
2. Hertz, J., A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Company, 1991, pp. 115–120.
3. Yang, H., T. Akiyama, and T. Sasaki. A Neural Network Approach to the Identification of Real Time Origin-Destination Flows from Traffic Counts. *Proc., International Conference on Artificial Intelligence Applications in Transportation Engineering*, 1992, pp. 253–269.

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

Estimating Destination-Specific Traffic Densities on Urban Freeways for Advanced Traffic Management

GARY A. DAVIS AND JEONG-GYU KANG

A continuous-time Markov compartment model of freeway traffic flow is presented and tested using simulated and real data. By using the method of large population approximation, the underlying stochastic process is approximated by the sum of a nonlinear deterministic process and a linear, time-varying Gaussian stochastic process. With this approximation a Kalman filter that tracks the density of a freeway section, broken down by destination, was derived. The filter was then tested using simulated data and actual freeway data obtained from Interstate 35W.

Advanced traffic management systems (ATMS) seek to combine an understanding of traveler route selection with improved real-time monitoring of traffic networks in order to alleviate the effects of traffic congestion without requiring substantial new roadway capacity. In particular, driver information and route guidance systems attempt to maximize existing roadway capacity by informing drivers of under- and overused routes or of temporary reductions in capacity.

The effective use of route guidance and driver information, however, requires the ability to forecast driver reactions, their tendencies to select new routes, departure times, modes, and so on in response to information; a number of researchers have developed models aimed at producing such forecasts. In principle, route diversion can be forecast using route selection models common in traffic assignment, but unlike traditional traffic assignment, for short-term (i.e., within-peak) forecasts, it generally will be insufficient to know the origin-destination (O-D) pattern of the traveling public. This is because a substantial component of the traffic, say, 15 min into the future will be composed of vehicles that were already en route when the information or guidance was made available.

Since route selection behavior depends on the particular origin and destination between which a driver is traveling, real-time diversion forecasting will require knowing the breakdown, by O-D pair, of the number of vehicles on each link of a roadway network. A simplification occurs when drivers can be assumed to follow a Markovian routing rule, in which one's future path depends only on one's destination and current location in the network. This condition occurs when modeling simple freeway sections or when route choice follows a logit assignment principle. In this case, knowing the number of vehicles and their distribution across destinations on each link of the road network is necessary for forecasting future route selection activities (*1,2*). For demand forecasting purposes, a vector containing these destination-specific vehicle counts can be considered the state of the traffic system.

Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive, S.E., Minneapolis, Minn. 55455.

Unfortunately, almost all traffic sensors provide data, such as the traffic volumes and lane occupancies provided by magnetic loop detectors, that are aggregated across the network's O-D-specific subflows, so that the traffic state must be estimated rather than measured directly. This is a filtering problem, which can be solved using the results of modern systems theory if one has at hand a unified, real-time model of traffic flow and assignment. Such models can be constructed using a class of stochastic process models called Markov compartment models (*1,3*). This paper describes the development and testing of such a model for traffic flow on a segment of urban freeway.

MARKOV COMPARTMENT MODEL OF FREEWAY TRAFFIC FLOW

A compartmental system is defined as "a system which is made up of a finite number of macroscopic subsystems, called compartments, each of which is well mixed, and the compartments interact by exchanging materials. There may be inputs from the environment into one or more of the compartments, and there may be outputs from one or more of the compartments into the environment" (*4*). Karmeshu and Pathria (*5*) proposed a Markov compartment model for highway traffic and provided an asymptotic analysis using a diffusion approximation. Here the material is composed of vehicles, and the stochastic nature of material transfer is caused by the random movement of vehicles according to a continuous-time Markov process. Now imagine that a segment of freeway has been divided into sections, such that on-ramps join the freeway only at the upstream boundaries of sections, off-ramps diverge from the freeway only at the downstream boundaries of sections, and mainline detectors are located at the downstream boundaries of sections. In addition, the number of lanes, grade, and other geometric characteristics are constant within the section.

Assume that the freeway has m origins, indexed by $i = 1, \ldots, m$; s destinations, indexed by $j = 1, \ldots, s$; and n sections, indexed by $k = 1, \ldots, n$. By convention, origin 1 is taken to be the upstream mainline boundary of the original freeway segment, while destination s is taken to be the downstream mainline boundary. Next, define the following variables:

$x_{oi}(t)$ = total remaining vehicles at origin i at time t,
$x_{dj}(t)$ = total vehicles that have exited the segment at destination j by time t,
$x_{kj}(t)$ = vehicles in section k destined for j at time t,
$y_l(t)$ = total vehicles counted at counter l up to time t.

Assume that the total number of vehicles in the system is fixed, so that

$$N = \Sigma_i x_{oi}(t) + \Sigma_k \Sigma_j x_{kj}(t) + \Sigma_j x_{dj}(t)$$

is constant at all times t. Let

$$\mathbf{x}(t)\ [x_{o1}(t), \ldots, x_{om}(t), x_{11}(t), x_{12}(t), \ldots, x_{ns}(t), x_{d1}(t), \ldots, x_{ds}(t)]^T$$

be a column vector containing the various compartment populations, and

$$\mathbf{y}(t) = [y_1(t), \ldots, y_p(t)]^T$$

be a column vector containing the count totals. Letting $\mathbf{e}_g$ denote a column vector with all elements equal to 0 except for position g, which is 1, and letting g, h index arbitrary elements of the vector $\mathbf{x}$, it will be assumed that over a very short time interval of length Δ, transitions of the form

$$\begin{bmatrix} \mathbf{x}(t+\Delta) \\ \mathbf{y}(t+\Delta) \end{bmatrix} - \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix} = \begin{bmatrix} \mathbf{e}_h - \mathbf{e}_g \\ \mathbf{H}\mathbf{e}_g \end{bmatrix} \tag{1}$$

occur with probability $x_g q_{g,h}[\mathbf{x}(t)]\Delta + o(\Delta)$, transitions with

$$[\mathbf{x}(t+\Delta)^T, \mathbf{y}(t+\Delta)^T]^T - [\mathbf{x}(t)^T, \mathbf{y}(t)^T]^T = \mathbf{0}$$

occur with probability $1 - \Sigma_{h \neq g} x_g q_{g,h}[\mathbf{x}(t)]\Delta + o(\Delta)$, and all other transitions have a probability that is $o(\Delta)$. Note that a $\mathbf{x}(t+\Delta) - \mathbf{x}(t) = \mathbf{e}_h - \mathbf{e}_g$ corresponds to the transition of a vehicle from compartment g to compartment h. By defining

$$\mathbf{H}_{lg} \begin{cases} = 1 \text{ if counter } l \text{ registers departure from } g \\ = 0 \text{ otherwise} \end{cases}$$

$\mathbf{y}(t+\Delta) - \mathbf{y}(t) = \mathbf{H}\mathbf{e}_g$ corresponds to an increment in the counter registering departures from g. The vehicle movements follow a closed, continuous-time Markov compartment model (or, equivalently, a nonlinear birth and death process), with the state vector augmented to include vehicle counts. The problem then is to use the counts at time t to produce estimates of the unobserved segment populations $x_{kj}(t)$. When the transition intensities $q_{g,h}[\mathbf{x}(t)]$ are not constant, the resulting filtering problem will be nonlinear and often intractable. Fortunately, given reasonable conditions on the functions $q_{g,h}[\mathbf{x}(t)]$, Lehoczky's argument (*6*) can be adapted to this case to show that as N, the total number of vehicles in the system, becomes large, the stochastic evolution of the random vectors $[\mathbf{x}(t)^T, \mathbf{y}(t)^T]^T$ can be approximated by the sum of a nonlinear deterministic process and a linear, time-varying Gaussian stochastic process. That is,

$$\begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix} \approx \begin{bmatrix} \bar{\mathbf{x}}(t) \\ \bar{\mathbf{y}}(t) \end{bmatrix} + \mathbf{z}(t) \tag{2}$$

where the deterministic, mean value process satisfies the ordinary differential equation

$$\frac{d\bar{x}_g(t)}{dt} = \Sigma_h \bar{x}_h(t) q_{h,g}[\bar{\mathbf{x}}(t)]$$

$$\frac{d\bar{y}_l(t)}{dt} = \Sigma_g H_{lg} \bar{x}_g(t) \Sigma_{u \neq g} q_{g,u}[\bar{\mathbf{x}}(t)] \tag{3}$$

and $\mathbf{z}(t)$ is a zero-mean, Gaussian random vector with covariance matrix $\mathbf{P}(t)$, which evolves according to the Ricatti equation

$$\frac{d\mathbf{P}(t)}{dt} = \boldsymbol{F}[\bar{\boldsymbol{x}}(t)]\,\boldsymbol{P}(t) + \boldsymbol{P}(t)\boldsymbol{F}[\bar{\boldsymbol{x}}(t)]^T + \boldsymbol{G}[\bar{\boldsymbol{x}}(t)] \tag{4}$$

Here $\mathbf{F}(.)$ denotes the Jacobian matrix of the right-hand side of Equation 3 with respect to $\mathbf{x}(t)$, while $\mathbf{G}(.)$ is a covariance term that depends only on $\mathbf{x}(t)$.

Given initial estimates $\mathbf{x}(0)$, $\mathbf{P}(0)$, $\mathbf{y}(0) = \mathbf{0}$, Equations 3 and 4 can be solved numerically to give approximate expected compartment totals and cumulative counts, along with variances and covariances for any future time t. When actual counts become available at some time T_k, the standard formulas for the Kalman filter (*7*) can be used to give a measurement update of compartment totals and their covariance terms. Equations 2 and 3 can then be restarted with $\mathbf{x}(0) = \mathbf{x}(T_k)$, $\mathbf{P}(0) = \mathbf{P}(T_k)$ and $\mathbf{y}(0) = \mathbf{0}$, and the recursion continued until the next count becomes available.

DETERMINING TRANSITION RATES

Implementation of the filter requires that appropriate functions are selected for the transition intensities. For the transitions from the origin sources to mainline sections, it is convenient to use transition intensities of the form $q_{oi}b_{ij}$, where q_{oi} equals the constant arrival intensity from on-ramp i, and b_{ij} is the probability that a vehicle is destined for off-ramp j, given it arrives at on-ramp i.

If the origin populations $x_{oi}(t)$ are large enough so that the number of total arrivals during the time period of interest is a small proportion of the original total, the quantity $x_{oi}(t)q_{oi}$ can be taken as a constant λ_{oi}, giving Poisson arrival rates at the freeway origins.

To obtain functions giving the transition rates within the mainline sections, assume that at time t the vehicles in section k have speeds assigned as independent, identically distributed random outcomes from a common speed distribution, and that distances from the downstream boundary of section k are assigned as independent, identically distributed outcomes from a uniform random variable with probability density $1/L_k$, where L_k is the length of section k. It is then straightforward to show that the probability of a randomly selected vehicle exiting section k during a short interval of length Δ is simply $\bar{u}_k(t)\Delta/L_k$, where $\bar{u}_k(t)$ gives the space-mean speed for section k at time t. The formulation can then be closed by requiring the space-mean speeds $\bar{u}_k$ to depend directly on $\mathbf{x}(t)$ via a form for the equilibrium speed-density relations of traffic flow theory, giving a version of the simple continuum model. As formulated though, this model will tend to lock up when the densities in a section rise above the critical density (*8*).

Although Markov traffic models can be extended to produce analogs of higher-order continuum models (*3,9*), a simpler solution is to use a device originally attributable to Szeto and Gazis (*10*) and allow the flow across the boundary of two sections to depend on both the upstream and downstream densities. The two-dimensional per lane flow-density relation (transition rate) used in this paper takes the form

$$Q(d_k, d_{k+1}) = d_k u_0 e^{-1/2(d_k/d_c)2} \left[1 - \left(\frac{d_{k+1}}{d_j}\right)^r\right] \qquad d_k \leq d_c$$

$$Q_0\left[1 - \left(\frac{d_{k+1}}{d_j}\right)^r\right] \qquad d_k > d_c \tag{5}$$

where

u_o = free-flow speed,
d_c = critical density,
Q_0 = capacity flow, and
d_j = jam density.

For a constant downstream density d_{k+1}, Equation 5 gives an increasing cross-boundary flow as the upstream density d_k increases, up to the point at which d_k equals the critical density. The cross-boundary flow then remains constant, thus modeling the upstream section as (approximately) an oversaturated finite-server queue. As the downstream density d_{k+1} approaches the jam density d_j, the cross-boundary flow goes to 0, with the sensitivity of this effect being governed by the exponent r. Figure 1 displays a plot of Equation 5 as calibrated for an actual segment of freeway.

The continuous-time Markov compartment (MARCOM) freeway traffic flow simulation model incorporating transition intensities can be expressed as the following simple recursive process, well-suited for computer simulation:

- *Step 0*: Given O-D splitting probabilities b_{ij} and destination-specific variables $x_{kj}(0)$, let $t = 0$; $i = 1, \ldots, m$; $k = 1, \ldots, n$; $j = 1, \ldots, s$.
- *Step 1*: Generate the next arrival time at origin i destined for j, Δ_{ij}, as an exponential outcome with parameter $\lambda_{oi}b_{ij}$, λ_{oi} = arrival rate at on-ramp i.
- *Step 2*: Calculate the mainline transition rates $x_{kj}q_{k,h}(t)$ from Equation 5, $q_{k,h}(t)$ = mainline transition intensity.
- *Step 3*: Generate the next transition time at each compartment k destined for j, Δ_{kj}, as an exponential outcome with parameter $x_{kj}q_{k,h}$.
- *Step 4*: Pick a minimum next arrival time $\Delta_{\min}$ among $(\Delta_{ij}, \Delta_{kj})$.
- *Step 5*: Let $t = t + \Delta_{\min}$, update state variable $x_{kj}(t)$:

$x_{kj}(t + \Delta_{\min}) = x_{kj}(t) + 1$, if it is a birth compartment;
$x_{kj}(t + \Delta_{\min}) = x_{kj}(t) - 1$, if it is a death compartment; and
$y_{lg}(t + \Delta_{\min}) = y_{lg}(t) + 1$, if detector l register departures, from g.

- *Step 6*: Go to Step 1.

PRELIMINARY TESTING OF MARCOM MODEL OF FREEWAY TRAFFIC FLOW

Behavior of MARCOM at a Lane Drop Bottleneck

Although the basic idea behind Equation 5 is not new, the traffic flow model that results is still somewhat novel, and it was first desired to see if Equation 5 could produce reasonable behavior at bottlenecks. To this end, a computer program implementing MARCOM was written and used to generate simulated flows for the hypothetical 3.5-mi freeway section shown in Figure 2. Here, the number of lanes is reduced from three to two behind the fifth of 12 subsections (the length of subsection was uniformly chosen to 1,500 ft). A 60-min simulation started with demand of 3,000 vehicles per hour (vph) and then increased to 4,800 vph, which exceeds the capacity of the two-lane section by approximately 20 percent, and finally decreased to 1,200 vph. The simulation results of this hypothetical case are depicted in Figures 3 and 4, which show the volume and density trajectories of the bottleneck section at 5-min intervals. As illustrated in these figures, the MARCOM provides a reasonable

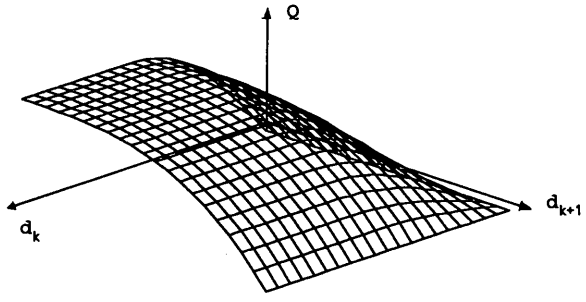


FIGURE 1 Two-dimensional flow-density relationship.

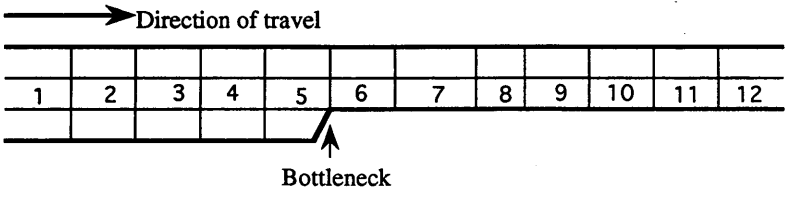


FIGURE 2 Geometrics of freeway section with bottleneck.

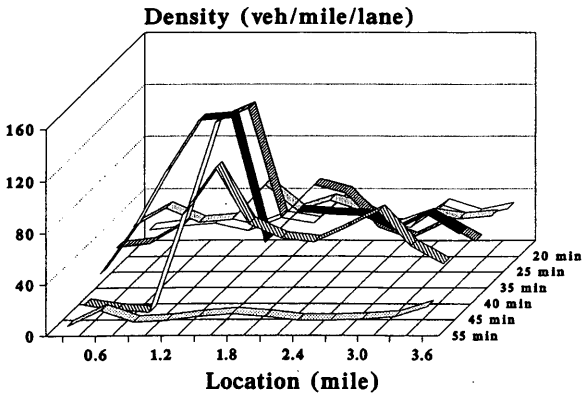


FIGURE 3 Density trajectories at a bottleneck.

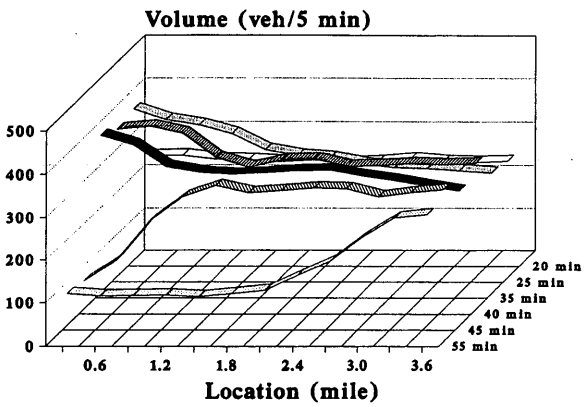


FIGURE 4 Volume trajectories at a bottleneck.

description of queue build-up and dissipation in that (a) congestion starts in front of the bottleneck and moves in upstream direction, while the density within the bottleneck section is around critical density; and (b) the volumes in the bottleneck section are limited to the capacity during congestion building and dissipation.

Calibration and Verification of MARCOM

As stated earlier, the ultimate objective of this research is to estimate destination-specific traffic densities on freeways. The solution strategy was to describe a Markovian traffic model, approximate the Markovian model with a linear stochastic model, and apply the theory of Kalman filtering to the linear model in order to estimate the destination-specific densities. Three questions then arise concerning this approach: (a) how reasonable is the underlying Markovian traffic model? (b) how accurate is the linear approximation? and (c) how well does the resulting Kalman filter perform? Since destination-specific densities are almost impossible to observe in practice, the accuracy of the Kalman filter must be assessed using simulated data. To this end, the MARCOM simulation program just described was calibrated using real data and run for model verification. Figure 5 depicts a seven-origin, four-destination segment of northbound Interstate highway I-35W that is 4.0 km (2.5 mi) long. Five-minute cumulative volume and lane occupancy measurements during a 3-hr morning peak period (6:00 to 9:00 a.m.) for mainline, on-ramp, and off-ramp stations were obtained from the Minnesota Department of Transportation (MNDOT).

To run the stochastic simulation model, MARCOM, it is necessary to know the on-ramp arrival rates λ_{oi}, the O-D splitting probabilities b_{ij}, and the parameters governing the flow-density relation in Equation 5. The arrival rates can simply be estimated as those values that reproduced the corresponding 5-min arrival counts allowing the arrival rates to vary for each 5-min interval.

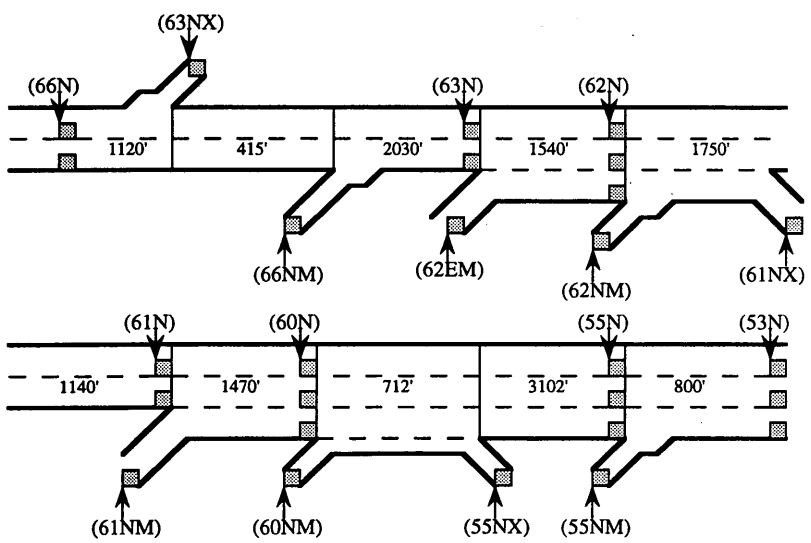


FIGURE 5 Geometrics of test section (I-35W northbound).

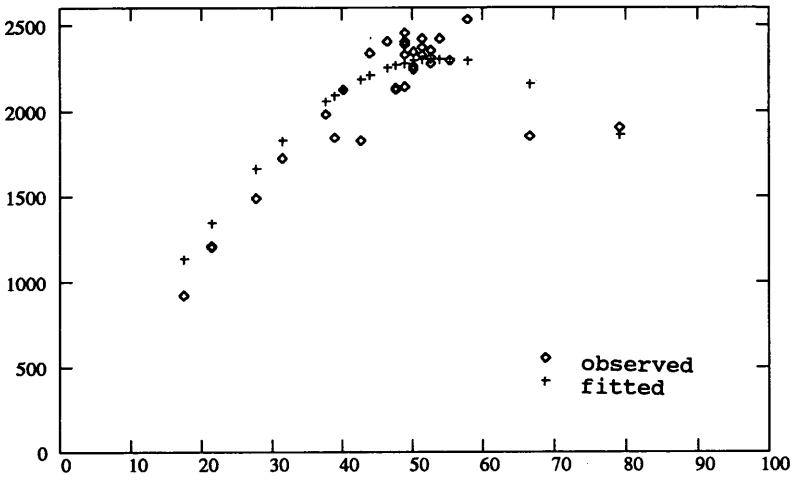


FIGURE 6 Fitted and observed steady-state flow versus density.

To determine the parameters for Equation 5, the lane occupancy measurements were converted to approximate density values and the parameters u_0, d_c, d_j, and r in Equation 5 were estimated using nonlinear least squares by setting $d_k = d_{k+1}$, corresponding to the notion of approximate homogeneous flow. Figure 6 shows the observed and fitted flow-density curve obtained for the estimates $u_0 = 66.6$ mph, $d_c = 64.5$ veh/lane/mi, $d_j = 120$ veh/lane/mi, and $r = 3$. Finally, the splitting probabilities b_{ij} were estimated by using the estimated traffic flow parameters to numerically solve the mean value Equation 3 given a trial set of b_{ij} values. For a given set of origin counts, this produced estimated destination counts, and those b_{ij} values that minimized the sum of squared errors between forecast and actual counts were obtained by embedding this routine in a nonlinear optimization program. This method produced reliable O-D parameter estimates in a recent research (*3*) when incorporating an accurate traffic flow model. These estimates were then used as inputs to a MARCOM that simulated the Markov compartment process described earlier to generate simulated traffic counts for various time intervals as well as destination-specific section populations, $x_{kj}(t)$.

The resulting comparisons of volume and density generally indicated good agreement between simulated and actual data. In order to evaluate the model performance quantitatively, two error measurements (mean absolute percentage difference and mean absolute error) are calculated. As indicated by the error measures in Table 1, MARCOM provided a reasonable reproduction of traffic flows.

TABLE 1 Mean Error of Simulated Volume Results (6:00 to 9:00 a.m.)

Detector Station	63N	62N	61N	55N	53N
MAPD[a]	2.2	2.0	2.0	2.1	2.2
MAE[b]	13	21	18	27	28

[a]Mean Absolute Percentage Difference (%) = $\sum_{k=1,N}(100*(\text{Measured-Simulated})_k/\text{Measured}_k)/N$

[b]Mean Absolute Error (veh/5 min) = $\sum_{k=1,N}(\text{Measured-Simulated})_k/N$ where N is the Number of Measured Points

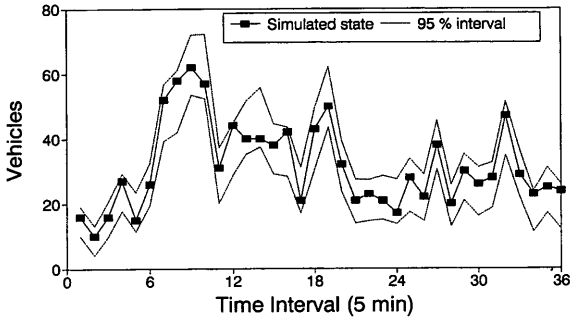


FIGURE 7 Simulated state and confidence interval (state variable: ×54).

ESTIMATING DESTINATION-SPECIFIC VEHICLES USING SIMULATED COUNT DATA

Next, the estimated parameter values were used to implement the density-tracking Kalman filter for the segment of I-35W depicted in Figure 5. First, instantaneous destination-specific volume counts at the end of every 5-min interval and 5-min cumulative volume counts at designated detectors were generated by MARCOM. Next, the Kalman filter was used to estimate destination-specific densities using simulated volume counts.

Figures 7 and 8 show the simulated destination-specific traffic densities along with the approximate 95 percent confidence produced by the Kalman filter. The two dotted curves in Figures 7 through 10 indicate the two-standard deviation envelope produced by the Kalman filter. In each case the estimation range tracks the simulated values reasonably well, with the larger volume flows being tracked somewhat better. This indicates that the filter is performing properly, although it is an approximation of the original process.

Finally, as an additional test of the model's accuracy, the Kalman filter was used to generate predicted mainline and off-ramp counts when fed by actual on-ramp counts. Figure 9 shows actual mainline counts along with the 95 percent prediction range for one of the detector stations, and Figure 10 compares actual ramp counts along

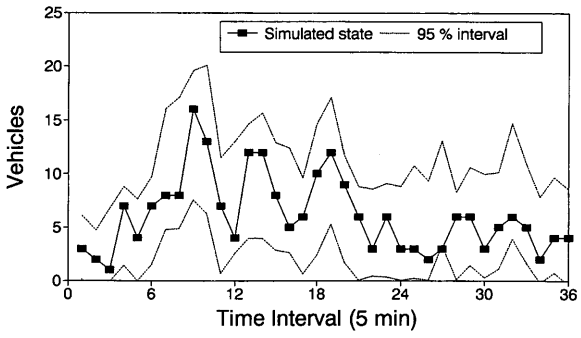


FIGURE 8 Simulated state and confidence interval (state variable: ×52).

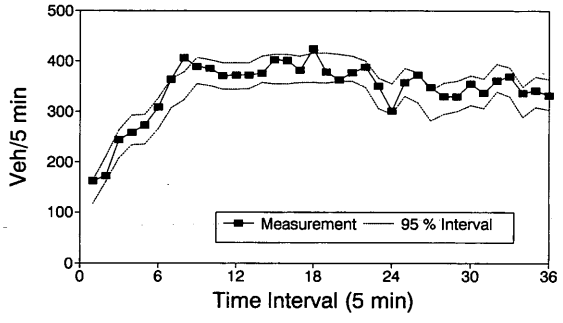


FIGURE 9 Actual volume and confidence interval (mainline: Station 61N).

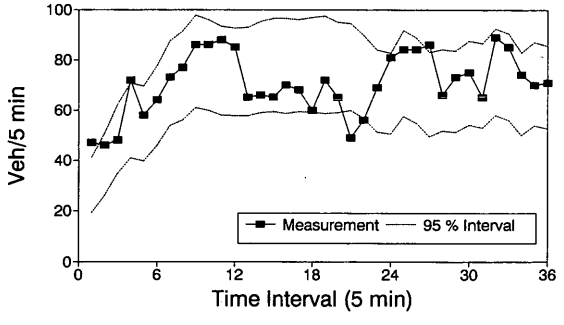


FIGURE 10 Actual volume and confidence interval (off-ramp: Station 61NX).

with the 95 percent prediction range. The mainline volumes are tracked reasonably well, and the Kalman filter appears to predict the mean value of the off-ramp count with accuracy but not its fluctuations as well as desired.

CONCLUSION

This paper began by arguing that a destination-specific breakdown of the traffic currently on a road network is an essential input to any route diversion forecasting method but that such information cannot be measured directly by existing surveillance systems. For the more tractable case of freeway segments, a Kalman filter that could produce such estimates was derived from a Markov compartment model of traffic flow and tested using data from an existing segment of freeway. Generally, the Markov model provided a reasonable description of freeway traffic flow, and the Kalman filter produced reasonable estimates of the destination-specific volumes, although for the lower-volume subflows, the proportion of error was somewhat greater.

Overall, it appears that this Kalman filtering approach provides the information needed for real-time diversion forecasting, at least for freeway segments. One obvious line of improvement would be to develop an adaptive filter by replacing the off-line parameter estimation procedures with their recursive equivalents. This would permit the tracking of slowly varying changes in the O-D pattern and possibly improve the accuracy shown in Figure 10. The main challenge, however, is to extend this filtering method to general traffic networks; this effort is currently under investigation.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation and in part by the Center for Transportation Studies at the University of Minnesota.

REFERENCES

1. Davis, G. Integrated Traffic Assignment and Flow Models Via Markovian Networks. Presented at 33rd ORSA/TIMS Joint National Conference, Orlando, Fla., 1992.
2. Papageorgiou, M., and A. Messmer. Dynamic Network Traffic Assignment and Route Guidance Via Feedback Regulation. In *Transportation Research Record 1306*, TRB, National Research Council, Washington, D.C., 1991, pp. 49–58.
3. Davis, G. Estimating Freeway Origin-Destination Parameters and Impact of Uncertainty on Ramp Control. *Journal of Transportation Engineering*, ASCE, Vol. 119, No. 4, 1993, pp. 489–503.
4. Jacquez, J. A. Kinetics of Distribution of Tracer-Labeled Materials. In *Compartmental Analysis in Biology and Medicine*. Elsevier Publishing Co., New York, 1972.
5. Karmeshu and R. K. Pathria. A Stochastic Model for Highway Traffic. *Transportation Research*, Vol. 15B, No. 4, 1981, pp. 285–294.
6. Lehoczky, J. Approximations for Interactive Markov Chains in Discrete and Continuous Time. *Journal of Mathematical Sociology*, Vol. 7, 1980, pp. 139–157.
7. Gelb, A. *Applied Optimal Estimation*. Analytic Sciences, Cambridge, Mass., 1974.
8. Ross, P. Traffic Dynamics. *Transportation Research*, Vol. 22B, No. 6, 1988, pp. 421–435.
9. Cremer, M., and A. May. *An Extended Model of Freeway Traffic*. Research Report UCB-ITS-RR-85-7. Institute of Transportation Studies, University of California, Berkeley, 1985.
10. Szeto, M., and D. Gazis. Application of Kalman Filtering to the Surveillance and Control of Traffic Systems. *Transportation Science*, Vol. 6, 1972, pp. 419–439.

All opinions and conclusions expressed here are solely the responsibility of the authors.

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

Estimation of Speeds from Single-Loop Freeway Flow and Occupancy Data Using Cusp Catastrophe Theory Model

ANNA PUSHKAR, FRED L. HALL, AND JORGE A. ACHA-DAZA

Many freeway management systems rely on single-loop detectors, which can measure only flow and occupancy, for information on freeway operating conditions. Although it is possible to estimate average speeds from those data by assuming a constant vehicle length, such estimates are not particularly good. The catastrophe theory model for these variables provides an alternative procedure for estimating average speeds. To apply it, different procedures are needed to calibrate the model. Such procedures are developed and their generality is tested, by applying them first across different days at the same (double-loop) station, then to other double-loop stations, and finally to single-loop stations. The first two tests allow for direct comparisons with measured average speeds; the final comparison can be made only with other estimated speeds, which is done on the basis of speed-flow diagrams. The results suggest that the catastrophe theory estimates are better than those made assuming a constant vehicle length. Estimates based on concurrent nearby measured vehicle lengths are similar to the catastrophe theory estimates on average, but the former overestimate the scatter and the latter underestimate it.

Many current freeway traffic management systems rely on single-loop detectors, which measure only volume and occupancy and can estimate only average speeds. Such systems would benefit from a model that could provide better estimates of average speed from these measurements, thereby avoiding the higher costs of installation and operation of double-loop detectors. Indeed, reliable values of traffic variables, and especially of system speeds or travel times, are becoming increasingly important as an input for intelligent vehicle-highway systems.

A recent paper compared calculations for speeds of 30-sec flow-occupancy observations on freeways given by catastrophe theory with those produced by a number of other traffic flow models (*1*). It was concluded that for the particular data sets used, the catastrophe theory model was better than any of the others. This paper picks up on that conclusion and addresses the question of whether the catastrophe theory model can be used in a freeway management system to provide reliable estimates of speed from single-loop detectors. This paper also attempts to resolve two difficulties noted in that earlier paper. First, the parameters used in the modeling depended heavily on specific extreme observations in each data set being modeled. Second, the model did not work well in all situations; in particular, it did not provide good predictions for queue discharge flow (i.e., where vehicles are accelerating away from a congested or stop-and-go situation).

A. Pushkar and F. L. Hall, Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7. J. A. Acha-Daza, Department of Civil Engineering, University of Texas, Austin, Tex. 78712.

BACKGROUND

The background discussion contains three components. First is a brief review of approaches currently used on freeway management systems for obtaining average speeds. Second, the catastrophe theory model is discussed briefly, with particular reference to Acha-Daza's methods for parameter identification (*2*) and potential changes. Third, the research task for this paper is explained.

Today the most widely used method for estimating average speeds at single-loop stations is based on the use of an average vehicle length. It can be shown that space-mean speed can be calculated on the basis of the equation

$$\text{sms} = \frac{\text{vol} \times 100 \times \text{length}}{\text{occ} \times T} \quad (1)$$

where

sms = space-mean speed (m/sec),
vol = volume measured over time T,
length = average vehicle length plus effective detector length,
occ = occupancy (%), and
T = interval length (20 sec on Highway 401).

Some errors are intrinsic to the use of a constant vehicle length, one of which relates to the effect of the variance of vehicle lengths (*3*). This difficulty is compounded by the fact that the amount of variation in vehicle lengths (and therefore in the error in using Equation 1) varies over lanes and over the day. In this study, this variance is minimized, as data are taken from only the leftmost lane, from which trucks are restricted.

A second method is available for those systems with some double-loop stations among the single-loop ones, namely, using the double-loop data to estimate vehicle lengths that will be arriving at the adjacent single-loop station in the next time interval. Such a system is used on the Highway 401 COMPASS freeway management system.

A third alternative, the catastrophe theory model, was first applied to traffic operations by Navin (*4*) and has subsequently been developed by Hall et al. (*5–8*). The cusp catastrophe model used in these applications can be represented as the three-dimensional surface given by

$$X^3 + aUX + bV = 0 \quad (2)$$

In converting this mathematical form to the traffic flow situation, it has been conventional to associate speed with the variable X, volume with U, and occupancy with V.

Acha-Daza's advance (*2*) on earlier work was to apply an axes rotation, as was first suggested by Forbes and Hall (*7*), along with an axes translation as proposed by Dillon and Hall (*6*). In doing the translation, Acha-Daza set the origin for the new axes system at the point of maximum observed flow, and the maximum observed occupancy occurring at that flow. The new axes were then rotated about this origin. For simplicity, the flow and occupancy values that determine this new origin will be referred to as the pivot point. The angle (theta) for the rotation was selected as that angle which minimized the number of misclassified observations, that is, the number of data that were classified as congested on the basis of a critical speed but uncongested (left of the *U*-axis) after the rotation, and vice versa. Acha-Daza selected the minimum speed observed at the pivot point as this critical speed. Acha-Daza's paper also uses a data-specific graphical factor to provide numbers of similar magnitude for calculating the trigonometric functions on which the rotation was based.

There are two problems with this method for selecting the pivot point, critical speed, and graphical factor. First, the parameter values are extremely dependent on the particulars of the data set being used. The overall goodness of fit of the model may well depend on whether the particular set of data being analyzed had observations near the "true" pivot point. Second, selecting parameters using the method identified by Acha-Daza makes it difficult, if not impossible, to generalize the results, and especially to identify appropriate parameters for speed estimation at single-loop stations. Hence one of the tasks in this paper is to designate a method for identifying the pivot point and critical speed that is not so data-specific and can therefore produce a more general set of parameters. This problem of data-dependent parameters is not unique to catastrophe theory models: Ceder (*9–11*) encountered similar problems with more conventional models, and data reported by Koshi (*12*) suggest the same difficulties. Four parameters—pivot point occupancy and volume, critical speed, and graphical factor—need to be set via experimentation. Given these four, the remaining three parameters—theta and the values of a and b in Equation 2—are found via analytical procedures.

The choice of best parameters is not simple, even with measured speeds from the paired loops available with which speed estimates from the model can be compared. The difficulty arises because the mean error obtained from the use of the catastrophe theory model is not 0. For this reason several goodness-of-fit measures are used: the mean of the error (the difference between the observed and predicted speed), the standard deviation of the error, and the average of the square of the errors (labeled average difference squared). This last measure is the sum of the squares of the previous two, and it is useful because the other two do not always reach their minima at the same set of parameter values. In addition to these numerical results, plots of predicted speeds versus observed speeds help in discovering trends and biases.

The task of evaluating estimates for single-loop stations is more difficult. Because there are no observed speeds with which to compare estimated speeds, the statistical measures just described cannot be used. Instead, estimated speed-volume plots based on catastrophe theory were compared with similar plots based on other estimation methods. However, it will still not be known which best indicates the realism of the speed estimates. To resolve this difficulty, the estimated speed-volume plots have also been compared with similar plots of observed variables from adjacent double-loop stations.

To summarize, then, there are two specific tasks in this paper. The first is to find a different way of calibrating the model, one that does not rely as heavily on extreme observations in a specific data set. The second task is to investigate how well the resulting parameters can be transferred or generalized—especially to single-loop stations.

AVAILABLE DATA

The data used for this study were made available by the Ministry of Transportation of Ontario (MTO) through its COMPASS traffic management system for Highway 401 in Toronto, which measures volume and occupancy over 20-sec intervals via inductive loops, located at stations 300 to 800 m apart on the freeway. Double-loop stations (roughly one-third of the total) measure speeds over the approximately 4.5 m between the loops. At single-loop stations, average speeds over the 20 sec are estimated using vehicle lengths obtained during the previous time interval at the next upstream double-loop station. In the rest of the discussion, the speed to be estimated is the average speed over the 20-sec interval.

This study involved the analysis of data from only the leftmost lanes on three-lane sections of the express systems. These data were collected for clear spring days in 1993 (May 12, June 3, and June 7). Although there were some differences in grade between the stations chosen, none was particularly steep. Results from five stations are included here (four westbound, one eastbound). Data from six other stations were also analyzed; the results support those presented here. Station identifiers used in this paper are a short form of the 12-character MTO station identifier. The MTO numbering starts from the control center, counting outward both eastbound and westbound. For example, E01W specifies the first station east of the traffic control center, in the westbound roadway.

The westbound stations discussed here lie in the region between Keele Street and Highway 400. W03W and W04W show queue discharge when the transfer lanes from the collector system between W02W and W03W cause upstream congestion. E02W, W02W, and W04W are double-loop stations; W03W is a single-loop station. Data for these stations were collected from 3:00 to 8:00 p.m. The eastbound station, W07E, is a double loop and shows queue discharge flow (QDF) when an entrance ramp creates a bottleneck situation upstream. Its data were collected from 5:00 to 10:00 a.m.

CALIBRATION TO PRODUCE GENERAL PARAMETERS

This section deals with calibration: what is a good general set of parameter values, and how can they be identified? The focus is primarily on selecting an appropriate combination of critical speed and pivot point. Given these values, the methods described by Acha-Daza and Hall (*1*) have been used to find theta and the coefficients a and b. In identifying good parameters, one regular station has been used for the primary analysis. Acha-Daza found problems in applying the model to stations that exhibit QDF (*2*); hence, data for one such station have also been investigated.

Exploratory analyses conducted prior to the ones reported here showed that the pivot point did not need to coincide with the highest observed volume, as was the previous practice. Further, in many cases it *should not* coincide with that point, because often the point with the highest volume does not fall on the line defining the bound-

ary between the congested and uncongested data. A pivot point that would always fall along the dividing line was needed.

If the maximum volume does not necessarily indicate the best pivot point, one might consider using points with lower volumes. One implication of lowering the pivot volume is that some of the transformed points would then have positive *U*-values (i.e., above the cusp of the catastrophe model). In this region outside the cusp, changes in predicted speed will be gradual rather than sudden. In fact, this might prove to be an advantage. With varying occupancy, perhaps smoother changes in speeds are exhibited at higher volumes (such as within QDF) than at lower volumes.

Recognizing that other parameters were unnecessarily data-dependent, a decision was made about the graphical factor, which was applied to the *V*-axis simply to make the *U*- and *V*-axes of similar scale and units. For simplicity and improved transferability, the graphical factor has been made a constant in this analysis, set at 0.25.

In seeking the best set of parameters, it was decided to investigate them independently—that is, first test critical speed using arbitrary pivot points, then, using the best critical speed, test pivot point values. If the initial arbitrary pivot points prove to be too far from the final pivot point, iteration is possible. For these tests, a single site was selected, Station E02W.

Critical Speeds

From the preliminary investigations, two pivot volumes lower than the maximum observed volume were chosen for the tests of critical speed: 13 and 8 veh/20 sec. Selecting constant occupancies at these pivot volumes would most likely create unnecessarily large numbers of misclassified points, thereby affecting the results. As Gilchrist and Hall (*13*) showed, speeds generally vary in bands that move somewhat diagonally across the volume-occupancy graph. The selection of pivot volume and critical speed together determine what the correct pivot occupancy should be. Pivot occupancies were identified to the nearest 0.5 percent to minimize the misclassified data points as critical speeds were changed.

Tables 1 and 2 show the results of the testing of several critical speeds at the two selected pivot volumes. With a pivot volume of 13 veh/20 sec (Table 1), all three criteria (standard deviation, average difference squared, and mean error) were at their minima for a critical speed of 85 km/hr. For a critical speed of 90 km/hr, however, the three measures are similar to their minima. Hence it may be that the results are not particularly sensitive to the critical speed over some range of values, although clearly a value above 70 km/hr is needed. In that respect, this result is consistent with the new speed-

TABLE 1 Effect of Changing Critical Speeds: Station E02W, May 12, Pivot Volume of 13 veh/20 sec

Critical Speed (km/h)	70	75	80	85	90	95
Pivot Occupancy (%)	16.5	16	15	15	14.5	14
Theta (Degrees)	-18.7	-17.3	-17.0	-17.0	-16.0	-16.0
Misclassified Points	6	3	2	2	3	8
a	273.1	245.9	225	184.7	136.4	95.98
b	8597	12133	15611	21910	29782	38460
Mean Error	11.09	7.28	4.50	**0.97**	-1.92	-4.13
Standard Deviation	9.94	9.14	8.5	**7.98**	8.04	8.58
Average Difference^2	222.0	136.6	92	**64.6**	68.3	90.78

TABLE 2 Effect of Changing Critical Speeds: Station E02W, May 12, Pivot Volume of 8 veh/20 sec

Critical Speed (km/h)	80	85	90	95	100	105
Pivot Occupancy (%)	9.5	9	8.5	8	7.5	7.5
Theta (Degrees)	-15.0	-16.0	-15.0	-14.0	-13.2	-12.9
Misclassified Points	3	2	3	6	9	15
a	277.0	266.8	227.8	191.6	168.5	144.0
b	39185	43389	47337	52344	59245	67531
Mean Error	7.03	3.68	**1.12**	**-1.12**	-2.99	-6.68
Standard Deviation	8.41	7.98	**7.70**	7.71	8.18	8.88
Average Difference^2	120.2	77.04	**60.31**	60.68	75.81	123.46

flow curves in the revised Chapter 3 of the *Highway Capacity Manual* (*14*), which have limiting speeds for uncongested data (i.e., speeds at capacity) ranging from 50 to 60 mph (80 to 100 km/hr).

Given the results of the testing in Table 1, Table 2 uses a higher range of critical speeds for the pivot volume of 8 veh/20 sec. For this volume, critical speeds of 90 and 95 km/hr had the best numerical results, with their measures of error being virtually identical. The results from Table 2 indicate that the best critical speed to use may depend on the particular pivot volume selected but, more importantly, that the error may be relatively insensitive to changes of up to 10 km/hr in the critical speed. For both pivot volumes, a critical speed of 90 km/hr fell in the optimal range. Hence it was chosen as a general value for subsequent analyses.

Varying the Origin

Using the critical speed of 90 km/hr, a range of pivot points was evaluated. Given the pivot volume and critical speed, occupancies were identified that minimized the number of misclassified data. The similarity in theta and in the number of misclassified points (Table 3) confirms that the pivot points lie on approximately the same line. The measures of error all point to a pivot volume of 9 veh/20 sec as being the best value. However, the measures of error for a pivot volume of 10 veh/20 sec are a very close second. Indeed, for the range of 8 to 12 veh/20 sec, the measures of error do not vary more than about 10 percent, and any of those values might be expected to produce reasonable results. Higher values, such as 17 veh/20 sec, are clearly not as good, although they do not have as deleterious an effect as using too low a critical speed (Tables 1 and 2).

Complications When Considering Queue Discharge Flow

Station E02W, used for the analyses in Tables 1 through 3, did not experience QDF but instead had data from within the stop-and-go conditions that constitute a queue on a freeway. In previous work, the catastrophe theory model did not reliably predict speeds at stations exhibiting QDF (*2*). It appeared reasonable to expect that modifying the choice of pivot point might overcome this problem. Several tests were run on one such station (W07E, May 12). It was found that the same critical speed (90 km/hr) produced reasonable results. However, for pivot volume, lower values were optimal: volumes of 5 and 6 minimized the standard deviation, 7 minimized the average difference squared, and 8 minimized the mean error.

In general, any of the volumes from 5 to 10 veh/20 sec produced reasonable results. The error measures are not especially sensitive to small changes of volume within the range of 5 to 10 veh/20 sec, which overlaps with the range of best pivots for E02W (8 to 12 veh/20 sec). Rather than using different parameters for different traffic flow regimes, a single compromise pivot volume was sought to encompass them all. A pivot volume of 10 veh/20 sec was chosen, implying the approximation that has occurred. Although this value is at the top of the range for QDF, comparing predicted and observed speeds for Station W07E (May 12) (Figure 1) gives graphical assurance that this pivot volume works well at this QDF station. For comparison, Figure 2 shows a similar plot of predicted versus observed speeds for the same station based on the use of Equation 1. For that method, length was set to 5.4 m (17.7 ft), the default vehicle length of 4.4 m used by the MTO plus a 1-m effective detector length. Although this figure shows that the method

TABLE 3 Effect of Changing Pivot Point: Station E02W, May 12, Critical Speed of 90 km/hr

Pivot Vol (veh/20-s)	0	6	8	9	10
Pivot Occupancy (%)	0	6	8.5	9.5	11
Theta (Theta)	-15.0	-15.0	-15.0	-15.0	-16.0
Misclassified Points	3	3	3	3	3
a	33.11	188.4	227.8	221.6	210.1
b	48689	54972	47337	42317	37623
Mean Error	2.24	3.71	1.12	**0.61**	-1.05
Standard Deviation	9.55	8.3	7.70	**7.56**	7.57
Ave. Difference^2	97.73	82.66	60.31	**57.63**	58.36

Pivot Vol (veh/20-s) Pivot	11	12	13	15	17
Occupancy (%)	12	13	14.5	16.5	19
Theta (Degrees)	-16.0	-16.0	-16.0	-17.7	-16.0
Misclassified Points	3	3	3	3	3
a	168.1	165.61	136.4	99.46	75.57
b	31003	31121	29782	28461	28181
Mean Error	-1.89	-1.17	1.92	-1.54	-1.62
Standard Deviation	7.86	7.78	8.04	8.46	8.79
Ave. Difference^2	59.43	61.98	68.3	74.01	79.86

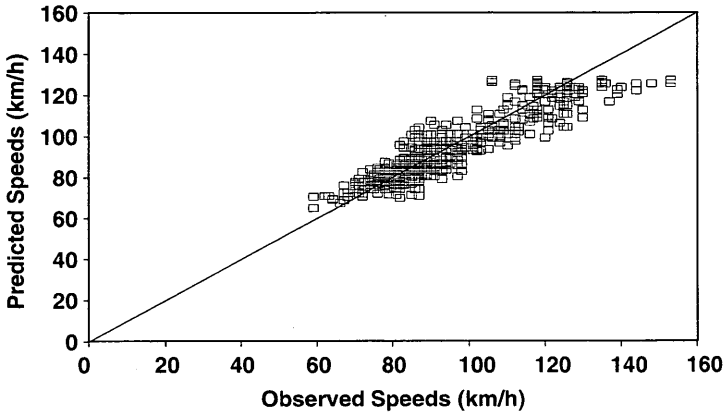


FIGURE 1 Comparison of observed average speeds with those predicted using catastrophe theory model: Station W07E, May 12.

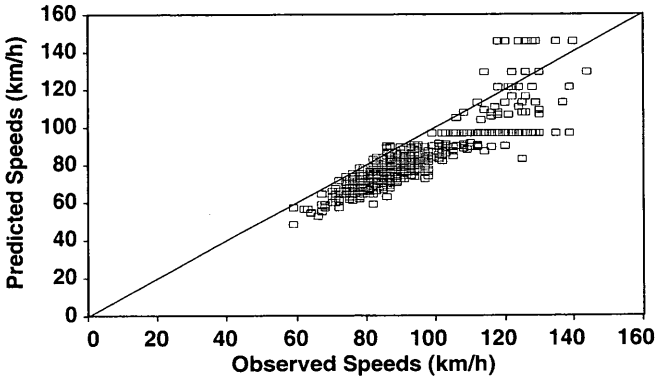


FIGURE 2 Comparison of observed speeds with those predicted on the basis of constant vehicle length: Station W07E, May 12.

produces biased estimates, which could be corrected easily by changing the value of length in Equation 2, other stations showed the opposite bias for the same constant length.

For the rest of the study, a pivot volume of 10 veh/20 sec will be used, with a critical speed of 90 km/hr. The pivot occupancy is the value that minimizes the number of misclassified points and will remain station-specific. Although these new parameters improve on previous work, the catastrophe model still produces a systematic error at some stations, as can be seen in Figure 3.

TESTING PARAMETER TRANSFERABILITY

There are three parts to the analysis of parameter transferability. The first investigates the stability of the parameters over different days at the same station. In previous studies (*9*), one of the troublesome results has been that distinct parameter sets were estimated for different days at the same site. It is essential to be able to overcome this if speed estimates are to be reliable. The second part of the analysis tests whether the ability of the parameter values to be generalized can be extended to other double-loop stations. Applying the values to such stations allows direct matching of speeds estimated using the catastrophe theory model with the measured speeds. The final part of the analysis is the extension to single-loop stations. Although this is the "acid test," it is the hardest to assess, which is why the two preceding sections are included.

Stability of Parameters over Time

The results of testing for parameter stability over different days, as displayed in Tables 4 and 5, were promising. Table 4 deals with a

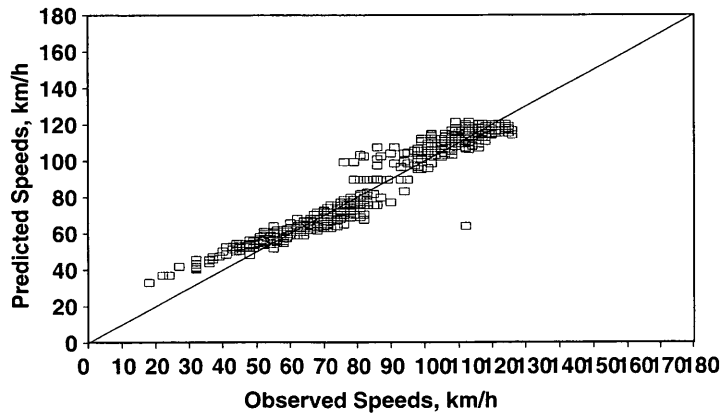


FIGURE 3 Comparison of observed average speeds with those predicted using catastrophe theory model: Station W04W, June 7.

regular station, W02W. The first set of values shows the results based on optimizing for each day separately, given the critical speed and pivot volume. (The rain day had a fairly steady drizzle, resulting in wet pavement, but it was not a heavy rain.) The second set of values shows that when the parameters are averaged across the clear-weather days (and considerably rounded off for *a* and *b*), the resulting values of average difference squared are only marginally higher than for the optimized parameters. Hence it would appear that for regular stations, a single set of parameters can perform well across days, including one with less-than-ideal weather.

For a QDF station, the results are also generally promising, but different in detail (Table 5). In one case, the averaged parameters are slightly worse than the optimized ones, as for the normal station. However, for two of the days, the average difference squared is even lower than it was in the optimized case. (The fact that it might be better than the optimized results signals the hazards of heuristic search and is not an indication of an error in the analysis or printing.) On the second day (June 3), the average difference squared is nearly twice its optimized value, but even so it is only of the same order as the optimized values for the regular station. It seems fair to say that these parameters are stable over time.

Transferability to Other Double-Loop Stations

A further test was done to determine the potential for finding a single set of parameter values that would produce reasonable results when applied to a number of stations. Parameters were determined for four stations, and rough approximations of the average values of these parameters were used in this test. The results (which are not shown in a table) ranged from reasonable to quite poor. For example, considering the increase in average difference squared between the "optimal" parameters and the common ones, reasonable results had error increasing only from 33.1 to 41.6; poor results almost tripled the error, to 144.8 from 54.5.

For another station, a small change in the pivot occupancy (from 11 to 12.5 percent), all other parameters being kept constant, reduced the average difference squared by more than half, from 212.5 to 97.6 (optimized results being 46.1). Hence, it may be best to put some initial effort into calibrating individual stations rather than attempting to find universal parameters, except at stations that are virtually identical.

Speed Prediction at Single-Loop Stations

In applying the catastrophe theory model to speed prediction at single-loop stations, parameters determined at adjacent double-loop stations were used. Because single-loop stations do not have measured speeds, predicted speeds and any observed speed cannot be compared numerically. Instead, to test the relative validity of speeds predicted using catastrophe theory, resulting speed-volume plots were compared with similar plots based on speed values given by two other methods of speed prediction: that given by Equation 1, and the method based on an average vehicle length from the nearest double-loop station, currently used by the MTO and available in the data base.

TABLE 4 Comparison of Parameter Stability Across Days, Station W02W

Critical Speed=90 km/h Pivot Volume=10 veh/20s		Clear Weather			Rain
		May 12	June 3	June 7	May 31
Optimized Results	Occupancy	11	11	11	11
	Theta	-16.0	-16.9	-16.0	-15.0
	a	221.9	258.5	260.0	253.8
	b	52857	52135	54281	55492
	Ave. Diff^2	46.87	100.52	54.49	45.52
Averaged Parameters	Ave. Diff^2	47.38	102.26	54.72	46.00
Occ = 11 Theta = -16.0 a = 250 b = 53100					

TABLE 5 Comparison of Parameter Stability Across Days, Station W04W

Critical Speed=90km/h Pivot Volume=10veh/20s		Clear Weather			Rain
		May 12	June 3	June 7	May 31
Optimized Results	Occupancy	11	11	11	11
	Theta	-14.0	-14.0	-14.0	-10.0
	a	88.90	133.15	88.34	72.56
	b	25923	26703	22779	18615
	Ave. Diff^2	26.65	29.44	33.08	32.30
Averaged Parameters	Ave. Diff^2	27.59	50.00	32.25	22.39
Occ = 11 Theta = 14.0 a = 100 b = 25100					

Five single-loop stations were analyzed for the feasibility of speed prediction by catastrophe theory. Here the results for W03W, June 7, are described; they are representative of those obtained for the other stations. Speeds similar to those at W03W would be expected at paired-loop detector W04W, located about 650 m downstream. Figure 4, the speed-flow curve based on speeds measured at W04W, provides a base for comparing the results from W03W.

In Figures 4 through 7, time-connected data, rather than simple scatter plots, are presented. The time-connected plots have three advantages. First, they give an indication of the number of times that a point occurs. Second, they show the distinction in the two regimes of traffic flow, which is not at all obvious from the scatter plot. And third, they show the order in which events occurred, in a general way. All three speed-flow plots for W03W (Figures 5, 6, and 7) indicate three types of operation—congestion, QDF, and uncongested flows—but the quality of the estimates varies considerably. One of the issues is the variation of speeds within uncongested operations. Figure 5 shows the speed-volume plot using the method based on constant vehicle length. Predicted speeds reach almost 200 km/hr and are regularly above 140 km/hr at flows below 10 veh/20 sec (which is a flow rate of 1,800 vehicles per hour). Uncongested speeds generally appear to be above 110 km/hr. A reduction in the vehicle-plus-detector length used would of course lower these numbers, but the combined vehicle and detector length are probably already underestimated. In addition, the scatter in the uncongested data is greater than normally seen with measured speed, as in Figure 4.

Figure 6 shows the plot resulting from predictions based on nearby length estimates. Here the speeds are more believable, surpassing 140 km/hr only a few times. There is a large amount of scatter in the uncongested region. This contrasts markedly with Figure 7, which shows a plot using the speeds predicted using catastrophe theory.

The parameters used in the application of the catastrophe theory model at W03W are a rough average of the parameters determined at Station W02W, just upstream of it, over 2 days. The predictions based on this model show more believable uncongested speeds, but the scatter in the uncongested regime is probably underestimated. Compared with Figure 4, Figure 7 shows too little scatter in the speed estimates, but Figure 5 shows too much. Hence it is not clear

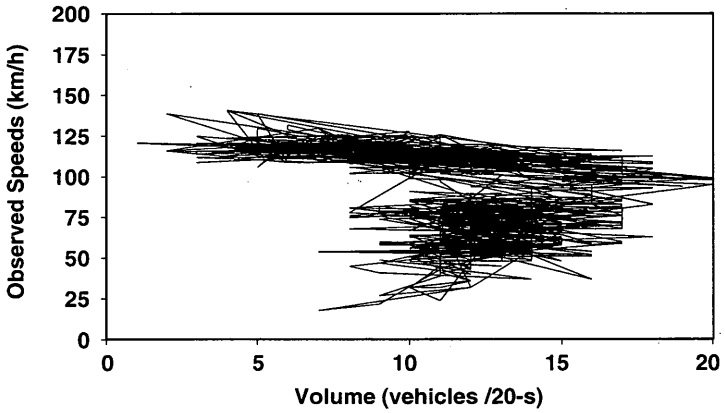


FIGURE 4 Speed-flow diagram using measured speeds: Station W04W, June 7.

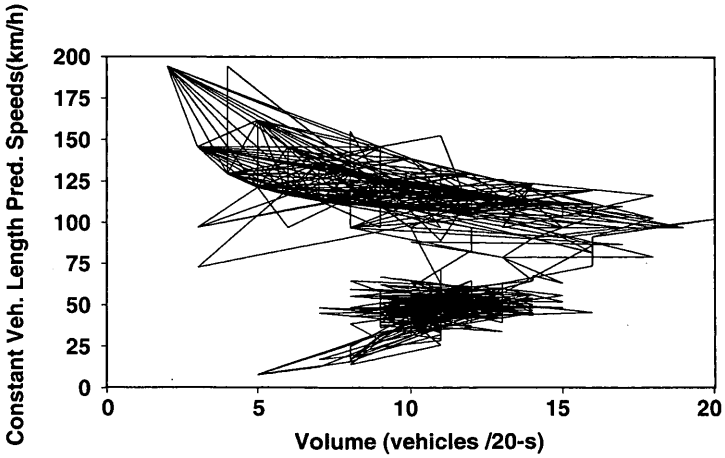


FIGURE 5 Speed-flow diagram using speeds calculated on the basis of constant vehicle length: Station W03W, June 7.

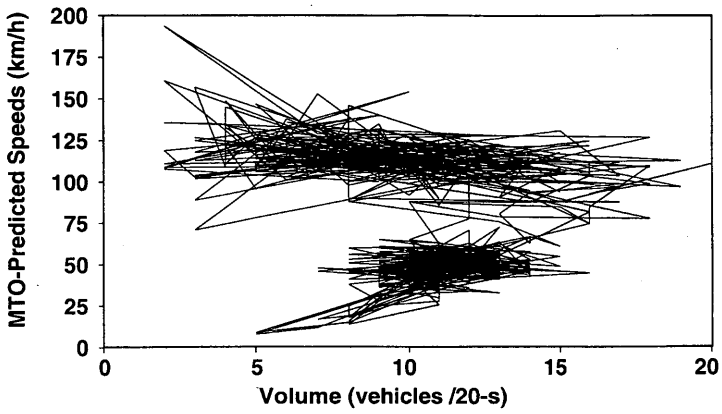


FIGURE 6 Speed-flow diagram using speeds calculated on the basis of vehicle lengths measured at adjacent stations: Station W03W, June 7.

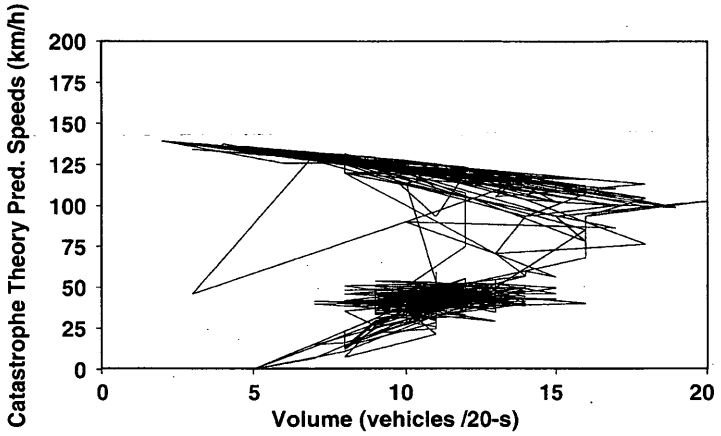


FIGURE 7 Speed-flow diagram using speeds calculated using catastrophe theory model: Station W03W, June 7.

that the catastrophe theory estimates are markedly better than those based on a nearby average. It is, however, clear that they are better than estimates based on a fixed vehicle length, especially below 10 veh/20 sec.

CONCLUSIONS

As part of this effort to use the catastrophe theory model to estimate speeds at single-loop stations, much was learned about the model itself. The axes transformations do not need to be as data-dependent as they were in previous work, and results occasionally improved when constant parameter values were used. A critical speed of 90 km/hr, together with a pivot volume of 10 veh/20 sec, produced reasonable results. The pivot occupancy would be approximately the occupancy that minimizes the number of misclassified data points based on the critical speed. Parameters for a particular station can be transferred quite well across days, but calibration should be fine-tuned across stations.

Applying the catastrophe theory model to speed prediction at single-loop stations gave encouraging results. The predicted speeds were clearly more reasonable, both in terms of levels of speed and scatter in the estimates, than were speeds predicted on the basis of a constant vehicle length. Catastrophe theory predictions of uncongested speeds have less scatter than they should, but speeds based on vehicle lengths measured nearby have more scatter than they should. The question then becomes which kind of error is preferred. Given that an estimate close to the mean is sought, the catastrophe theory model may be preferable.

ACKNOWLEDGMENTS

The work described in this paper was financed by a grant from the Natural Sciences and Engineering Research Council of Canada, whose support is gratefully acknowledged. The assistance of MTO in making the data available was much appreciated. In this regard, special mention should be made of Emanuel Morala of the Freeway Management Section of MTO.

REFERENCES

1. Acha-Daza, J. A., and F. L. Hall. Graphical Comparison of the Predictions for Speed Given by Catastrophe Theory and Some Classic

Models. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993, pp. 119–124.
2. Acha-Daza, J. A. *The Application of Catastrophe Theory to Traffic Flow Variables*. Master's thesis. Department of Civil Engineering, McMaster University, Hamilton, Ontario, Canada, June 1992.
3. Hall, F. L., and B. N. Persaud. Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.
4. Navin, F. P. D. Traffic Congestion Catastrophes. *Transportation Planning and Technology*, Vol. 11, 1986, pp. 19–25.
5. Hall, F. L. An Interpretation of Speed-Flow Concentration Relationships Using Catastrophe Theory. *Transportation Research*, Vol. 21A, No. 3, 1987, pp. 191–201.
6. Dillon, D. S., and F. L. Hall. Freeway Operations and the Cusp Catastrophe: An Empirical Analysis. In *Transportation Research Record 1132*, TRB, National Research Council, Washington, D.C., 1988, pp. 66–76.
7. Forbes, G. J., and F. L. Hall. The Applicability of Catastrophe Theory in Modelling Freeway Traffic Operations. *Transportation Research*, Vol. 24A, No. 5, 1990, pp. 335–344.
8. Acha-Daza, J. A., and F. L. Hall. Application of Catastrophe Theory to Traffic Flow Variables. *Transportation Research B* (in preparation).
9. Ceder, A. *Investigation of Two Regime Traffic Flow Models at the Micro- and Macroscopic Levels*. Ph.D. dissertation. University of California, Berkeley, 1975.
10. Ceder, A. A Deterministic Traffic Flow Model for the Two-Regime Approach. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 16–30.
11. Ceder, A., and A. D. May. Further Evaluation of Single and Two-Regime Traffic Flow Models. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 1–15.
12. Koshi, M., M. Iwasaki, and I. Ohkura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory* (V. F. Hurdle, E. Hauer, and G. N. Steuart, eds.), University of Toronto Press, Ontario, Canada, 1983, pp. 403–426.
13. Gilchrist, R. S., and F. L. Hall. Three-Dimensional Relationships Among Traffic Flow Theory Variables. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 99–108.
14. *Special Report 209: Highway Capacity Manual*, 3rd ed. TRB, National Research Council, Washington, D.C., 1994.

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

Toward the Use of Detector Output for Arterial Link Travel Time Estimation: A Literature Review

VIRGINIA P. SISIOPIKU AND NAGUI M. ROUPHAIL

The ability to estimate travel time on-line for use in signal timing optimization and route guidance and vehicle navigation applications is becoming necessary. Detector data are considered a valuable source of information on traffic conditions in transportation facilities. The development of models that use detector information to estimate travel time in urban networks is traced. Existing research efforts are briefly described and interrelated. A comparative discussion of the alternatives is aimed at providing a qualitative evaluation of a range of options available today for developing arterial travel time functions. The need for further calibration and validation of existing models is identified, and enhancements to improve their quality and applicability are recommended.

Road networks can be made more efficient through the implementation of advanced traffic management and control systems, as well as by giving drivers more accurate information to help them avoid traffic congestion. In intelligent vehicle-highway systems (IVHS) parlance, these capabilities refer to advanced traffic management systems (ATMS) and advanced traveler information systems (ATIS), respectively. In both approaches the need for reliable information on the current traffic situation is essential.

Loop vehicle detector systems are a valuable source of information for studying and monitoring the performance of traffic networks. The output from loop detectors contains information on traffic volumes, occupancy levels, and arrival patterns. These data may be applied directly or may be used in functions relating them to other important parameters defining the performance of a road network, such as travel time, safety, and comfort.

This paper reviews previous research efforts aimed at the development of models for estimating travel time from detector output under various traffic and road conditions. Travel time is a key parameter required to pinpoint trouble spots both for immediate use and for planning and reporting purposes. It can be used as an indication of the overall road system performance, as a real-time measure of congestion, as the means for assessing traffic management strategies, as a planning tool, and as an input to project evaluation (*1*). As a measure of link performance, travel time effectively allows the traffic performance of different links within a network to be evaluated and compared, which greatly facilitates the identification of critical links in a network and provides an important input into the planning process.

V. P. Sisiopiku, Intelligent Vehicle-Highway Systems Laboratory, University of Illinois at Chicago, Department of Electrical Engineering and Computer Science (M/C 154), 1120 Science and Engineering Offices, Box 4348, Chicago, Ill. 60680. N. M. Rouphail, Department of Civil Engineering, North Carolina State University, P.O. Box 2908, Raleigh, N.C. 27695.

Travel time also provides an excellent measure of the effectiveness of specific projects because any improvement can be readily quantified. This is particularly useful to operators of coordinated traffic signal systems for assessing whether changes in signal control strategies or timing have been effective. Besides, it can be used for evaluating the ability of dynamic (real-time) in-vehicle guidance systems to improve both decisions on route selection and the performance of traffic systems.

The purpose of this paper is twofold:

1. To review past research on travel time estimation based on detected flows and occupancy levels on signalized arterial links, and
2. To study the ability of the existing formulations to provide accurate estimates of travel time and identification of potential improvements on both the estimation procedures and the models themselves.

OVERVIEW

Several studies have attempted to develop relationships between travel time (measured in the field or simulated) and surveillance detector data (flows, occupancies, or both). Some of these studies examined the impact of the location of the detector on a link, and a few used elements of traffic control to better model the travel time variations observed in urban networks. The vast majority of existing work focuses on the use of regression analysis to estimate travel time in terms of some or all of the factors outlined previously (*2*).

The primary motivation of this work is the need for better management of signals in road traffic computer-control systems. In this context, travel time is usually viewed as the single most important criterion in optimizing the signal settings process. On the other hand, recent ATIS applications created the need for accurate estimation of travel times for route planning. This challenge gives a new dimension to the investigation of the interrelationships between travel time and detector output that is expected to be advanced in the coming years.

In the following, worldwide research on identifying the relationship between travel time on arterial links and loop detector information is documented briefly; the unique characteristics of each approach are emphasized, and the limitations and shortcomings of each are summarized. The review is basically organized chronologically, except in instances in which the chronological order is altered to introduce related work that provides better insight into the interrelations between the approaches. Formulations of selected approaches are also reported. However, the reader is encouraged to consult the references for definitions of the parameters and further

details. Recent modeling efforts on travel time estimation based on flow/occupancy data are compared using level of aggregation selected, data sources used, factors considered, type of model developed, model variables, limitations and reliability. Finally, concluding comments and recommendations for future research and development are given.

LITERATURE REVIEW

Basic Concepts

Gipps (*3*) was one of the earliest advocates of using detector occupancy and arrival time at the detector to develop regression estimates of link travel time based on simulated data. His plots of vehicle travel time on a link against arrival time at a detector showed a clear discontinuity. To overcome this difficulty, he decided to choose another zero point and conveniently defined "register time" so that, on average, undelayed vehicles that passed over the detector at register time zero will reach the stopline at the time the signal indication turns red. He realized the need for incorporating the effects of the signal settings, number of lanes, and changes in the link length into the parameters and offered a suggestion on ways to untangle the effect of the correlations on the parameter estimates.

Gault and Taylor improved Gipps's initial model by discarding parameters of low importance, taking into account the correlations between variables, and calibrating the model for a two-lane highway on a lane-specific basis (*4,5*). Gault also observed a linear relationship between travel time and detector occupancy up to occupancies of approximately 70 percent. She chose to ignore higher occupancies and formulated a model that reflected the effects of occupancy levels, cruise time, degree of saturation, and signal settings on link travel time. Among her conclusions were that the optimum detector positioning is 120 ft upstream of the traffic signals and that aggregation of detector output over 20-min (as opposed to 5-min) periods does not have a significant impact on the accuracy of the travel time prediction.

Strobel treated the estimation of link travel times as a problem of system identification (*6*). His objective was to find an appropriate relationship between the input and output time series of traffic flow and to estimate the values of the parameters that identify this relationship. The input and output to the transfer function were time series of traffic volumes collected from an upstream and a downstream detector, respectively. He also suggested how the concept could be used for on-line applications.

Later, Luk tested Strobel's formulation with both traffic flow and with data on wheelbase length collected on an urban arterial road (*7*). His motivation for using wheelbases came from the rural road traffic studies of Hoban (*8*). Luk confirmed the validity of the input-output framework for platoon travel time estimation and found that journey times are insensitive to the congestion level. This characteristic was attributed to the difficulty in estimating the journey times of those vehicles at the tail end of a platoon that could not pass an intersection in one green phase. The observation that the platoon journey time is insensitive to congestion level can also be concluded from the results of Gault and Taylor (*5*).

Abours attempted to study the relationship between occupancies obtained from detectors and travel times measured by floating cars using a polynomial relationship that, however, was not reported (*9*). She also suggested the use of substitution detectors—that is, additional detectors placed on the same links—to provide occupancy data when a detector failure occurs and studied the impact of such substitutions. Comparisons of computed and measured travel time show a consistent overestimation of travel time.

Lin and Percy (*10*) and Lin and Shen (*11*) emphasized the importance of adequately representing the vehicle-detector interactions in any simulation model used in analyzing traffic-actuated control. In their work on vehicle-detector interactions, they calculated delay as a function of vehicle interval and flow rate for motion control and as a function of detector length, extension interval, and flow rate for presence control.

Usami et al. proposed a formulation for travel time estimation on an oversaturated link (*12*). Travel time is expressed as a function of link length, traffic volume, and traffic density, treating density as a linear function of volume. The procedure was validated through a license plate survey.

Luk and Cahill proposed a scheme by which system performance can be monitored with stopline detectors (*13*). Link flows were first estimated by a recursive least-squares algorithm from stopline departure flow profiles collected upstream. Platoon delay was then estimated from the predicted arrival and the actual departure profiles. The scheme introduced modeling into a signal control system such as SCATS and could be applied for optimal selection of offsets. Results based on simulated data indicate that the scheme is practicable.

Young verified Gault's earlier observation of the existence of a linear relationship between mean occupancy per vehicle and mean delay per vehicle given that queues clear the most distant detector during green phases (*14*). His results showed that the delay-occupancy relationship contains a linear segment and that the range of this segment is related to the length of roadway covered by detectors. Young emphasized the role of the detector layout in the validity of the argument for linearity and discussed his findings without, however, providing a calibrated model.

All the research work presented thus far examined the relationships between travel time and a variety of factors on a link-specific basis. The recent focus on IVHS, however, increased the interest in addressing travel time and delay on a section-specific basis. Toward this direction, Bohnke and Pfannerstill introduced a system that uses inductive loop detectors and pattern recognition principles to reidentify platoons of vehicles after they have traversed a specific road section and obtain the journey time for the platoon from the instant of reindentification (*15*).

In recent work, Takaba et al. also referred to section travel times but treated them simply as the summation of travel times for those links composing the section (*16*). Link travel times were estimated from link detector information including traffic flow and queue length based on regression analysis. They framed two models, each based on the summation of link travel time for the uncongested and congested part of the link, using the formulation developed earlier by Usami et al. (*12*) for the latter component. However, the approach they suggest for estimating section travel times is of low value as it still requires calculation of individual link travel times, neglects the dependency of travel times between consecutive links, and requires detectors to be located on every link of a section.

Most of these studies suffer from limited calibration and validation as well as a neglect of such factors as link length, distribution of traffic between movements, traffic composition, and driver behavior, any of which may influence the estimation of travel time significantly. Thus, a generalization of the results without further testing and recalibration of the model parameters would be inadvisable.

The models reviewed herein concentrated on travel time estimates for all movements on a link, thus the differences in travel time values among the various turning movements are not reflected in them. This issue needs to draw further attention as travel times of left-turn movements, for instance, are considerably higher than those experienced by through vehicles, especially when the flows opposing the turning movement are heavy.

Formulations

Gipps used a linear regression model in quadratic form to describe travel time in terms of register time (as defined earlier) and occupancy level (*3*). The initial model was

$$T = a + (1 - \delta)(b_{10}t^* + b_{01}\phi + b_{20}t^{*2} + b_{11}t^*\phi + b_{01}\phi^2) + \delta(c_{10}t^* + c_{01}\phi + c_{20}t^{*2} + c_{11}t^*\phi + c_{01}\phi^2) + \epsilon \quad (1)$$

where

T = travel time,
$t^* = t - (C - G + \text{lag})$,
ϕ = occupancy level,
t = register time,
C = cycle length at downstream signal,
G = green time at downstream signal,
R = red time at downstream signal,
lag = average time for a vehicle to travel from detector to stopline,
ϵ = random variable from $N(0, T^2)$,
$a, b_{10}, b_{01}, b_{11}, b_{20}, b_{02}, c_{10}, c_{01}, c_{11}, c_{20}, c_{02}$ = parameters, and

$$\delta = \begin{cases} 0 \text{ when } t \le R \\ 1 \text{ when } t > R. \end{cases}$$

This model is reduced in stages to a simpler form that provided a fit nearly as good as the original. The final model reported was of the form

$$T + (1 - \delta)t^* = a + (1 - \delta)b_{01}\phi + \delta(c_{10}t^* + c_{01}\phi) \quad (2)$$

which led to an estimate of travel time for a single vehicle of the form

$$T = \begin{cases} (a + R) - t + b_{01}\phi & \text{for } t \le R \\ (a - c_{01}R) + c_{01}t + c_{01}\phi & \text{otherwise} \end{cases} \quad (3)$$

By pursuing the same initial model as Gipps (see Equation 2) but discarding parameters not proven important and taking into account the correlations between variables, Gault and Taylor (*5*) further reduced the model initially proposed by Gipps to

$$T = (1 - \delta)at^* + \delta g^{1.6} + K \quad (4)$$

where a, g, and K are parameters described as functions of the offset (off), undelayed time (undt = link length/desired speed), and degree of saturation (x).

From multiple regression analysis of the results from 60 simulations, the relationships for a single lane were found to be

$$a = 0.0168 \text{ off} - 0.0266 \text{ undt} - 0.375x - 0.609$$

$$g = -0.00027 \text{ off} + 0.00077 \text{ undt} + 0.0104x - 0.00386$$

$$K = 0.392 \text{ off} + 0.832 \text{ undt} + 11.35x - 4.13 \quad (5)$$

Similar results are reported by Gault for a two-lane case in which each lane is calibrated separately (*4*). Gault also derived an occupancy model of the form

$$\bar{t} = aO + b \quad (6)$$

where

$\bar{t}$ = average link travel time,
O = average detector occupancy,
$a = f(\text{undt}, x, P_d)$,
$b = g(\text{undt}, x, P_d)$, and
P_d = percentage of green time at downstream signals.

Gault's research indicated that P_d/P_u is a more appropriate parameter on which the relationship between detector occupancy and travel time depends, with P_u being the green time at the upstream signals. She calibrated the parameters for a and b as

$$a = 0.33 - 0.004 \text{ undt} - 0.057x + 0.294(P_d/P_u)$$

$$a = 9.95 - 1.42 \text{ undt} - 0.996x - 10.5(P_d/P_u) \quad (7)$$

and used Equation 6 to predict travel time.

Usami et al. (*12*) considered the congested section of the road to be divided into subsections i where there is no inflow or outflow of vehicles and suggested that travel time for the congested section be expressed generally as

$$T = \sum_i \left(\frac{L_i}{H_i}\right)\left(\frac{1}{Q_i}\right) \quad (8)$$

where

T = travel time (sec),
L_i = length of (congested) section i (m),
H_i = average space headway (m/veh), and
Q_i = traffic volume (veh/sec).

They then modified Equation 8 by letting traffic density K (i.e., the inverse of H) be a linear function of the traffic volume, Q, of the form

$$K = k_m - kQ_i \quad (9)$$

yielding the following formula

$$T = \Sigma_i L_i(k_m - kQ_i)\left(\frac{1}{Q_i}\right) = k_m \Sigma_i L_i \left(\frac{L_i}{Q_i}\right) - k \Sigma_i L_i \quad (10)$$

where k_m and k are constants with preassigned values. Calibration using travel time data obtained by a license plate survey yielded values for k_m and k of 0.107 and −0.181, respectively.

Takaba et al. used the same approximation as Usami et al. treating the relationship between density, K, and flow volume, Q_i, as linear under congested conditions (*16*). The model that they developed (the so-called sandglass model) estimates travel time for link i as the summation of the travel time in the congested section and the travel time in the uncongested section as

$$T_i = \frac{N_i}{Q_i} + \frac{(L_i^0 - L_i)}{v_a} \quad (11)$$

where

T_i = travel time of link i (sec),
N_i = number of vehicles in queue,
Q_i = flow volume (veh/sec),
L_i^0 = length of link i (m),
L_i = queue length (m), and
v_a = desired speed (m/sec).

By introducing traffic density, K, where $K = N_i/L_i$ and assuming the linear approximation given in Equation 9, Equation 11 can be rewritten as

$$T_i = \left(\frac{k_m L_i}{Q_i} - kL_i\right) + \frac{(L_i^0 - L_i)}{v_a} \quad (12)$$

where the travel time estimation for the congested part is identical to that of Usami et al. Notice that k_m is the jam density and k_m, k are regression coefficients.

In addition to the sandglass model, Takaba et al. proposed a delay model that actually converges to the sandglass model if the regression coefficient, k, is set to $k_m/s - 1/v$ with s being the saturation flow and v the running speed. They defined travel time in the congested sections as the summation of delay and running time. Delay is expressed as

$$D_i = (C - G_i)\left(k_m \frac{L_i}{Q_i C}\right) = C\left(1 - \frac{Q_i}{s}\right)\left(k_m \frac{L_i}{Q_i C}\right) \quad (13)$$

where C is the cycle length in seconds and G_i is the effective green time in seconds. Notice that the first term of Equation 13 corresponds to the delay occurring per congested cycle and the second reflects the duration of the congestion for link i, in number of cycles. From Equation 13 and for running time in the congested section equal to L_i/v, the delay time model suggested by Takaba et al. becomes

$$T_i = \left[\frac{k_m L_i}{Q_i} - L_i\left(\frac{k_m}{s} - \frac{1}{v}\right)\right] + \frac{(L_i^0 - L_i)}{v_a} \quad (14)$$

COMPARISON OF ALTERNATIVE PROCEDURES

Introduction

The previous sections focused on a presentation of the general concepts and basic formulations of the procedures developed to assess travel time and delay in urban networks using detector data. This presentation was meant to familiarize the reader with the literature available on the topic. Here the procedures are compared in attempts to provide an in-depth analysis of their characteristics, present their advantages and shortcomings, highlight their differences, and address their validity and applicability.

The possibility of comparing the alternative procedures with actual data was first considered. Such an approach would have been useful for future researchers in selecting the models that showed the most reliable performance and the closer fit to the actual data, but several major problems were encountered. First, all models currently available are site-specific. As often recognized by the researchers themselves, the transferability and applicability of their models under different conditions is limited. Moreover, differences in the estimation methods do not allow for comparisons under a general study design. For example, decisions on issues such as detector location, type of control, and patterns of traffic demand are required when designing the settings of the general experiment. These parameters should remain fixed for all alternative models tested, which poses a problem because of the assumptions involved in each model or range of operations for which it has been developed. For example, Luk's work demands stopline detectors, whereas all other models assume that detectors are placed in various locations upstream of the traffic signals. Lin and Percy studied the case of actuated traffic control, whereas Gipps and Gault assumed fixed traffic settings. Finally, the work of Usami et al. and Takaba et al. is indented for oversaturated conditions, whereas Gault suggested bounding the models under such conditions (for occupancies of more than 50 to 70 percent).

Because of such difficulties, the idea of comparing the various methods using the same data set was abandoned. Instead, the models have been compared in terms of their scope, characteristics, and limitations. This comparison is organized in table form. First, some general information about the models is provided, including the measure of performance selected (travel time versus delay), the key variables used to relate travel time to detector output (flow, occupancy, or both), the level of aggregation selected (link-movement, link, section), and the data sources used to collect or generate the data. This is presented in Table 1.

The model characteristics are given in Table 2; they include the type of model proposed, factors varied in the analysis, and variables used for the model development. Table 3 focuses on the validity and applicability of each approach and briefly presents the limitations of the procedures, the validation process, and some statistical measures indicative of the prediction accuracy.

Discussion of Results

The review indicates that substantial research is required to investigate the relationship between travel time and flow or occupancy on arterial links, because the factors involved are numerous and complex. Basic observations on the nature of these relationships have been reported, and a few formulations have been derived for simplified situations. However, more work is needed to calibrate and validate the proposed link travel time functions before they are implemented on a larger scale.

Most researchers selected travel time as measure of performance and, thus, developed formulations using travel time as the dependent variable. They agreed that travel time is more manageable than delay, which, being the difference between two values, is an awkward quantity to assess. Furthermore, the use of delay is complicated by the existence of several possible definitions.

Several of the approaches preferred the use of flow over occupancy as the key independent variable; this is partly because of the tradition of expressing link travel time as a function of flow in link performance functions, extensively used in planning applications. Among these functions, the equation developed by the U.S. Bureau of Public Roads (*17*) and the formula proposed by Davidson (*18*) and later revised and extended by Akçelik (*19,20*) are the ones more often used in practice. Another possible reason for using traffic flow as the key explanatory variable is the ease in collecting vehicle counts from loop detectors in the field: several types of detector do

TABLE 1 Scope of General Models

Model	Date	Dependent Variable	Key Independent Variable	Level of Aggregation	Data Source
Gipps	1977	Travel Time	Occupancy	1-Lane Link	Simulation
Gault et al[a]	1981	Travel Time	Flow	Lane	Simulation
Gault[b]	1981	Travel Time	Occupancy	Lane	Simulation
Abours	1981	Travel Time	Occupancy	Link	Floating Car
Luk et al	1986	Delay	Flow	Link	Simulation
Usami et al	1986	Travel Time	Flow	Link	Simul./Lic. Plates
Young	1988	Delay	Occupancy	Link	License Plates
Luk	1989	Travel Time	Flow	1-Lane Link	Wheelbase Match
Takaba et al[c]	1991	Travel Time	Flow/Speed	Link/Section	License Plates
Takaba et al[d]	1991	Travel Time	Flow/Speed	Link/Section	Vehicle Detectors

[a]Arrival type model
[b]Occupancy model
[c]Sandglass model
[d]Delay model

not provide occupancy information. However, the review indicates that occupancy may be a better predictor for travel time than flow. Further investigation on developing link travel time functions using occupancy data from loop detector systems is a major task for further research.

It is worth noting that traffic flow and occupancy were never used simultaneously in any of the models reported in the literature. Although such an option has not explicitly been explored so far, it is believed that the high correlation between the two may restrict their coexistence in a regression formulation.

Several other variables were used as independent variables in the equations suggested for link travel time estimation. These variables include signal settings (cycle length, red time), queue length, dispersion parameter, and speeds (running, desired). See Table 2 for an enumeration of the variables used in each model.

All regression relationships reported in the literature are site-specific, that is, the models are calibrated for each link and travel times are then estimated on a link-by-link basis. Generalization of the models so that they can apply to groups of links with similar characteristics needs further research.

All alternative procedures depend on the appropriate placement of enough vehicle detectors in the traffic lanes approaching the junction. Several researchers study the optimal placement of the detectors, and there is general agreement that detector location can affect the results significantly. The most interesting work on this issue is reported by Young (*14*).

As noted earlier, the vast majority of the research deals with the development of link-specific functions. The work by Takaba et al. (*16*) addressed travel time estimation on a section-specific basis in a very simplistic way. This issue needs further study.

Several of the researchers used simulation models to study the relationships between travel time and flow/occupancy. A number of simulation runs were performed in each study. Selected factors were varied to better represent traffic conditions encountered in real urban networks. Among them, traffic volumes, offsets, and cycle length were the most popular factors.

Various techniques were used for gathering travel time data for validation, including license plate matching, floating cars, and wheelbase data matching. It should be noted, however, that validation of the models with field data was limited and most approaches were validated primarily with simulated data that yielded better results (within 10 to 20 percent of the mean). A review of the validation procedure applied in each case and the main shortcomings of each model are presented in Table 3.

CONCLUDING REMARKS

In this survey, the authors have reviewed and interrelated various developments pertaining to travel time estimation based on loop detector information. The main findings and conclusions follow:

1. The available research on converting fixed detector output to arterial travel times is limited because of the complexity of modeling traffic phenomena under interrupted travel flow conditions.
2. Most existing models are link-specific. Site dependency limits the applicability and transferability of the models under different demand, control, and geometric configurations.
3. None of the existing models accounts for the differences in travel times due to movement type. Movement-specific models are expected to enhance the quality of arterial travel time predictions.
4. In an urban environment, factors such as link length, distribution of traffic between movements, traffic composition, platoon dispersion, and driver behavior play a large role in estimating travel time. All of these factors have been disregarded in the models currently available; further attention in future model development efforts is needed.
5. The methods reviewed in this paper vary considerably in terms of assumptions made, variables involved, and range of traffic operations covered. Therefore, a comparison of the various procedures using the same set of actual data, although very valuable, is not practicable.
6. Recent interest in ATIS applications increases the need for estimating travel times at a section (as opposed to link) level. The literature review indicates a great need for more research toward this direction.

TABLE 2 Model Characteristics

Model	Type of Model	Factors Varied	Independent Variables
Gipps	Linear Regression; Quadratic Form	Cycle Length Offset Traffic Volume	Occupancy Level Register Time Red Time
Gault et al[a]	Multiple Linear Regression	Cycle Length Offset Traffic Volume	Register Time Red Time
Gault[b]	Multiple Linear Regression	Cycle Length Offset Vehicle Flow Vehicle Speed Link Length	Occupancy
Luk et al	Input-Output; Platoon Dispersion	Offset	Flow Profiles Signal Settings Undelayed Time Dispersion Parameter
Usami et al	Analytical; Sandglass	N/A	Queue Length Traffic Volume
Luk[c]	Computer Program	N/A	N/A
Luk[d]	Input-Output	N/A	Flow
Takaba et al[e]	Analytical; Sandglass	N/A	Queue Length Output Flow Rate
Takaba et al[f]	Analytical; Delay Model	N/A	Queue Length Output Flow Rate Running Speed Desired Speed

[a]Arrival type model
[b]Occupancy model
[c]Wheelbase matching technique
[d]Input-Output model
[e]Sandglass model
[f]Delay model

7. Most studies performed on travel time estimation from arterial detector output suffer from limited calibration and validation. In particular, field validation is generally missing. This considerably limits the applicability of the models under general traffic and road conditions.

Related issues that should be addressed in future research are summarized in the following:

1. Improvements in Modeling Framework

–*Development of movement-specific models*: even though through-movement travel time models may be suitable for right-turning travel time estimation, caution is advised if trying to apply them for left-turn treatments. Additional factors substantially affect travel time estimation on left-turning links (such as opposing flow) and must be incorporated in the models.

–*Estimation of section travel times*: knowledge of section travel times is often more valuable than link travel times for ATIS applications. The estimation of section travel times, given that several links in the path are detectorized, is a challenging issue for future research.

–*Development of generalized models*: link-specific models are site-dependent and need to be calibrated for every link they apply. To overcome this difficulty, generalized models should be developed. If they are to provide reasonable travel time estimates for the links to which they are applied, generalized models should include variables accounting for variations in geometric, flow, and control characteristics.

TABLE 3 Model Assessment

Model	Limitations	Validation
Gipps	- Lack of empirical validation - Signal settings/geometry not considered - Correlation of the parameters exists	With simulated data only; MSE[a]= 10-15%
Gault et al[b]	- Underestimates travel time for occ.>50% - Lack of empirical validation	With simulated data only; Within 10% of the mean
Gault[c]	- Bounded (occ. should be $\leq$ 70%) - Not appropriate for oversaturation	With video tape data; Within 10% (rarely up to 50%)
Abours	- Signal settings are ignored - Formulation not reported	With floating car data; RMSE[d]= 13%
Luk et al Usami	- Requirement of stop-line detectors - Applicable for oversaturation only	Not reported With simulation & field data RMSE = 10-19%
Luk	- Flow conservation assumption - More suitable for freeway environment - Requirement of stop-line detectors	With wheelbase data Within 10% of the mean
Takaba	- Linearity assumption between travel time & flow in congestion - Neglect of dependency between links	Error ratio = 12-24%

[a]MSE: Mean Square Error
[b]Arrival Type Model
[c]Occupancy Model
[d]RMSE: Relative Mean Square Error

2. Enhancements to Model Structure

–*Availability of real-time data*: an interesting application of the models estimating travel time from detector data is an ATIS framework. In that respect, the on-line availability of the data requested by the models should be a determinant during the model formulation process.

–*Calibration of model parameters using empirical data*: most of the models reported use simulation to study the relationships between travel time/delay and detector output. Recalibration of the existing model forms using empirical data is expected to enhance the quality of the models as the actual traffic behavior encountered in the field can be reflected.

–*Revision and expansion of current model structures*: as mentioned, several factors affecting link travel time estimation have been disregarded, including traffic composition, driver behavior, and platoon dispersion. Further experimentation of the model forms selected is encouraged to improve the quality, accuracy, and credibility of travel time prediction.

3. Improvements in Validation Procedure

–*Validation with field data*: tests of accuracy based on simulated data have their limitations since the comparisons are performed between predicted and observed data sets that depend on the simulation model itself. On the other hand, validation with empirical data can show whether the prediction models accurately reflect real-world conditions.

–*Comparative application of alternative models*: models that follow similar assumptions or are developed for application under similar traffic demand conditions should be compared. Doing so will facilitate the selection of the most reliable model forms in future applications.

ACKNOWLEDGMENTS

The authors wish to thank Joseph Raj of the U.S. Department of Transportation for his contribution to the preliminary research phase of this study. The financial support of this research by FHWA and the ADVANCE IVHS demonstration project in Chicago is greatly appreciated.

REFERENCES

1. Longfoot, J. An Automatic Network Travel Time System—ANNTS. *Proc., Society of Automotive Engineers*, Vol. 2, 1991, pp. 1053–1061.
2. Sisiopiku, V. P. *Arterial Link Travel Time Estimation from Loop Detector Output*. Technical Report DTFH61-92-P-40029. FHWA, U.S. Department of Transportation, 1993, pp. 13–29.
3. Gipps, P. G. *The Estimation of a Measure of Vehicle Delay from Detector Output*. Research Report 25. Transport Operation Research Group, University of Newcastle upon Tyne, England, 1977.
4. Gault, H. E. An On-Line Measure of Delay in Road Traffic Computer-Controlled Systems. *Traffic Engineering and Control*, Vol. 22, No. 7, 1983, pp. 384–389.

5. Gault, H. E., and I. G. Taylor. *The Use of the Output from Vehicle Detectors to Assess Delay in Computer-Controlled Area Traffic Control Systems*. Research Report 31. Transport Operations Research Group, University of Newcastle upon Tyne, England, 1977.
6. Strobel, H. *Traffic Control Systems Analysis by Means of Dynamic State and Input-Output Models*. RR-77-12. International Institute of Applied Systems Analysis, Laxemberg, Austria, 1977.
7. Luk, J. Y. K. *Modeling and Monitoring the Performance of Urban Traffic Control Systems*. SR 43. Australian Road Research Board, 1989, pp. 29–42, 54–60.
8. Hoban, C. J. *Vehicle Analysis and Classification from Axle Time Data*. AIR 359-13. Australian Road Research Board, 1984.
9. Abours, S. Estimation of Travel Times from Occupancy on an Urban Network: An Experiment in Paris, France. *Proc., 2nd International Conference on Road Traffic Control*, Institute of Electrical and Electronics Engineers, 1986, pp. 137–139.
10. Lin, F.-B., and M. C. Percy. Vehicle-Detector Intersections and Controls. Presented at 63rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1984.
11. Lin, F.-B., and S. Shen. Relationships Between Queuing Flows and Presence Detectors. *ITE Journal*, Vol. 61, No. 8, 1991, pp. 41–45.
12. Usami, T., K. Ikenoue, and T. Miyasako. Travel Time Prediction Algorithm and Signal Operations at Critical Intersections for Controlling Travel Time. *Proc., 2nd International Conference on Road Traffic Control*, Institute of Electrical and Electronics Engineers, 1986, pp. 205–208.
13. Luk, J. Y. K., and L. W. Cahill. On-Line Estimation of Platoon Delay. *Proc., 13th Australian Road Research Board*, Vol. 13, Part 7, 1986.
14. Young, C. P. A Relationship Between Vehicle Detector Occupancy and Delay at Signal-Controlled Junctions. *Traffic Engineering and Control*, Vol. 29, 1988, pp. 131–134.
15. Bohnke, P., and E. Pfannerstill. A System for the Automatic Surveillance of Traffic Situations. *ITE Journal*, Vol. 56, No. 1, 1986, pp. 41–45.
16. Takaba, S., T. Morita, and T. Hada. Estimation and Measurement of Travel Time by Vehicle Detectors and License Plate Readers. *Proc., Vehicle Navigation and Information Systems Conference*, Vol. 1, Society of Automotive Engineers, 1991, pp. 257–267.
17. *Traffic Assignment Manual*. Bureau of Public Roads, U.S. Department of Commerce, 1964.
18. Davidson, K. B. A Flow-Travel Time Relationship for Use in Transportation Planning. *Proc., Australian Road Research Board*, Vol. 3, Melbourne, 1966, pp. 183–194.
19. Akçelik, R. A New Look at Davidson's Travel Time Function. *Traffic Engineering and Control*, Vol. 19, No. 10, 1978, pp. 459–463.
20. Akçelik, R. On Davidson's Flow Rate/Travel Time Relationship. *Australian Road Research*, Vol. 8, No. 1, 1978, pp. 41–44.

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

Analysis of Correlation Between Arterial Travel Time and Detector Data from Simulation and Field Studies

VIRGINIA P. SISIOPIKU, NAGUI M. ROUPHAIL, AND ALBERTO SANTIAGO

The effectiveness of control strategies applied to alleviate traffic congestion depends heavily on the accuracy and credibility of the data sources used. Among data sources now available, loop detector systems can provide large quantities of high-quality data. The feasibility of using detector data to improve the performance of advanced traveler information systems functions, such as arterial travel time estimation, is examined. The relationships between travel times and flow/occupancy information are assessed using simulation techniques and field data. A common study area is used for both types of experiments. The NETSIM model was selected as the best simulation tool available. Recent enhancements of the model, expected to be incorporated in its next release, allow for the simulation of surveillance detector information such as vehicle counts, percentage occupancy values, and average spot speed. Several experiments are performed to incorporate variations in entry flows and turning movement percentages, as well as the randomness of traffic phenomena. Besides the simulation experiments, field studies are carried out as part of a validation effort. These studies include on-site travel time data collection and concurrent detector output consideration for detectorized links in the study area. The explanatory analysis presented indicates that both approaches support the following conclusions: (a) travel time is independent from both flow and occupancy under conditions of low traffic demand, and (b) generalized regression equations can be fitted for certain ranges of occupancies to properly model the relationships between travel time and detector data.

The main goal of the intelligent vehicle-highway systems (IVHS) program is to develop and implement state-of-the-art vehicle-highway management techniques and control systems that will reduce congestion by best using the existing infrastructure (*1*). IVHS progress will be achieved by successfully integrating advanced technology and information with conventional infrastructure to provide an expanding set of services (*2*). This evolution increases the interaction among traffic management, traveler information, and vehicle control systems calling for functional integration of IVHS components, particularly advanced traffic management systems (ATMS) and advanced traveler information systems (ATIS).

In this context, data from loop detector systems, traditionally used to optimize traffic signals, can be also used to assist several components of ATIS technologies. Loop detectors can facilitate the creation of historical data profiles, supplement instrumented vehicle (probe) reports with on-line information on traffic conditions (especially on links for which such reports are infrequent), provide a sound alternative data source to be used in the absence of vehicle reports, and improve the accuracy of data fusion (*3*). Data fusion is a mechanism engaged in combining data from various sources (on- or off-line) in order to provide improved travel time estimates. The latter will be used to advise trip makers on route choices or reroute vehicles around incidents and traffic congestion.

This paper addresses the effectiveness of applying loop detector information to route navigation systems. The main interest of the research is to indicate appropriate ways to convert detector output into arterial travel times. Loop detectors currently provide flow and occupancy data, as opposed to all other data sources involved in data fusion, which provide travel time information. Simulated and empirical data are used to study the correlation between arterial link travel times and flow and occupancy data. The observed relationships are interpreted, and general guidelines are given regarding the formulation of models capable of transforming detector data into travel times. It is cautioned that the factors involved are numerous and complex, and substantial research is needed to investigate the relationship completely. Detailed models expressing such relationships will be considered in future research.

The main goal of this paper is to justify the use of information from closed-loop signal systems, in order to enhance the performance of critical ATIS functions, such as estimating arterial link travel times. The paper aims at reporting correlations between through-movement link travel time and detector flow and occupancy observed for arterial streets. This is expected to enhance knowledge on such relationships, which is currently limited as indicated in the literature research (see the paper by Sisiopiku and Rouphail in this Record). Another objective is to offer specific guidelines and suggestions on the development of models that convert flow/occupancy data from fixed detectors on arterial streets to travel time estimates.

Herein, NETSIM capabilities are described briefly, and the parameters of interest are defined. Basic guidelines for the experiment designs are provided, including specification of the test area characteristics, description of the simulation experiments, and organization of field studies. The results obtained from the simulation analysis are demonstrated and interpreted, and simulated and observed traffic patterns are compared. Finally, concluding remarks and directions for future research are presented.

APPROACH

The first step toward studying the correlation between arterial link travel time and detector output was to collect several sets of simulated data using an appropriate simulation model. A wide array of

V. P. Sisiopiku, Intelligent Vehicle-Highway Systems Laboratory, University of Illinois at Chicago, Department of Electrical Engineering and Computer Science (M/C 154), 1120 Science and Engineering Offices, Box 4348, Chicago, Ill. 60680. N. M. Rouphail, Department of Civil Engineering, North Carolina State University, P.O. Box 2908, Raleigh, N.C. 27695. A. Santiago, Traffic Systems Branch, Federal Highway Administration, 6300 Georgetown Pike, McLean, Va. 22101.

inexpensive, widely disseminated, and highly elaborate packages is available. The primary function of these packages is to support the analysis, design, and evaluation of activities related to traffic systems operation and control. On the basis of the specifications and requirements of this study, the NETSIM simulation model was selected for application.

Briefly, NETSIM is a FORTRAN-based simulation model that describes in detail the operational performance of vehicles traveling in an urban traffic network. It is a microscopic computer software program that simulates individual vehicular behavior in response to factors such as traffic volumes, signal operations, turning movements, intersection configurations, bus operations, lane closures, and more (*4*). NETSIM simulates individual vehicle movements according to car-following, queue discharge, and lane-changing laws (*5*).

The input requirements of NETSIM include network supply features, traffic demand patterns, and traffic control information. The network is made up of directional links and nodes, and physical features of each link must be specified (*6*). The traffic demands are entered as input at entry nodes and turning proportions at intersections. The output provides traffic performance characteristics including travel times, delays, number of stops, vehicle queues, and environmental measures. Link- and movement-specific data are available.

Recent enhancements to NETSIM (Version 4.0) allow surveillance detectors to be simulated. The position and type of detector (passage or presence) is set by the user, as is the desired frequency for surveillance detector intermediate output reports. Surveillance statistics (vehicle counts, cumulative on-time, percentage occupancy values, and average speeds) are calculated for every detector during each evaluation period. All information is collected the instant that a vehicle actuates the detector. Surveillance detector capabilities are expected to be available to NETSIM users in the next release.

The increasing necessity of, and reliance on, traffic simulation as a tool for evaluating and designing advanced traffic control requires the systematic reevaluation of the underlying relations. Data collection is therefore desirable to calibrate the various relations now in use in various simulation models (*7*). For this reason, NETSIM was calibrated for an arterial segment in Chicago's northwest suburban area. The test segment operates under closed-loop signal control, is extensively detectorized, and offers a representative spectrum of physical features and traffic demand patterns.

The actual arterial segment was also used to carry out field data collection studies. This is necessary in order to improve the acceptability of and increase the confidence in the implementation of simulation research results. The focus of the data collection plan was to gather simultaneously through movement travel times and flow/occupancy data from a number of loop detector systems on the test segment. In addition to serving validation needs, the developed data base allowed preliminary testing of the actual correlations between travel time and detector output and enabled comparisons between real data and simulation output.

DEFINITIONS

Several of the parameters of interest in this paper have been defined in many ways in earlier research works. The definitions used herein follow.

Time period is defined as the amount of time during which data describing traffic flow control and characteristics are assumed to remain constant. A period of 15 min is selected as the observation period for both simulation and field studies. *Flow* is defined as the number of vehicle counts collected from one detector over each 15-min period. This definition holds for both simulated and empirical data. Flow is expressed in vehicles per lane per 15 min.

Percentage occupancy determines the vehicle presence within a detection zone. It is defined as the percentage of time that a detector is occupied by vehicles over each 15-min period.

Travel time in seconds per vehicle is defined as the time that it takes the average through vehicle to travel a distance equal to one link length. Travel time is viewed as the summation of cruise time and average delay (due to signal control existing in the intersection). *Cruise time* is the idealized travel time of an average vehicle if all vehicle trips on the link were performed at the mean free-flow speed.

DESIGN OF EXPERIMENTS

Study Area

A segment of Dundee Road, in the northwest Chicago suburbs, was selected for both simulation and field studies. The segment covers a total length of approximately 2.85 mi, including 11 signal-controlled intersections, 2 of which are considered entry and exit points for the simulated network. Information on link attributes such as link lengths, number of lanes, existence and length of left-turn pockets, and detector locations are given in Tables 1 and 2. The geometric characteristics of the study area are illustrated graphically in Figure 1.

The study segment provides a variety of geometric and traffic demand conditions. Two of its intersections are at expressway ramps, two at arterial cross streets, and seven at collector cross streets and frontage roads. Daily traffic counts on Dundee Road range from 32,000 to 47,000 vehicles per day (*8*); morning peak occurs in the eastbound direction. Most of the links include left- or right-turn pockets with a minimum of two lanes in the midblock. The entire network is on a level grade, and the posted speed limit on Dundee links is 40 mph. Sixteen presence-type loop detectors exist in the field within the limits of the area of interest. Real-time vehicle counts and occupancy data are gathered from Dundee Road closed-loop signal system detectors daily on a routine basis over 15-min intervals. In the field, the system operates under semi-actuated signal control with two-way progression.

Several data sources are used to gather the information needed for NETSIM calibration. Information on the physical properties of Dundee Road is collected by on-site visits. Such information includes link lengths (measured from stopline to stopline), number of lanes, lane channelization, existence of turning pockets, length and type (right or left) of turning pockets, and locations of detectors. Phasing information used during the coding process is obtained from an on-line data base made available by the Illinois Department of Transportation (IDOT). Signal timing plans were updated as of December 1992. Offsets are determined by a signal coordination and timing study (*8*), which also provided information on traffic flow, and turning movement percentages.

Some assumptions regarding free-flow speed, ideal saturation flow rate, driver behavior, and traffic composition were made, since field calibration of such parameters was unavailable. The accuracy of these assumptions—as well as the use of data from various sources collected at different times and for several purposes—raises

TABLE 1 Study Links Attributes

Dundee Corridor - Eastbound Links					
Link ID	Length (ft)	Lanes in Midblock	Turn Pocket Type	Turn Pocket Length	Detector Setback[a](ft)
(31,1)	1210	2	Left	300	
(1,2)	660	2	Left	300	
(2,3)	1270	2	Right[b]	140	350
(3,4)	610	3	Left	300	350
(4,5)	1190	2	Left	300	
(5,6)	1230	2	Left	140	
(6,7)	1710	2	Left	200	
(7,8)	2680	2	Left[c]	300	2334
			Right[b]	300	
(8,9)	2490	2	Left	80	
(9,39)	2030	2	Left	80	350

Dundee Corridor - Westbound Links					
Link ID	Length (ft)	Lanes in Midblock	Turn Pocket Type	Turn Pocket Length	Detector Setback[a](ft)
(39,9)	2030	2	Left	100	1676
(9,8)	2490	2	Left	120	
(8,7)	2680	2	Left	300	350
(7,6)	1710	2	Left	120	
(6,5)	1230	2	Left	100	
(5,4)	1190	2	Left	160	350
(4,3)	610	2	Right[b]	300	
(3,2)	1270	3	Left	300	
(2,1)	660	2	Left	300	350
(1, 31)	1210	2	Left	300	

[a]Distance Between Detector and Stopline
[b]Channelized
[c]Dual Left Turn Lane

some concern about the compatibility of the data used and their agreement with actual field conditions. These issues should be kept in mind when comparing simulated output with real data.

Simulation Experiments

A set of simulation experiments is performed to provide appropriate data for the study of the correlation between link travel time and detector output. Two factors are selected for consideration, namely, entry link flows and turning movement percentages.

A base case scenario is first constructed assuming entry link flows and turning movement percentages similar to those obtained from turning movement counts collected for an IDOT SCATS study (*8*). Using the base case as a starting point, seven different levels of each factor are considered. Changes to the entry flows by ± 10, ± 20, and ± 30 percent relative to the base case scenario yield seven flow levels. Similarly, the percentage of left-turn movements initially considered for each link is varied by ± 10, ± 20, and ± 30 percent. It is assumed that the percentage of right turns remains unaffected throughout the simulation experiments and thus any changes in left-turn movement percentages directly reflect changes in through movement percentages.

A full-factorial design is applied. Since there are two factors with seven levels each, a 7^2 factorial design was performed yielding in 49 combinations of the different factor levels considered in the experiment. On the basis of recommendations reported by Sisiopiku (*3*), three simulation replications with a different random number seed are performed for each run to account for the randomness of traffic phenomena. Each run simulates 1 hr of operations based on data from the morning peak (7:30 to 8:30 a.m.). To model the observed variability of traffic flows within the peak hour, the simulation interval is divided into four 15-min periods, and entry flows and turning movement percentages are given for each 15-min period. Traffic conditions are assumed to remain stationary within each 15-min period.

Overall, 147 NETSIM runs are performed, providing a large amount of output. The output is reduced in size, reorganized, and compressed using two custom-built computer programs. Automation of the data-handling mechanism produces great advantages including convenience, increase of processing speed, and guaranteed accuracy. Link-by-link analysis followed. The results obtained are discussed in a later section.

Field Studies

Empirical travel time data were also collected through on-site surveys on the Dundee Road segment. Concurrent detector data were gathered from inductive loop detector systems located on four through links, namely, links (2,1), (2,3), (3,4), and (5,4). Detector logs were made available through IDOT, which has on-line access to information for several closed-loop systems on freeways and arterials in the Chicago metropolitan area.

A variety of operational conditions was desired, from free flow to highly congested flow. A review of historical flow and occupancy profiles and a reconnaissance study indicated that the morning and afternoon peaks normally extend from 7:00 to 9:00 a.m. and from 4:30 to 6:30 p.m., respectively (*9*). To incorporate the transition to and from congested states, data are collected in 3-hr blocks from 7:00 to 10:00 a.m. or from 3:30 to 6:30 p.m. for five typical weekdays, yielding twelve 15-min periods a day (or 60 intervals).

Initial travel time data collection took place using the average-car technique as defined elsewhere (*10*). While the driver was traveling on the average speed of the traffic stream, an on-board observer recorded link travel times using a stopwatch.

Two problems were encountered using this data collection effort:

1. As in all floating car techniques, the driver's perception of the average speed introduces some bias on the data collected. For example, a shift of ±1 standard deviation is easy to effect by varying slightly the rules by which acceleration and deceleration decisions are made.
2. The test-vehicle data collection provided two to three travel time observations per 15-min interval for each target link. Conclusions based on small sample sizes of data are generally of limited value, and initial analysis indicated that high fluctuations occur within each 15-min period. Averages based on limited observations often do not represent the population mean, and conclusions derived from them may be misleading.

To overcome these problems, supplementary data collection was performed using license plate matching. This method allows for collection of a significant amount of data for each 15-min interval. When this approach was used, sample size increased considerably (10 to 46 observations per interval), allowing confidence in arguing that the average values obtained per 15-min interval are representative of the population mean.

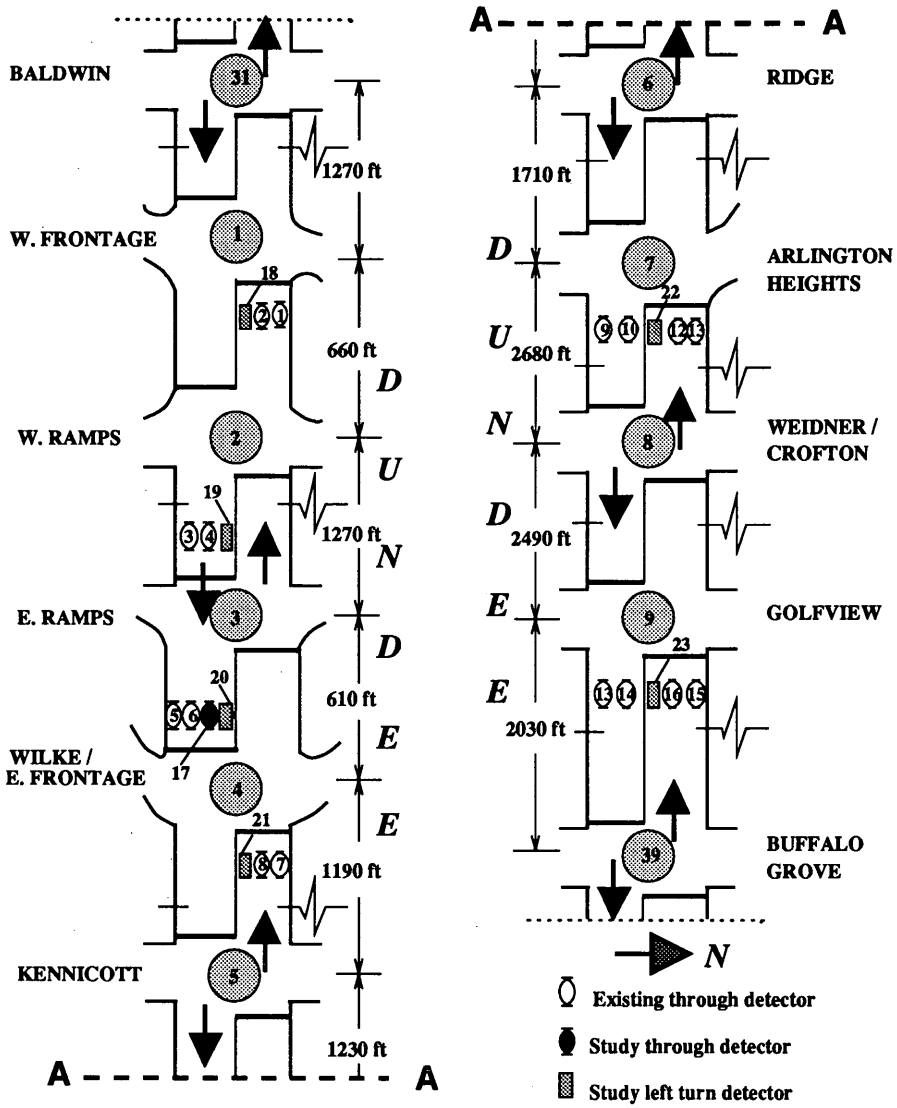


FIGURE 1 Schematic representation of test area.

RESULTS

Simulation Analysis

The analysis focuses on simulated data for through movements on detectorized links. As shown in Figure 1, eight such links are available in the study area and cover a variety of traffic demand conditions. Link (5,4) covers virtually the entire spectrum of simulated occupancies (from 4 to 96 percent) and is chosen as reference. Even though detailed qualitative assessment of simulated data is offered for this link only, final conclusions are derived on the basis of observations from all study links.

So that the relationships between simulated travel time and detector output could be observed, several graphs have been prepared, including plots of travel time versus flow and occupancy, and flow and spot speed versus occupancy. Each data point in these plots corresponds to the population average of all through vehicles on the link over a 15-min period. Data are collected on a detector-by-detector basis. At the end of each 15-min period, the data for each detector are averaged. Then the mean value from all through detectors on the same link is calculated. This value is reported in the plots as the average for the subject link and the subject 15-min interval.

Figure 2 depicts the relationship between flow and percentage occupancy. Under uncongested flow conditions, flow increases linearly with occupancy. In the near-capacity flow regime, flow stabilizes around link capacity value while occupancy increases. Finally, in the forced flow regime, flows actually drop as occupancy increases. For simulated occupancies of more than 90 percent, flow decreases at a higher rate. Occupancies in that range are indicative of queue spillback onto the detector due to highly congested conditions exacerbated by signal delays or incidents.

The relationship between travel time and flow observed from simulation is graphically illustrated in Figure 3. Under uncongested conditions, travel time is practically independent from flow. This becomes evident when the fairly dense and almost horizontal data pattern that exists under conditions of low traffic demand is considered. As demand increases, travel time increases with flow (albeit slightly) until capacity conditions are reached. After that point, travel times increase rapidly but flows drop. This observation is in compliance with traffic stream fundamentals, according to

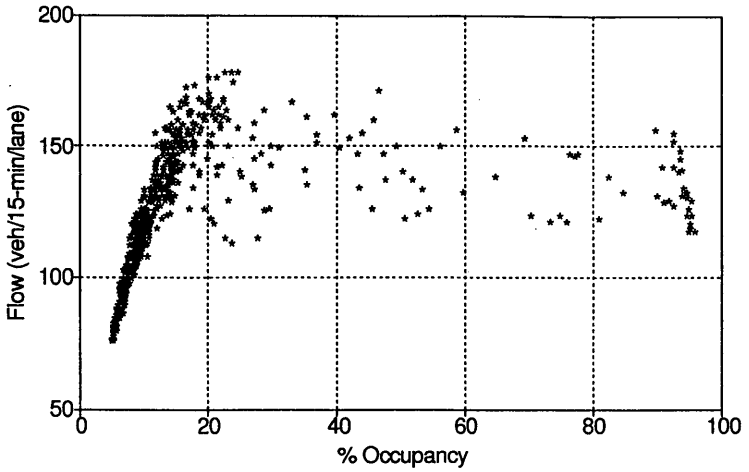


FIGURE 2 Flow versus occupancy (simulated data).

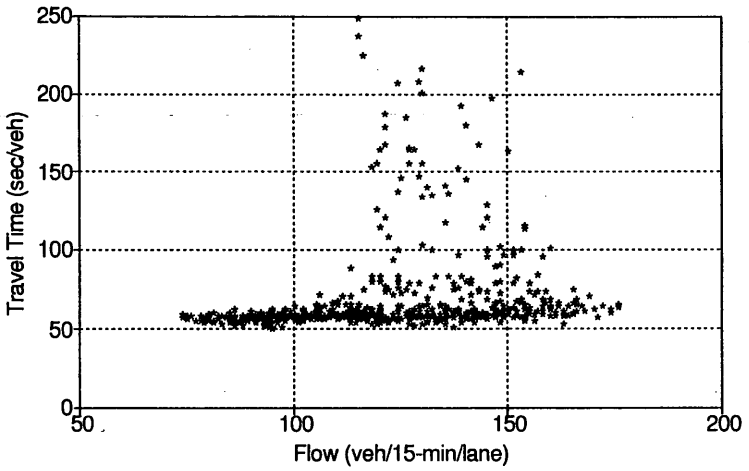


FIGURE 3 Travel time versus flow (simulated data).

which low flow rates can result from either very low concentration (low demand, uncongested conditions) or very high concentration (demand exceeding capacity, congested operations). Overall, a nonlinear pattern is observed.

Figure 4 depicts the relationship between travel time and percentage occupancy as observed from simulation runs. Under conditions of low demand, no significant correlation between travel time and occupancy values is observed. As occupancy increases, a linear dependency between travel time and occupancy level becomes clear. This relationship collapses for high occupancy values (more than 90 percent), for which detectors are actually blocked by standing vehicles and travel time is very unstable and practically unpredictable.

An exponential relationship is observed from the plot of average detector spot speed against occupancy, as shown in Figure 5. Speed drops as occupancy increases with a higher rate for occupancy values in the uncongested regime. Under oversaturation, speed stabilizes near a value of approximately 10 mph. Further reduction is observed for occupancies of more than 90 percent due to vehicles actually stopped over the detection zone.

Validation with Empirical Data

Empirical data collected from floating car and license plate matching techniques were used to gain an insight on the actual relationships between travel time, flow, and occupancies. In addition, the field data enabled comparisons between simulated and real traffic conditions, assisting performance evaluation of the simulation model.

The relationships between flow, occupancy, and travel time, both from simulation and field studies, are shown in Figure 6. When attempting to interpret the results, one should keep in mind variations in the amount of data available and the range of conditions represented in simulation and field runs. Simulated data are available for 588 15-min intervals and for a variety of traffic flow and movement percentage combinations, but test vehicle field data are limited to 60 15-min observations under uncontrolled traffic patterns.

Another issue worth noting is related to the differences in stability between simulated and observed data. Plots based on simulated output show patterns much more dense than those from field data,

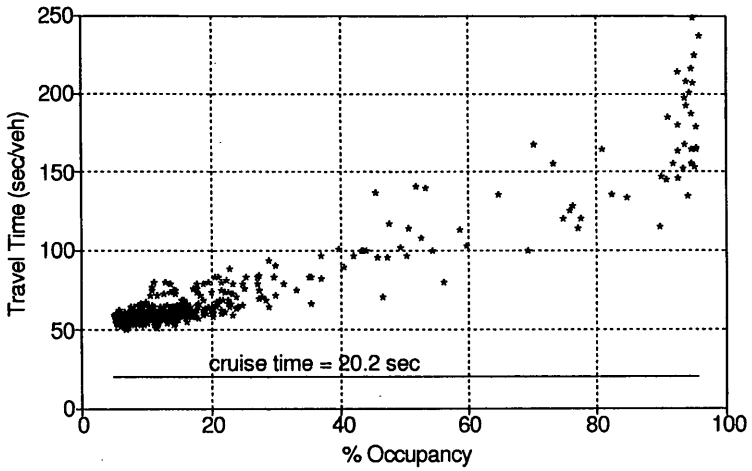


FIGURE 4 Travel time versus occupancy (simulated data).

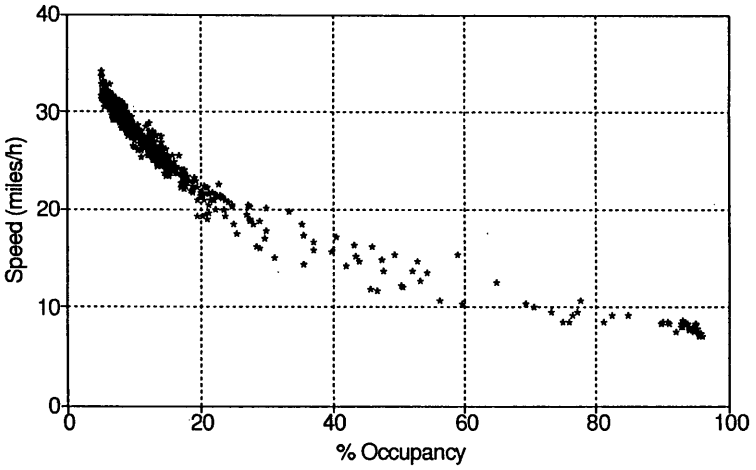


FIGURE 5 Speed versus occupancy (simulated data).

simply because in the former, travel times are averages of the entire population, whereas in the latter, they are based on a small sample of observations.

Careful consideration of the plots displayed in Figure 6 identifies a conservative tendency of NETSIM model. Observed capacity values are much higher than simulated, observed flows are higher than simulated for similar occupancy values, and simulated delays are consistently higher than observed. Although differences in absolute values exist, NETSIM manages to simulate traffic relationships between arterial link travel times and flow/occupancy that are similar to those observed from field data.

Thus, observed differences between simulated and observed values are due primarily to insufficient model calibration rather than actual deficiency in the simulation code to properly model traffic processes. However, further investigation is required to identify the main cause of such behavior.

The plot of field flow data against corresponding occupancy values indicates a strong correlation between flow and occupancy values. The nonlinear relationship between simulated travel times and flows is also supported by the empirical data.

Observed travel times under uncongested traffic conditions are close to or even below the assumed cruise travel time. This shows that the assumption of fixed link cruise time calculated from a mean free-flow speed equal to the posted speed limit is not valid. Calibration of the actual cruise travel time value is expected to narrow considerably the deviations between observed and simulated travel times under free-flow operations.

Empirical data indicate that travel time is linearly related to percentage occupancy for occupancies of approximately 17 to 60 percent. For occupancies below the lower bound, travel time is virtually independent from the occupancy value; however, occupancies of more than 60 percent have not been observed in the field and therefore no conclusions can be drawn.

As data are averaged over 15-min periods, a number of phenomena, often encountered under interrupted traffic flow conditions, may be difficult to recognize. Effects from cycle failures, short-term events, congestion built up downstream of the subject link, and so forth cannot be detected using 15-min observation periods. The main reason for selecting 15-min periods instead of 5- or even 1-min periods was the desire to comply with real data currently

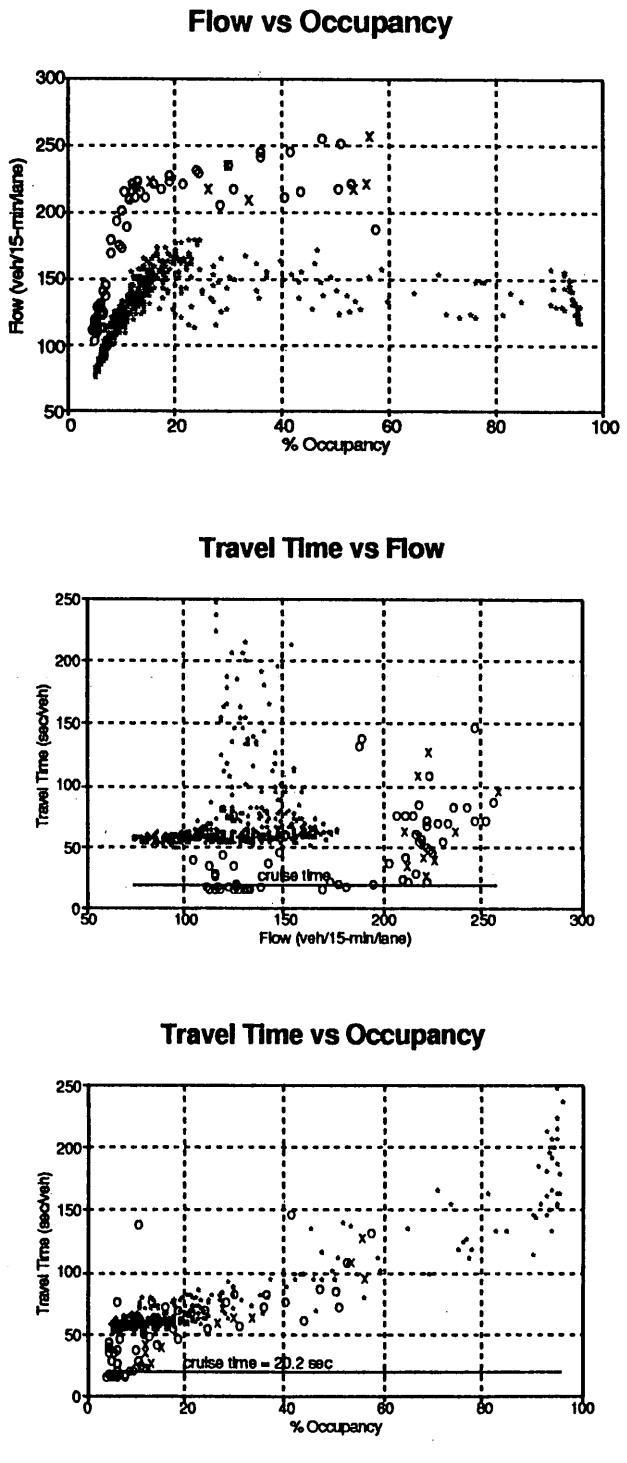


FIGURE 6 Flow, occupancy, and travel time plots (simulated and empirical data).

available through existing loop detector systems. Such systems provide flow and occupancy information as averages over 15-min intervals. However, for real-time ATIS applications the inherent assumption of traffic flow stability over a 15-min period is questionable and requires further proof. This is a topic recommended for future research.

CONCLUSIONS AND RECOMMENDATIONS

The purpose of this investigation was to test the feasibility of using detector data to develop arterial travel time models for ATIS applications. The availability of real-time data from closed-loop signal system detectors is expected to enhance the quality and validity of travel time and incident information distributed by a traffic information center. Because real-time data from probe vehicles are often sporadic, fixed detector data will constitute an important source of periodic information for estimating dynamic travel time (*11*).

The sharing of traffic data between traffic surveillance and control and traveler information systems is an excellent example of the benefits of functional integration of IVHS components. Loop detector data traditionally used for optimizing traffic signals can play an important role in predicting dynamic travel times. Traffic volume counts and occupancy data from closed-loop systems allow for a more comprehensive understanding of overall traffic conditions within a roadway corridor, which in turn enables system operators to make improved decisions to facilitate traffic flow (*12*).

This paper studied the relationships between arterial through link travel times and detector output based on simulation and field studies. The observed relationships should be formulated mathematically to provide models capable of converting detector data (flows and occupancies) into travel times. Travel times based on detector information will be used with probe vehicle reports and historical data in order to derive improved estimates of travel time through data fusion.

Study of the correlations between arterial through link travel times and flow/occupancy values reveals the following:

1. Travel time is independent of both flow and occupancy under conditions of low traffic demand. Under such conditions, detector data have limited significance to the process of forecasting arterial link travel times.
2. As percentage occupancy increases, the correlation between travel time and occupancy becomes more significant. For occupancies above certain threshold values, detector data can be used to supplement travel time information on detectorized urban links.
3. Regression equations can be fitted for certain ranges of occupancies to properly model the observed relationships between travel time and detector occupancy.
4. Both simulation and field data indicate a strong correlation between flow and occupancy for certain ranges of values, restricting the use of both parameters in the same regression model.
5. Percentage occupancy is viewed as a better predictor for link travel time than traffic flow and is thus recommended as a superior explanatory variable.
6. In general, travel time and detector occupancy are related linearly. If travel time and flows are considered instead, nonlinear terms must be present in the regression equation.
7. For occupancies over a threshold value of approximately 90 percent, travel time predictions are not possible, as travel time is unpredictable due to the queues that persist over the detector location.

8. Relationships between travel time and detector flow/occupancy data, based on field observations, show similar patterns to those from simulation for the same ranges of operation.
9. Proper calibration of NETSIM prior to simulation runs is mandatory if the program is used to provide the data base used to model the relationships between travel time, flow, and occupancies.
10. Empirical data bases are superior to simulated ones for model calibration, given that adequate data samples are obtained to ensure that the population average conditions are represented properly.

Although the relationships between travel time and detector information are studied and described in detail, mathematical formulations are not given in this paper. Development of statistical models based on the previous observations and recommendations needs further study and will be addressed in future research. If the models under development are intended for ATIS applications, the complexity of the forms used and on-line data sources available should be limited.

Several model forms can be applied. Current knowledge indicates that linear regression equations with travel time or delay as the dependent variable and occupancy as the independent variable can give reasonable fits for occupancies for a wide range of occupancy values as far as the models are link-specific. Otherwise, the effects of other parameters need to be also incorporated including link length, detector location, effective green time, progression quality, etc.

It is also recommended that empirical data be used to calibrate the regression model parameters. Empirical data should be used when the data collection methods provide samples that are representative of the population average. Alternatively, NETSIM simulation model can be used but only after extensive calibration and validation. This is recommended to enhance the credibility of simulation results by representing the actual traffic conditions more realistically.

The analysis performed in this study focuses on observations from through movements. Turning movements are expected to show some differences in the relationships between travel time and detector data and need to be studied separately. For example, left-turning vehicles proceeding under permissive phasing are expected to suffer higher delays than through vehicles for similar flow and occupancy levels. This is due to additional factors that significantly affect travel time, such as opposing flow and driver behavior.

Research is also required toward the development of travel time functions for estimating route travel times based on information from loop detectors located on a number of links composing the route.

ACKNOWLEDGMENTS

The financial support of this research by FHWA and the ADVANCE project is greatly appreciated.

REFERENCES

1. Santiago, A. J. ATMS Technology—What We Know and What We Don't Know. Presented at the 2nd Annual Meeting of IVHS America, 1992.
2. *Strategic Plan for Intelligent Vehicle Highway Systems in the United States*. IVHS America, Washington, D.C., 1992.
3. Sisiopiku, V. P. *Arterial Link Travel Time Estimation Based on Loop Detector Output*. Final Report, FHWA Grant DTFH61-92-P-40029. FHWA, U.S. Department of Transportation, 1993.
4. Wong, S.-Y. TRAF-NETSIM: How It Works, What It Does. *ITE Journal*, Vol. 60, No. 4, 1990.
5. Skabardonis, A., 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.
6. May, A. D. *Traffic Flow Fundamentals*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1990.
7. Mahmassani, H. S. *Traffic Simulation Models*. ORSA Workshop Report, Orlando, Fla., 1992.
8. Bucher, Wills, and Ratliff. *1990 Signal Coordination and Timing (SCATS) Study*. Job D-91-136-90. District 1, Illinois Department of Transportation, Aurora, 1991.
9. Rouphail, N. M., and V. P. Sisiopiku. *Travel Time and Loop Detector Output Analysis on Dundee Road Closed-Loop Signal Systems*. ADVANCE Working Paper Series 24. ADVANCE, Chicago, Ill., 1993.
10. *Manual of Traffic Engineering Studies*. ITE, Arlington, Va., 1976.
11. Bhat, C. R., F. S. Koppelman, R. Laver, and S. S. Shbaklo. *Initial Static Travel Time Data Base Development: Probe Vehicle Coverage Intensity Plan and Data Base Assembly Procedures*. Final Report AVD-BD-02. Evanston, Ill., 1993.
12. Arch, J. *The Role of Traffic Signal Systems in IVHS*. PB Network, Trenton, N.J., 1992.

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

Development and Comparative Evaluation of High-Order Traffic Flow Models

ANASTASIOS S. LYRINTZIS, GUOQING LIU, AND PANOS G. MICHALOPOULOS

Five high-order continuum traffic flow models are compared: Payne's original model, Papageorgiou's improved model, the semiviscous model and the viscous model, a proposed high-order model, and the simple continuum model based on the pipeline cases. The stability of the high-order models is analyzed and the shock structure investigated in all models. In addition, the importance of the proper choice of finite-difference method is addressed. For this reason, three explicit finite-difference methods for numerical implementation—the Lax method, the explicit Euler method, and the upwind scheme with flux vector splitting—are discussed. The test with hypothetical data and the comparison of numerical results with field data suggest that high-order models implemented through the upwind method are more accurate than the simple continuum model. For congested cases, the proposed high-order model appears to be more accurate than the other high-order models for all cases tested.

Since Lighthill and Whitham (*1*) first applied a simple continuum model to describe the characteristics of traffic flow in 1955, much progress has been made in the development and application of macroscopic continuum traffic flow models, especially with the introduction of the high-order continuum models. For example, since 1985, Michalopoulos et al. (*2–4*) have developed a microcomputer simulation program, KRONOS, based on the simple continuum model. KRONOS has been used by the Minnesota Department of Transportation for simulating freeway traffic. In 1971 Payne developed a high-order continuum model that includes the effects of the drivers' reaction and acceleration (*5*). Later he applied this high-order model to the computer simulation program, FREFLO (*6*). Since then, researchers in traffic flow theory have developed a few new high-order continuum models. Examples are Papageorgiou's improved high-order model (*7,8*), the semiviscous and viscous high-order models (*9,10*), and others (*11–13*).

As is well known, high-order continuum models are more sophisticated than the simple continuum model because the simple continuum model is based only on the conservation equation, but high-order models include not only the conservation but also the momentum equation, which accounts for the dynamic effects of inertia and acceleration of traffic mass. However, it is unknown whether in practice high-order continuum models produce better results than those of the simple continuum model.

Although it is well understood that a finite-difference method can affect the computational accuracy of continuum traffic flow models, the importance of the proper choice of finite-difference method was not addressed properly in the past, and some improper finite-difference methods were applied to the continuum models. Only recently have other finite-difference methods been applied to the continuum models. For example, some implicit methods for the simple continuum model and the semiviscous model are discussed by Chronopoulos et al. (*14,15*). Leo and Pretty (*16*) used an upwind method for the simple continuum model and Payne's model. In addition, Lyrintzis et al. surveyed the application of upwind methods including the total variation diminishing (TVD) method to the simple continuum model (*17*). Although the implicit first-order upwind scheme is strongly recommended for the simple continuum model, it is not clear which finite-difference method should be used with the high-order continuum models to achieve a higher computational accuracy. The purpose of this paper is to address these questions.

The five high-order continuum models have been investigated and compared with the simple continuum model for the pipeline cases. These five high-order models are Payne's high-order model (*5,6*), Papageorgiou's improved high-order model (*7,8*), the semiviscous model, and the viscous model (*9,10*) as well as a new high-order model developed here. The stability of the high-order models is analyzed and the shock structure investigated in all models. Three explicit finite-difference methods—the Lax method (*18*), the explicit Euler method, and the upwind scheme with flux vector splitting (*19,20*)—are discussed. Through mathematical analysis, testing with hypothetical data, and comparison of numerical results with field data, it is demonstrated that high-order models implemented through the upwind scheme with flux vector splitting can perform better than the simple continuum model. Furthermore, the proposed high-order model appears to be more accurate than the other high-order models.

CONTINUUM TRAFFIC FLOW MODELS: AN OVERVIEW

Simple Continuum Model

The simple continuum model proposed by Lighthill and Whitham (*1*) consists essentially of a conservation equation for the pipeline case

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{1a}$$

supplemented by the definition of the flow rate

$$q = uk \tag{1b}$$

and a speed-density (u-k) relationship

$$u = u_e(k) \tag{1c}$$

A. S. Lyrintzis, Department of Aerospace Engineering and Mechanics, University of Minnesota, 107 Akerman Hall, 110 Union Street, S.E., Minneapolis, Minn. 55455. G. Liu and P. G. Michalopoulos, Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive, S.E., Minneapolis, Minn. 55455.

where

k = traffic density (veh/km),
q = flow rate of traffic stream (veh/hr),
u = space-mean speed (km/hr),
t = time,
x = space, and
$u_e(k)$ = equilibrium relationship between speed and traffic density.

It is well-known that Equation 1a is nonlinear and dominated by convective effects. Therefore, the simple continuum model (Equation 1) always leads to discontinuous solutions so that a smooth solution can exist only for a finite time, even when the initial condition is arbitrarily smooth. However, actual traffic flow changes smoothly. This means that, from a theoretical point of view, the simple continuum model does not accurately describe the traffic dynamics.

It should be noted that the numerical treatment of the simple continuum model can introduce numerical dissipation to smooth the discontinuous solution. For example, the Lax method (*18*) can be used because it introduces a strong dissipation effect to the simple continuum model (*2,19*). Although the dissipation effect might be introduced to the simple continuum model by using the other finite-difference methods, there is still another drawback to the simple continuum model. That is, changes in speed in the simple continuum model occur instantaneously and fluctuations of speed about equilibrium values are not allowed. This problem cannot be overcome by using a finite-difference method. Nevertheless, the simple continuum model usually captures the basic shock wave structure and gives reliable results for various test cases and geometries (*4*).

Original High-Order Model

Payne proposed a more attractive high-order continuum traffic flow model in which a momentum equation is included (*5*). This model is called the original high-order model. The momentum equation in this model was derived from car-following theory. The state equations of the original high-order model for the pipeline case are

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{2a}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{1}{T}\left\{u_e(k) - u - \frac{\nu}{k}\frac{\partial k}{\partial x}\right\} \tag{2b}$$

$$q = uk \tag{2c}$$

where T is the constant reaction time and ν is an anticipation coefficient that is the function of the density with the following form

$$\nu = -\frac{1}{2}\frac{du_e}{dk} \tag{3}$$

It should be noted that a constant anticipation coefficient was later suggested by Payne (*6*).

Since the momentum equation (Equation 2b) is included in the original high-order model, some new features emerge. First, by using the linearized theory to the original high-order model, it can be seen that an equilibrium state $[k_0, u_0 = u_e(k_0)]$ exists in the original high-order model if the following condition holds (*21*):

$$u_0 + \sqrt{\frac{\nu}{T}} > c_0 > u_0 - \sqrt{\frac{\nu}{T}} \tag{4}$$

where c_0 is the kinematic wave speed and

$$c_0 = u_0 + k_0\left(\frac{du_e}{dk}\right)_{k_0} \tag{5}$$

Another new feature is that there is a smooth shock. That is, the shock can be represented by

$$\xi = x - Ut \tag{6}$$

where U is the constant speed of the smooth shock. To see this, substituting Equation 6 into Equation 2 (for the pipeline case) yields a single equation for the density

$$[\nu - T(u - U)^2]\frac{dk}{d\xi} = k[u_e(k) - u] \tag{7}$$

It has been proved (*21*) that Equation 7 has a unique solution if and only if

$$\nu > T(u - U)^2$$

that is

$$u - \sqrt{\frac{\nu}{T}} < U < u + \sqrt{\frac{\nu}{T}} \tag{9}$$

In other words, Equation 6 does represent the smooth shock for the original high-order model if the condition (Equation 8) holds. Moreover, the coefficient $[\nu - T(u - U)^2]$ determines the shock thickness that represents the space containing the shock. The larger the value of the coefficient, the thicker the shock, and vice versa. It should be pointed out that if the condition (Equation 8) does not hold, the smooth shock does not exist but a discontinuity occurs.

From this discussion, it can be seen that the original high-order model is superior to the simple continuum model conceptually. Unfortunately, because the explicit Euler-like finite-difference method was applied to the original high-order model (*5*), application of this model does not show the superiority. Indeed, applying the explicit Euler-like method to the original high-order model (Equation 2) yields

$$k_j^{n+1} = k_j^n + \frac{\Delta t}{\Delta x}[q_j^{n+1} + q_{j+1}^{n+1}] \tag{9a}$$

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x}u_j^n[u_j^n - u_{j-1}^n] + \frac{\Delta t}{T}\left\{u_e(k_j^n) - u_j^n - \frac{\nu}{k_j^n \Delta x}[k_{j+1}^n - k_j^n]\right\} \tag{9b}$$

$$q_j^{n+1} = \frac{1}{4}[k_{j-1}^n + k_j^n][u_{j-1}^n + u_j^n] \tag{9c}$$

From this discretized form, it is evident that the original high-order model cannot work at the smaller values of the density because of the term $\nu[k_{j+1}^n - k_j^n]/k_j^n \Delta x$. Since this discretized form does not come from the conservation form of the system, it cannot produce the correct shock intensities (*22*). Moreover, this discretized form is

unstable from the computational point of view. To see this, investigate the truncation error associated with Equation 9a. Here only those terms that involve a second-space derivative of the density k are needed, since they are the only ones that contribute to a diffusion. The effective diffusion coefficient for Equation 9a, through terms of order Δt^2 and Δx^2, is

$$-\frac{\Delta t}{2}u^2\left(1-\frac{\Delta t}{T}\right)-\frac{\Delta t}{2}\frac{v}{T}\left(1-\frac{\Delta t}{3T}\right)$$
$$+\left\{\frac{\Delta t^2}{6}\left(9u^2+\frac{7v}{T}\right)-\frac{\Delta x^2}{4}\right\}\frac{\partial u}{\partial x}-\frac{\Delta t^2}{6}\frac{u}{T}\left(5k\frac{du_e}{dk}+3u_e\right) \quad (10)$$

It has been proved that instabilities can occur wherever a diffusion coefficient is negative (*23*). From Equation 10, it is easy to see that the first two terms are always negative; the third term will be negative when traffic becomes congested. Thus, under congested flow, the computed solutions provided by this discretized form of the original high-order model become unstable. Therefore, in order to implement the original high-order model effectively, an alternative finite-difference method is needed.

Although another finite-difference method can be applied to the original high-order model to improve its performance, there is still a problem in the model: reaction time. As car-following theory suggests, the reaction time is the time measured from the time at which the lead driver initiates a stop until the second driver initiates his or her own stopping maneuver. After such a time, the velocities of the two vehicles are assumed equal (*24*). This would mean that the second vehicle has a jump in speed, but this is not the case. In fact, after the reaction time, there is still a process of adjusting speed for the second vehicle, which is called the relaxation process. Such a relaxation process is not included in the original high-order model because only the reaction time is taken into account. For this reason, the authors propose the new high-order continuum model presented in a later section.

Improved High-Order Model

On the basis of the original high-order model, Papageorgiou (*7,8*) proposed an improved high-order continuum model. The equations of this improved high-order model for the pipeline case are

$$\frac{\partial k}{\partial t}+\frac{\partial q}{\partial x}=0 \quad (11a)$$

$$\frac{\partial u}{\partial t}+u\zeta\frac{\partial u}{\partial x}=\frac{1}{T}\left\{[u_e(k)-u]-\frac{v}{k+\kappa}\frac{\partial k}{\partial x}\right\} \quad (11b)$$

$$q=uk \quad (11c)$$

where κ and ζ are constants. The improved high-order model was based on the Euler-like discretized form of the original high-order model. So to improve the computational effect of the original high-order model, κ was added to keep the third term on the right-hand side of Equation 9b limited when the density k becomes small; ζ was added only for the numerical computation of the model.

To see the difference between the improved and the original high-order models, linearize the improved high-order model (Equation 11) for small perturbations about the state $[k_0, u_0 = u_e(k_0)]$. Thus it can be seen that the state $[k_0, u_0 = u_e(k_0)]$ is equilibrium if the following condition holds:

$$u_0+\sqrt{\frac{k_0v}{k_0+\kappa T}+\frac{1}{4}u_0^2(1-\zeta)^2}-\frac{1}{2}u_0(1-\zeta)>c_0>u_0$$
$$-\sqrt{\frac{k_0v}{k_0+\kappa T}+\frac{1}{4}u_0^2(1-\zeta)^2}-\frac{1}{2}u_0(1-\zeta) \quad (12)$$

Comparing with the stable condition (Equation 4) of the original high-order model, when $\zeta = 1$ and $\kappa \neq 0$, the range of stability of Equation 12 is less than that of Equation 4; when $\zeta > 1$ or $\zeta < 1$, the range of stability of Equation 12 shifts right or left relative to the range of stability of Equation 4.

Now the shock structure of the improved high-order model will be investigated. Using the same method as in the original high-order model,

$$\left[\frac{kv}{k+\kappa}-T(u\zeta-U)(u-U)\right]\frac{dk}{d\xi}=k[u_e(k)-u] \quad (13)$$

Thus, the smooth shock exists if and only if

$$\frac{kv}{k+\kappa}>T(u\zeta-U)(u-U) \quad (14)$$

Moreover, comparing Equation 14 with Equation 7, the shock thickness of the improved high-order model is less than or equal to that of the original high-order model because the coefficient of the left term of Equation 14 is less than or equal to that of the left term of Equation 7.

From this discussion, the improved high-order model might have more accurate computational results than the original high-order model. However, this conclusion depends on the choice of the parameters κ and ζ. Since the improved high-order model was developed on the basis of the Euler-like discretized form of the original high-order model, the discretized form of the improved high-order model still suffers the same instability problem. Moreover, the upwind scheme cannot be used with flux vector splitting (see next section) to overcome the instability problem of this model because the Jacobian is not homogeneous.

Semiviscous Model

Michalopoulos et al. (*9,10*) proposed two high-order continuum models: the semiviscous model and the viscous model. The equations of the semiviscous model for the pipeline case are

$$\frac{\partial k}{\partial t}+\frac{\partial q}{\partial x}=0 \quad (15a)$$

$$\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}=\frac{1}{T(k)}[u_f(k)-u]-\alpha k^\beta\frac{\partial k}{\partial x} \quad (15b)$$

$$q=uk \quad (15c)$$

where

$u_f(x)$ = free-flow speed,
α = positive constant (and $\sqrt{\alpha}$ has the dimension of velocity), and
β = parameter, usually chosen as -1.

Note that the first term on the right side of Equation 15b represents relaxation, which is the process whereby drivers adjust their speeds to the free-flow speeds. Thus $T(k)$ is the relaxation time, which is a function of density and is given as

$$T(k) = T_0\left[1 + \frac{\gamma k}{k_{\text{jam}} - \gamma k}\right] \tag{16}$$

where $T_0 > 0$ and $0 < \gamma < 1$ are constants and k_{jam} is the jam density. It should be noted that this relaxation term can contribute to Equation 15b only when $u_f(x)$ is changed from one section of the roadway to another.

In comparison with the previous models, the main feature of the semiviscous model is that it does not require an explicit equilibrium speed-density relationship. The semiviscous model appears to be more appealing for field applications, but, because of the simplification, new problems occur. First consider a pipeline case with a fixed free-flow speed where the relaxation term has disappeared. The semiviscous model is reduced to the perfect gas dynamic model:

$$\frac{\partial k}{\partial t} + \frac{\partial (uk)}{\partial x} = 0 \tag{17a}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -\frac{\alpha}{k}\frac{\partial k}{\partial x} \tag{17b}$$

where the value of β is chosen as -1. It has been shown that for an originally continuous compression wave, the system described by Equation 17 always yields a discontinuity (*25*). In fact, Equation 17b is Greenberg's one-dimensional fluid state equation (*26*). Thus, when the free-flow speed is fixed for the pipeline, the semiviscous model produces the same results as the simple continuum model.

Next, consider a pipeline with two free-flow speeds. In this case, one must use the the full form of the semiviscous model to describe traffic flow. If the free-flow speeds are decreasingly distributed on the pipeline, then the contribution of the first term on the right side of Equation 15b to the upstream always represents acceleration. Clearly, this is not the case. This means that the relaxation process in which the free-flow speed serves as the desired state for the adjustment of speed is incorrect. Hence, some modifications to the semiviscous model are needed.

Nevertheless, when combined with the upwind scheme with flux vector splitting (which will be referred as the "upwind method") (*19,20*), the semiviscous model appears to be working more effectively than the simple continuum model. This is because the upwind method introduces physical propagation properties in the discretization process of the semiviscous model. That is, a forward difference is used for an upstream moving wave and a backward difference for a downstream moving wave. Moreover, the upwind method still introduces a numerical viscosity into the discretized form so that shocks can be smeared out. It should be pointed out that the semiviscous model should be modified when the free-flow speed is not constant.

Viscous Model

The viscous model discussed here was proposed by Michalopoulos et al. (*9*). The equations of the viscous model for the pipeline case are

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{18a}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = -\alpha k^{\beta}\frac{\partial k}{\partial x} + \eta k^{\rho}\frac{\partial^2 k}{\partial x^2} \tag{18b}$$

$$q = uk \tag{18c}$$

where η is the viscous parameter and ρ is a dimensionless constant. The first term on the right-hand side of Equation 18b represents anticipation. The second term on the right side of Equation 18b is the viscosity term, which is used to address traffic friction. It should be noted that the viscous term always exists in the model regardless of the geometry of the freeway. In addition, the viscous model does not use the equilibrium speed-density relationship.

When the semiviscous model is compared with the viscous model, the viscous model can be derived from the semiviscous model if the relaxation term is replaced by the viscous term for the pipeline case. Indeed, both relaxation and viscosity have the same effect—smearing out of the shock. However, from gas dynamics it is known that only when the relaxation time is small can the effect of the relaxation be replaced by a corresponding bulk viscosity (*27*). As will be seen, the relaxation time in the congested traffic flow is small, whereas the relaxation time in the uncongested traffic flow is large. Therefore, the relaxation process cannot be totally replaced by viscosity. Hence, the viscous model could lead to inaccuracies.

Finally, since the Euler method was used with the viscous model, the discretized form of the viscous model is unstable because this discretized form lacks a positive mass diffusion, even though there is a viscous term in the momentum equation.

PROPOSED HIGH-ORDER MODEL

As mentioned earlier, the original high-order model considers only the reaction time and ignores the relaxation time. A question that may arise is whether the relaxation property does in fact exist in a macroscopic sense. Clearly, from the microscopic point of view, there is a process of adjusting speed for the second vehicle after the reaction time. Moreover, it has been suggested that drivers have different behavior at different density levels. For example, at low densities, interaction between drivers becomes negligible, but at high densities, the interaction becomes strong. Hence, from the macroscopic point of view, the process of adjusting speed can be considered as the process of relaxation of drivers' speed to the equilibrium speed, and the relaxation time at a high density level should be shorter than that at a low density level in order to avoid a collision. Therefore, the author proposes the following high-order continuum model for the pipeline case:

$$\frac{\partial k}{\partial t} + \frac{\partial q}{\partial x} = 0 \tag{19a}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \frac{1}{T(k)}[u_e(k) - u] - \sigma k^{\beta}\frac{\partial k}{\partial x} \tag{19b}$$

$$q = uk \tag{19c}$$

where β is a parameter. To use the upwind scheme with flux vector splitting for this model, β is chosen as -1 in order to make the Jaco-

bian homogeneous. Thus σ is an anticipation constant and $\sqrt{\sigma}$ has the dimension of velocity. $T(k)$ is the relaxation time, which should be a function of density k. Since the relaxation time at a high density level is shorter than that at a low density level, the following general function for $T(k)$ is suggested:

$$T(k) = T_0\left[1 + \frac{1}{\left(\frac{k}{k_m}\right)^{\theta}}\right] \tag{20}$$

where

k_m = critical density,
T_0 = constant reaction time, and
$\theta > 0$.

Thus, when $k \to 0$, $T(k) \to \infty$; when $k > k_m$, $T(k) \to T_0$. This means that at high density levels, the relaxation time is equal to the driver's reaction time. This formula of $T(k)$ is physically acceptable. Moreover, for simplicity, the following equilibrium speed can be used:

$$u_e(k) = u_f\left[1 - \left(\frac{k}{k_{\text{jam}}}\right)^{\psi}\right] \tag{21}$$

where ψ is a positive parameter. Other forms of the u-k relationship can also be used with this model (i.e., the model is independent of the choice of u-k relationship). The previous form of Equation 21 was selected for easy parameter calibration.

In comparison with the original high-order model, the proposed high-order model takes the relaxation process into account and treats the relaxation time as a function of the density. In addition, the relaxation time in the new model appears only at the first term on the right side of Equation 19b, making the new model more reasonable from the physical point of view and easier to be treated by sophisticated finite-difference methods.

A comparison of the proposed high-order model with the semiviscous model shows that the difference between them is that a different relaxation process is adopted by each model. When traffic becomes congested, the relaxation process adopted by the proposed high-order model does not produce the incorrect speed change that occurs in the semiviscous model.

To see a detailed difference between the proposed and the original high-order models, linearize the proposed high-order model (Equation 19) for small perturbations around the state $[k_0, u_0 = u_e(k_0)]$ when $\beta = -1$. Thus, the state $[k_0, u_0 = u_e(k_0)]$ is equilibrium if the following condition holds:

$$u_0 + \sqrt{\sigma} > c_0 > u_0 - \sqrt{\sigma} \tag{22}$$

Compared with the stable condition (Equation 4) of the original high-order model, if $\sigma = v/T$, then Equation 22 is equal to Equation 4; if $\sigma > v/T$, then the range of stability given by Equation 22 is larger than that given by Equation 4; if $\sigma < v/T$, then the range of stability given by Equation 22 is smaller than that given by Equation 4.

Now the shock structure of the proposed high-order model for the case of $\beta = -1$ will be investigated. Using the same method as for the original high-order model,

$$T(k)[\sigma - (u - U)^2]\frac{dk}{d\xi} = k[u_e(k) - u] \tag{23}$$

Thus, the smooth shock exists when the following condition holds:

$$\sigma > (u - U)^2 \tag{24}$$

In addition, comparing Equation 23 with Equation 8, it is seen that when traffic is uncongested, the shock thickness of the proposed high-order model is larger than that of the original high-order model when the value of $T(k)$ is large; when traffic is congested, the shock thickness of the proposed high-order model may be equal to that of the original high-order model because the value of $T(k)$ will approach the constant reaction time.

To implement the proposed high-order model, the upwind method is applied to the model [see details elsewhere (*28*)]. It should be noted that preliminary results show that a high-order TVD method is computationally very expensive and less accurate than the first-order upwind method used, because the former results in shock waves sharper than they really are. However, implicit methods have some merits (*15*) and should be investigated further.

TEST RESULTS

Testing with Hypothetical Data

The continuum models discussed earlier are investigated on the basis of a hypothetical case in order to find the model that can produce a reasonable description of traffic when an incident occurs downstream. For this reason, assume the hypothetical case (Case 1) described next. The freeway geometry for this case is a three-lane, 1,828-m-long pipeline section as shown in Figure 1 (*top*). The analysis period is 15 min. Traffic flow during the first 5 min is assumed to be uncongested. After the first 5 min congestion occurs at the downstream end and continues for 5 min. Then the incident is removed from the downstream end. Figure 1 (*bottom*) gives the flow patterns at the upstream and downstream boundaries. For comparing the results produced by the proposed high-order model, the simple continuum model and the original high-order model as implemented in the CORFLO (*29*) as well as the improved high-order model, the equilibrium speed-density relationship of KRONOS (*4*) was used for implementing these models. This relationship is

$$u_e(\text{km/hr}) = \begin{cases} 105 & \text{for } k \le 9 \text{ (veh/km)} \\ -\frac{1,155}{729}k + \frac{83,160}{729} + \frac{34,020}{729k} & \text{for } 9 \le k \le 36 \\ -\frac{21}{64}k + \frac{189}{8} + \frac{6,699}{4k} & \text{for } 36 \le k \le 116 \end{cases}$$

Now look at the simulation results of speed because speed is one of the variables that have the shock behavior. The simulation results of 5-min average speed at time interval [5,10] produced by the simple continuum model with the Lax method, the semiviscous model with the upwind method, and the proposed high-order model with the upwind method are shown in Figure 2. From Figure 2, it is clear that the simple continuum model, the semiviscous model, and the proposed high-order model capture the shock wave propagation. However, the proposed high-order model is more accurate in capturing the shock wave than the other two models because the shock wave produced by the proposed high-order model backward propagates the same as the theoretical value.

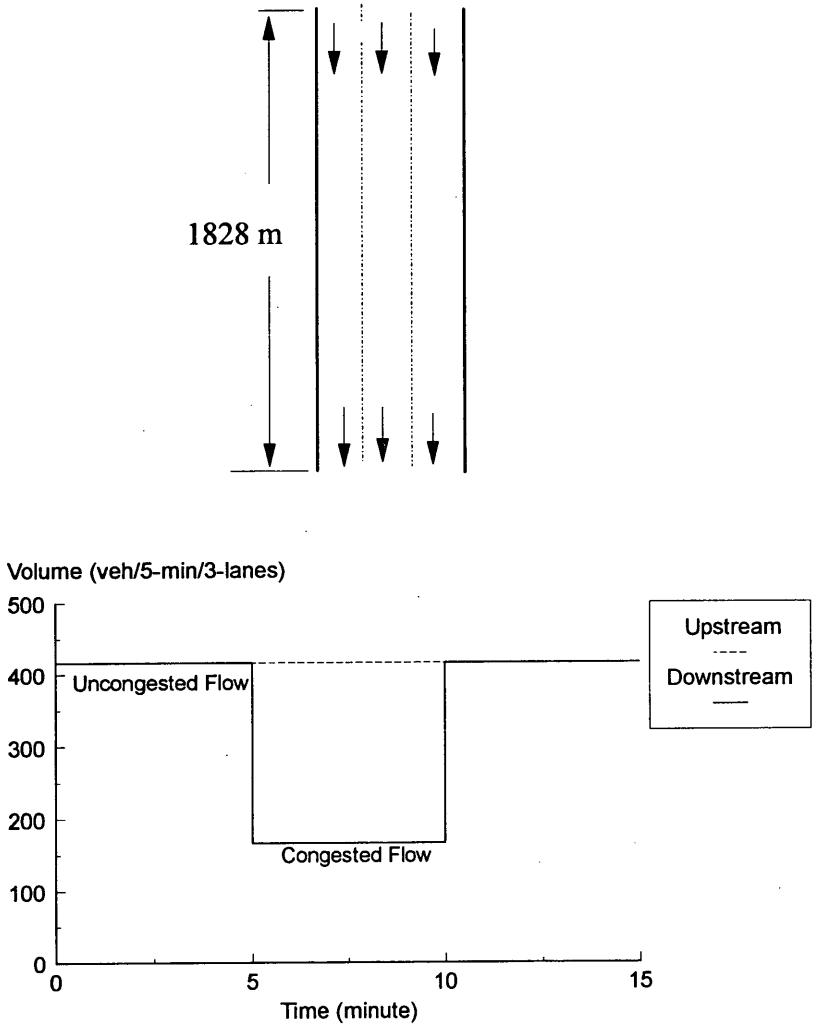


FIGURE 1 Geometry (*top*) and volume at upstream and downstream boundaries (*bottom*), Case 1.

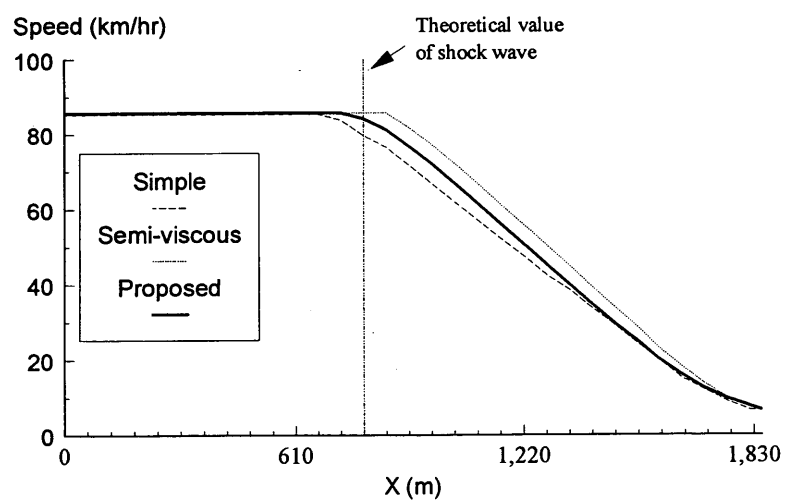


FIGURE 2 Five-minute average speed at time interval [5,10] produced by simple continuum model (Lax method), semiviscous model (upwind method), and proposed high-order model (upwind method), Case 1.

Unfortunately, when Case 1 was investigated by using the original high-order model (CORFLO), the improved high-order model, and the viscous model, these three models did not produce the correct shock wave as shown in Figure 2. This is because as shown earlier, the explicit Euler method is adopted by these three models. To demonstrate this assertion, the authors applied the upwind method to the original high-order model. The results showed that the upwind discretized form of the original high-order model did capture the shock wave propagation. Moreover, the original high-order model (CORFLO) and the improved high-order model for the Euler method appear to be very sensitive to the parameters chosen (*28*). Details of the outputs from this case as well as results from several other hypothetical cases can be found elsewhere (*28*).

Parameter Calibration

It has been shown that all the high-order continuum models include parameters. So before the models are used with field data, these parameters must be calibrated. In the past, parameters were calibrated by trial and error. Such a process is very time-consuming and requires great effort. To minimize the effort, the authors have developed a procedure that has been incorporated into their simulation program without user interface beyond the supply of field data.

This parameter calibration is an optimization problem in which the objective function is defined as follows:

$$\min f(x_1, x_2, \ldots, x_p) = \sum_{i=1}^{n} \{MSE_i(V) + MSE_i(S)\} \tag{25}$$

where

$x_j (j = 1, 2, \ldots, p)$ = parameters to be calibrated,
n = number of checking points, and

$$MSE(y) = \frac{1}{N} \sum_{i=1}^{N} [y_i^o - y_i^c] \tag{26}$$

where

y = volume or speed,
y^o = observed data,
y^c = computed result,
MSE = mean squared error, and
N = number of observations.

The optimization procedure is based on the Fletcher-Reeves conjugate method (*30*). The gradients in this method are evaluated by a finite-difference approximation in the procedure. Thus, by using this optimization procedure in parameter calibration, the minimization of the objective function, Equation 25, at least local minimization, yields an optimized set of parameters. Other more sophisticated optimization strategies (e.g., Monte Carlo methods) will be explored in the future.

Testing with Field Data

In this subsection, two test cases with field data are presented [others can be found elsewhere (*28*)]. Case 2 is based on a two-lane pipeline freeway of the Minneapolis I-35W between the 76th and 70th Streets. Traffic data used in Case 2 were the uncongested northbound traffic from 6:30 to 8:30 a.m. on November 7, 1989. The roadway geometry and arrival and departure traffic patterns for Case 2 are shown in Figure 3. The checking point is at the middle of the freeway. Case 3 is based on a four-lane pipeline freeway from I-35W close to downtown Minneapolis, starting from 26th Street and ending at 31st Street. Traffic data used by Case 3 were the congested southbound traffic from 4:00 to 6:40 p.m. on November 14, 1989. Congestion starts at 4:05 p.m. at the downstream boundary and lasts 2 hr 15 min. The geometry and arrival and departure patterns for Case 3 are shown in Figure 4. The checking point for Case 3 is also at the middle of the freeway.

To evaluate each model quantitatively, the following statistics are calculated to get the error indexes based on the deviations of simulation results from the field observations:

$$MAE = \frac{1}{N} \sum_{1}^{N} |\text{observed} - \text{computed}|$$

$$MPE = \frac{1}{N} \sum_{1}^{N} \frac{|\text{observed} - \text{computed}|}{\text{observed}}$$

$$MSE = \frac{1}{N} \sum_{1}^{N} (\text{observed} - \text{computed})^2$$

$$\text{Std. deviation} = \sqrt{\frac{1}{N-1} \sum_{1}^{N} (\text{observed} - \text{computed})^2}$$

where

MAE = mean absolute error,
MPE = mean percentage error, and
N = number of observations (i.e., the number of time intervals).

In these two cases, six models are investigated, namely, the simple continuum model, the original high-order model, the improved high-order model, the semiviscous model, the viscous model, and the proposed high-order model. $\Delta x = 61$ m and $\Delta t = 1$ sec are adopted for each model except CORFLO, in which the step size in space and time are determined internally ($\Delta x = 31$ m and $\Delta t = 1$ sec). The u-k curve wherever it applies was obtained from data collected from I-35W. CORFLO has built in three types of u-k curves to choose from; all three types have been tried and the best results are presented. Except for the simple continuum model and CORFLO, the parameters in the other four models were calibrated by using the optimization procedure mentioned earlier.

Results for the two test cases are presented in Tables 1 and 2. It can be seen that

1. When there is no downstream congestion (as in Case 2), all models including the simple continuum model performed at about the same error level except CORFLO.
2. When downstream congestion begins, different models produce different results. The simple continuum model gave good results that were better than CORFLO. Comparing the results produced by CORFLO and the improved high-order model, which are solved with the same numerical method, it is seen that the improved high-order model was more accurate than the original high-order model. It is clear that the proposed high-order model was the over-

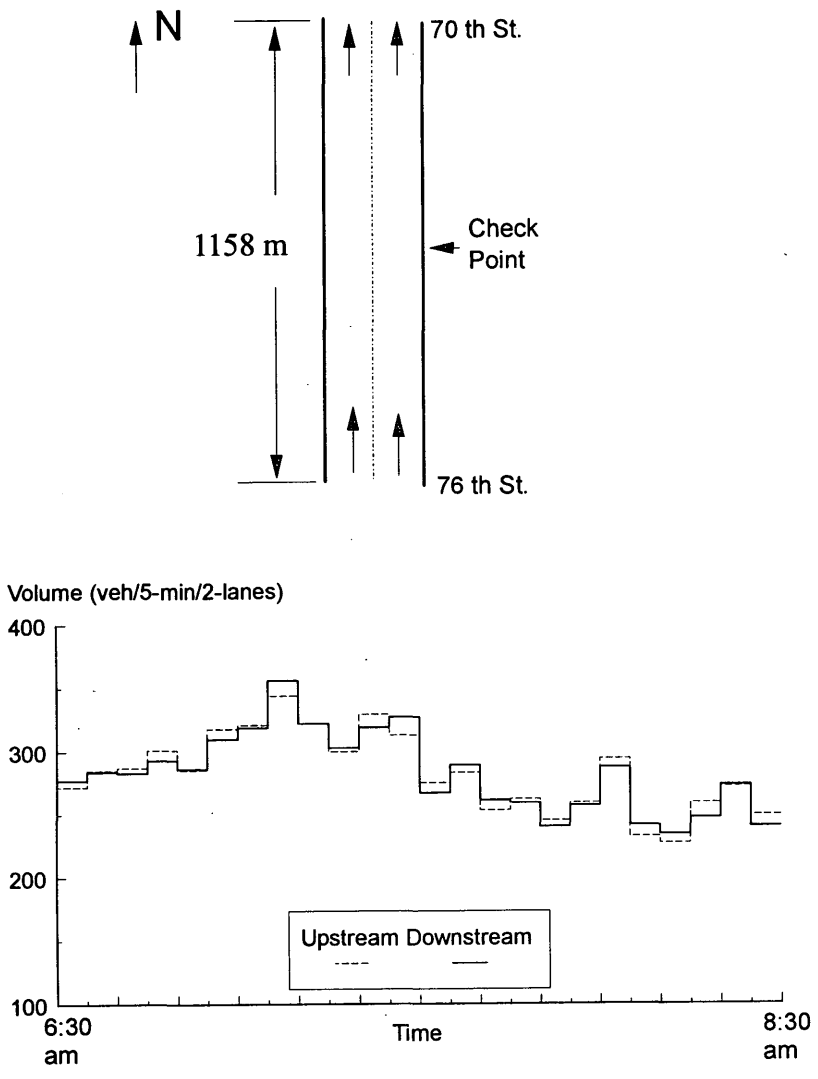


FIGURE 3 Geometry (*top*) and volume at upstream and downstream boundaries (*bottom*), Case 2.

all best in terms of accuracy, with an *MSE* of 290. The viscous model always produces a large error in *MSE* than the other high-order models (except CORFLO).

3. All the high-order models except the proposed high-order model use the different values of parameters for Cases 2 and 3 in order to get good results. This means that the proposed high-order model is the easiest to calibrate.

4. From the tested cases (*28*), the results from the simple continuum model appear to be very sensitive to the choice of the speed-density relationship. The proposed high-order model is less dependent on the choice of the speed-density relationship.

CONCLUSIONS

Five continuum models and a proposed high-order model have been reviewed. Merits and limitations of the various formulations were identified. Preliminary comparative testing of the models was also undertaken. From the hypothetical case, the simple continuum model, the semiviscous model, and the proposed high-order model properly capture the shock wave structure. The ability of CORFLO (the original high-order model), the improved high-order model, and the viscous model to capture shock waves accurately is limited.

In the authors' preliminary testing with field data, all models including the simple continuum model give reliable results. For uncongested cases tested, no apparent merit of high-order modeling versus simple continuum modeling was found. For congested cases tested most high-order models show some error reductions. For all the cases tested, the proposed high-order model produces a smaller error than the other models. Moreover, the proposed high-order model has the strong robust property of parameters for various cases. This property is very important for implementing the proposed high-order model in practice because one can use only a set of precalibrated parameters.

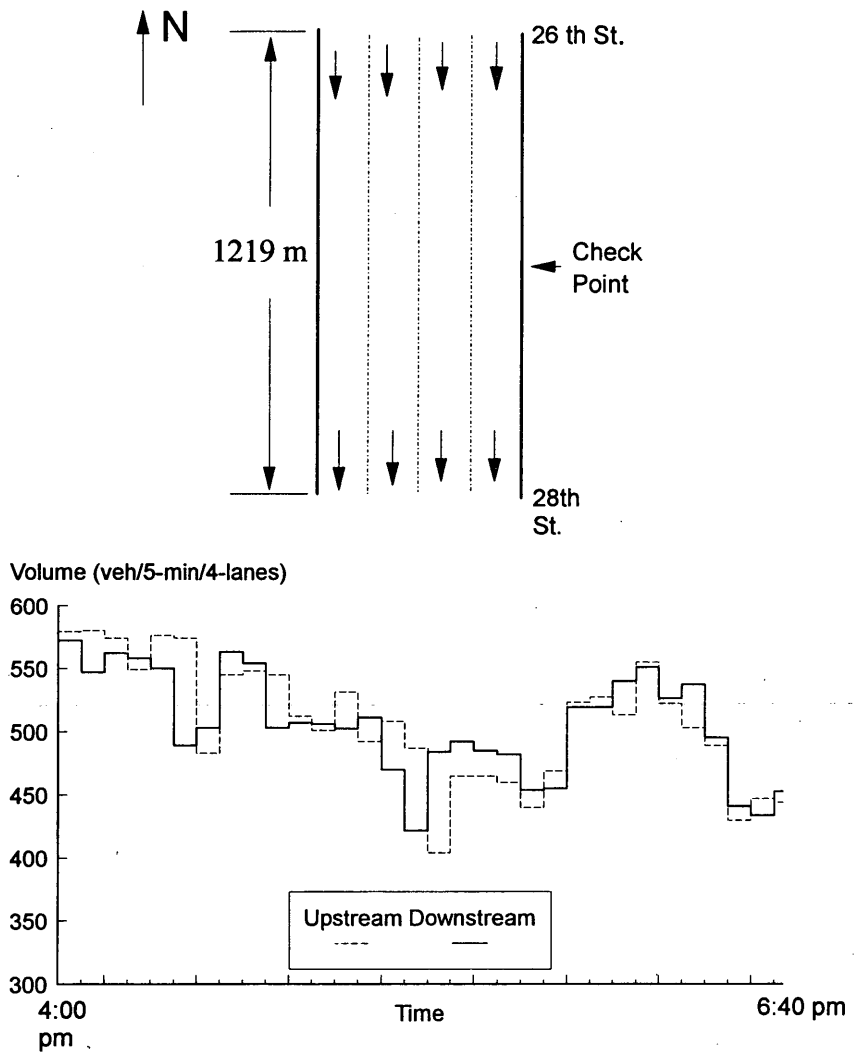


FIGURE 4 Geometry (*top*) and volume at upstream and downstream boundaries (*bottom*), Case 3.

TABLE 1 Error Indexes for Case 2

Models (method)	Simple continuum model (Lax)	Improved high-order model (Euler)	CORFLO (Euler)	Original high-order model (Upwind)	Semi-viscous model (Upwind)	Viscous model (Euler)	Proposed high-order model (Upwind)
MAE	4	4	8	4	4	4	4
MPE	2	2	2	2	2	2	2
MSE	27	27	52	27	22	27	22
Std.Dev	5	5	7	5	5	5	5

(1). MAE and Std.Dev: Veh/5-minutes;
(2). MSE: $(\text{Veh/5-minutes})^2$.

For finite-difference methods, the Euler method is not good for the numerical implementation of traffic flow models. The upwind scheme with flux vector splitting is recommended for computational accuracy and efficiency.

Finally, the simple continuum model is very sensitive to the choice of the speed-density relationship, whereas most high-order models are less sensitive or not sensitive at all.

ACKNOWLEDGMENTS

This research was sponsored by the Center for Transportation Studies of the University of Minnesota. The authors would like to thank P. Yi from the Minnesota Department of Transportation for his help during the course of this work and for going over the final manuscript.

TABLE 2 Error Indexes for Case 3

Models (method)	Simple continuum model (Lax)	Improved high-order model (Euler)	CORFLO (Euler)	Original high-order model (Upwind)	Semi-viscous model (Upwind)	Viscous model (Euler)	Proposed high-order model (Upwind)
MAE	18	15	24	17	19	15	15
MPE	4	3	5	3	4	3	3
MSE	522	416	999	370	508	511	290
Std.Dev	23	21	32	20	23	23	17

(1). MAE and Std.Dev: Veh/5-minutes;
(2). MSE: (Veh/5-minutes)^2.

REFERENCES

1. Lighthill, M. H., and G. B. Whitham. On Kinematic Waves II: A Theory of Traffic Flow on Long Crowded Roads. *Proc., Royal Society of London Series A*, Vol. 229, 1955, pp. 317–345.
2. Michalopoulos, P. G., and J. Lin. A Freeway Simulation Program for Microcomputers. *Proc., 1st National Conference on Microcomputers in Urban Transportation*, ASCE, California, 1985, pp. 330–341.
3. Michalopoulos, P. G., E. Kwon, and J.-G. Kang. Enhancement and Field Testing of a Dynamic Freeway Simulation Program. In *Transportation Research Record 1320*, TRB, National Research Council, Washington, D.C., 1991, pp. 203–215.
4. Michalopoulos, P. G., E. Kwon, C.-F. Lee, G. Mahadevan, and J.-G. Kang. *Development of an Integrated Simulation Package for Freeway Design, Operations and Adaptive Traffic Management (Phase 1: Enhancement of the KRONOS Simulation Program)*. Final Report. Center for Transportation Studies, University of Minnesota, Minneapolis, 1992.
5. Payne, H. J. Models of Freeway Traffic and Control. *Proc., Simulation Council, Mathematical Models of Public Systems*, Vol. 1, No. 1, 1971, pp. 51–61.
6. Payne, H. J. FREFLO: A Macroscopic Simulation Model of Freeway Traffic. In *Transportation Research Record 722*, TRB, National Research Council, Washington, D.C., 1979, pp. 68–77.
7. Papageorgiou, M. A Hierarchical Control System for Freeway Traffic. *Transportation Research*, Vol. 17B, No. 3, 1983, pp. 251–261.
8. Papageorgiou, M., J. M. Blosseville, and H. Hadj-Salem. Macroscopic Modelling of Traffic Flow on the Boulevard Peripherique in Paris. *Transportation Research*, Vol. 23B, No. 1, 1989, pp. 29–47.
9. Michalopoulos, P. G., P. Yi, D. E. Beskos, and A. S. Lyrintzis. Continuum Modeling of Traffic Dynamics. *Proc., 2nd International Conference on Applications of Advanced Technology in Transportation Engineering*, Minneapolis, Minn., Aug. 1991, pp. 36–40.
10. Michalopoulos, P. G., P. Yi, and A. S. Lyrintzis. Development of an Improved High Order Continuum Traffic Flow Model. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1993, pp. 125–132.
11. Phillips, W. F. *A New Continuum Model for Traffic Flow*. Report DOT-RC-82018. Utah State University, Logan, 1979.
12. Kühne, R. Macroscopic Freeway Model for Dense Traffic: Stop-Start Waves and Incident Detection. *Proc., 9th International Symposium on Transportation and Traffic Theory*, VNU Science Press, 1984, pp. 21–42.
13. Ross, P. Traffic Dynamics. *Transportation Research*, Vol. 22B, No. 4, 1988, pp. 421–435.
14. Chronopoulos, A., P. G. Michalopoulos, and J. Donohoe. Efficient Traffic Flow Simulation Computations. *Mathematical Computation and Modelling*, Vol. 16, No. 5, May 1992, pp. 107–120.
15. Chronopoulos, A., A. S. Lyrintzis, P. G. Michalopoulos, C. Rhee, and P. Yi. Traffic Flow Simulation Through High-Order Traffic Modelling. *Mathematical Computation and Modelling*, Vol. 17, No. 8, Aug. 1993, pp. 11–22.
16. Leo, C. J., and R. L. Pretty. Numerical Simulation of Macroscopic Continuum Traffic Models. *Transportation Research*, Vol. 26B, No. 3, 1992, pp. 207–220.
17. Lyrintzis, A. S., P. G. Michalopoulos, A. Chronopoulos, P. Yi, G. Liu, and C. Rhee. *Development of Advanced Traffic Flow Models and Implementation in Parallel Processing (Phase I)*. Final Report. Center for Transportation Studies, University of Minnesota, Minneapolis, 1992.
18. Lax, P. D. Weak Solution of Nonlinear Hyperbolic Equations and Their Numerical Computations. *Communications on Pure Applied Mathematics*, Vol. 7, 1954, pp. 159–173.
19. Hirsch, C. *Numerical Computation of Internal and External Flows. Vol. 2: Computational Methods for Inviscid and Viscous Flows*. John Wiley & Sons, Chichester, England, 1990.
20. Steger, J. L., and R. F. Warming. Flux Vector Splitting of the Inviscid Gas-Dynamic EquationMSEs with Applications to Finite Difference Methods. *Journal of Computational Physics*, Vol. 40, 1981, pp. 263–293.
21. Whitham, G. B. *Linear and Nonlinear Waves*. John Wiley & Sons, New York, 1974.
22. Hirsch, C. *Numerical Computation of Internal and External Flows. Vol. 1: Fundamentals of Numerical Discretization*. John Wiley & Sons, Chichester, England, 1988.
23. Hirt, C. W. Heuristic Stability Theory for Finite-Difference Equations. *Journal of Computational Physics*, Vol. 2, 1968, pp. 339–355.
24. Gerlough, D. L., and M. J. Huber. *Special Report 165: Traffic Flow Theory*. TRB, National Research Council, Washington, D.C., 1975.
25. Becker, E. *Gas Dynamics*. Academic Press, New York, 1968.
26. Greenberg. H. An Analysis of Traffic Flow. *Operations Research*, Vol. 7, No. 1, 1959, pp. 79–85.
27. Talbot, L., and S. M. Scala. Shock Wave Structure in a Relaxing Diatomic Gas. *Advanced Applied Mechanics*, Supplement 1 (H. L. Dryden and T. Von Karman, eds.), Academic, New York, 1961, pp. 603–622.
28. Lyrintzis, A. S., P. G. Michalopoulos, G. Liu, and R. P. Rangiah. *Development of Advanced Traffic Flow Models and Implementation in Parallel Processing (Phase II)*. Final Report. Center for Transportation Studies, University of Minnesota, Minneapolis, 1994.
29. *TRAF User Reference Guide*. FHWA, U.S. Department of Transportation, 1992.
30. Luenberger, D. G. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, New York, 1973.

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

Effect of Adverse Weather Conditions on Speed-Flow-Occupancy Relationships

AMAL T. IBRAHIM AND FRED L. HALL

The effect of adverse weather conditions on the flow-occupancy and speed-flow relationships is studied. The data used in the analysis were obtained from the Queen Elizabeth Way Mississauga freeway traffic management system. Regression analyses were performed to select proper models representing the flow-occupancy and speed-flow relationship for uncongested operation. Then dummy variable multiple regression analysis techniques were used to test for significant differences in traffic operations between different weather conditions. It is concluded that adverse weather conditions reduce the slope of flow-occupancy function and cause a downward shift in the speed-flow function. Adverse weather conditions also reduce the maximum observed flow rates.

This paper addresses the effect of adverse weather on freeway operations, a topic that is of interest for several reasons. Most of the data and analyses presented in standard reference works such as the *Highway Capacity Manual* (HCM) (*1*) deal solely with "correcting" for other departures from ideal conditions. No corrections are provided for weather effects, in part perhaps because little is known in detail. The introduction of intelligent vehicle-highway systems (IVHS) will require much more detailed knowledge of the operational characteristics of freeways under various conditions. If adaptive control of freeways is to become a reality, one of the factors that must be adapted to is weather. Intuitively, snow or heavy rain decreases speeds and perhaps volumes. The topic of this paper is not simply whether that intuition is correct, but what quantities can be put on those intuitive expectations.

BACKGROUND

Three main issues are addressed in the literature review. The first is the findings of previous studies on the effect of adverse weather conditions on speed-flow-occupancy relationships. The second is the nature of the functions for those relationships, to focus the analysis. The third issue concerns using a dummy variable multiple regression analysis technique that provides a means of testing for significant differences between data sets.

A computerized bibliographic search was conducted through the Transportation Research Information Service records and the Engineering Information Index to find related previous work. The search showed that six items dealt with the effect of weather on roadway traffic operations. One of the six, by Hall and Barrow (*2*), discussed the effect of weather on the relationship between flow and occupancy; the other five (*3–7*) focused on the relationship between adverse weather and road safety. Two results of those studies were helpful for this analysis. Andrey and Yagar found that collision risk returns to normal immediately after rain stops (*7*). Hence, it appears important to confine the analysis to the occurrence of adverse weather in order to indicate clearly the impact of weather on traffic operations. Salonen and Puttonen found that darkness reduces the operating speed by 5 km/hr (*3*). Thus, the weather effect should be studied keeping day- and nighttime traffic data separate.

In addition to those found through the computerized bibliographic search, there are three more references mentioned in Chapter 6 of the HCM. Jones and Goolsby reported a reduction in capacity of 14 percent during rain, but the severity of the rain is not noted (*8,9*). Kleitsch and Cleveland reported an average reduction of 8 percent but emphasized the variation in the reduction that was associated with rainfall intensity (*10*).

The literature can also be helpful on a second topic, the shape of the function to use in the analysis. Several researchers (*11–14*) have reported on the occurrence of gaps or discontinuities in freeway speed-density and flow-density data. They suggested that discontinuous functions are necessary to properly describe the observed traffic behavior. Given the recent changes in understanding of the shape of the speed-flow curve (*15*), rather than use functional forms as specified in the earlier literature, regression analysis will be used to identify forms from the data. To simplify the analysis, only the uncongested operations will be analyzed.

The dummy variable multiple regression analysis technique used by Hall and Barrow (*2*) will be used for the current study. That technique used a dummy variable with values of 1 and 0 to distinguish between two data sets. For instance, if a linear function is used to represent the flow-occupancy relationship (for the uncongested data), the general equation used for the dummy variable regression analysis will be of the form

$$\text{Flow} = a + b * \text{occupancy} + c * \text{dummy} + e * \text{dummy} * \text{occupancy}$$

If the coefficient of the dummy variable is significantly different from 0, there is a significant difference for the value of the intercept between the two data sets by an amount equal to the estimated coefficient c. If the coefficient on the product of the dummy and occupancy is significant, this means there is a difference in the slope with a value equal to the coefficient e. When both coefficients involving the dummy variable are significant, both slope and intercept for the two functions are different. Under those conditions, for the data set that has a dummy value equal to 0,

$$\text{Flow} = a + b * \text{occupancy}$$

For the data set that has a dummy value equal to 1,

$$\text{Flow} = (a + c) + (b + e) * \text{occupancy}$$

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

DATA

There are two issues to address with respect to the data for this study: first, where they come from and how the sites and times were selected for analysis; second, how the weather information was obtained, how well it represents actual conditions at the site, and how it affects the sample selection.

The traffic data available for this study were from the freeway traffic management system (FTMS) for the Queen Elizabeth Way (QEW) in Mississauga, Ontario. The data are recorded 24 hr/day at 30-sec intervals. Three variables are available: volume, occupancy, and speed.

Three criteria were chosen for site selection. The first was the existence of double loop detectors, which provided measured speeds. The second was that the study site should not be influenced by ramp or weaving sections. The third was the data quality. Only Stations 14 and 21 met all criteria; both are used for the analysis.

The study was performed for the median lane and for the average data across three lanes. The median lane has the highest flow rates of the three lanes and includes passenger cars only. Any differences between that and the three-lane average data could indicate the effect of adverse weather on the behavior of trucks and slow vehicles.

The comparison between weather conditions was limited to the same time of day under each condition for three reasons. First, driver behavior differed from daytime to dark for the same weather condition, as found by Salonen and Puttonen (*3*). Second, regression equations representing the flow-occupancy relationship for different periods proved to be significantly different. Third, data sets from various times of day included a different range of occupancy and flow, which itself made any cross-period comparison difficult.

Detailed weather records for Pearson International Airport were obtained from the Atmospheric Environment Service in Downsview, Ontario, and were compared with the operators' log book at the FTMS center in Mississauga to ensure that the records for the airport accurately reflected weather conditions at the QEW freeway in Mississauga. The information available in the operators' log book agreed with the airport weather records.

Three factors were considered in selecting days to use in the study. The first was to include days with different types and intensities of weather conditions: light rain, heavy rain, light snow, snow storms, and clear. The visibility criterion was used to identify the intensity of snow, and the rate of fall was used to identify the intensity of rain. To increase the likelihood of adverse weather conditions, the months of October, November, and December 1990 and January and February 1991 were considered. Clear days were taken from the same months.

The second factor considered in choosing the days was day of the week. It was hoped to exclude Saturdays, Sundays, and holidays because of possible changes in travel patterns, but Saturday, December 29, 1990, was used for rainy weather conditions because it had 6 hr of rainfall during the relevant time of day, and there was a lack of good data for rainy conditions during other days.

As stated, the comparison was limited to the same period in all days. To enable a focus on uncongested data, the period from 10:00 a.m. to 4:00 p.m. was chosen. If adverse weather did not last for the whole 6 hr, observations before or after adverse weather were deleted from the data file. The result is that the days used for the analysis constitute not a probability-based sample but all of the days with consistent adverse weather in the period investigated, together with an arbitrarily chosen representative set of days with clear weather (Table 1).

TABLE 1 Selected Days for Different Weather Conditions

Date (yymmdd)	Weather cond.	Duration of weather condition	Number of Good Data Points			
			Station 14		Station 21	
			median lane	Avg over 3 lane	median lane	Avg over 3 lane
901002	clear	10:00 - 16:00	297	634	706	720
901015	clear	10:00 - 16:00	720	604	702	720
901115	clear	10:00 - 16:00	601	567	593	601
901116	clear	10:00 - 16:00	697	695	694	690
901205	clear	10:00 - 16:00	697	720	719	720
901210	clear	10:00 - 16:00	694	720	720	720
901105	rain	10:00 - 16:00	677	695	658	663
901122	rain	10:00 - 16:00	674	588	710	712
901221	rain	10:00 - 16:00	720	713	707	711
901229	rain	10:00 - 16:00	720	688	720	722
901012	rain	11:49 - 16:00	495	388	465	494
910214	snow	10:00 - 16:00	713	720	719	720
910215	snow	10:00 - 15:13	556	553	521	528
910108	snow	10:00 - 14:00	333	332	308	403
910111	snow	10:00 - 16:00	661	571	647	618
901204	snow	10:00 - 16:00	720	720	693	696
901203	snow	10:00 - 14:30	414	390	423	400

SELECTION OF MODELS AND PRELIMINARY ANALYSIS

Regression analyses were conducted on the clear weather data to select models for the uncongested part of the flow-occupancy and speed-flow relationships. From a visual inspection of a plot of 30-sec flow-occupancy data, two functional forms appeared plausible: a linear function or a quadratic function. For the linear model, flow = $a + b *$ occupancy, the results are as follows:

St. 14 (median): flow = $1.4 + 1.14 *$ occ $\quad R^2 = .9592$
St. 14 (average): flow = $2.2 + 0.85 *$ occ $\quad R^2 = .8434$
St. 21 (median): flow = $4.8 + 0.75 *$ occ $\quad R^2 = .6165$
St. 21 (average): flow = $2.5 + 1.06 *$ occ $\quad R^2 = .8581$

The regression analysis showed significant values for the intercept and slope at the 5 percent level and respectable values of R^2, indicating a good fit of the model to the data. However, three of the intercept values are large, which is meaningless in practical terms. It is possible to have a value of 1 or 2 veh/30 sec at zero occupancy since the occupancies are truncated, but it is not possible to have a value of 4.8, and even 2.5 is unlikely.

For the second model, flow = $a + b *$ occupancy + $c *$ occupancy^2, the results are as follows:

St. 14 (median): flow = $0.8 + 1.29 *$ occ $- 0.009 *$ occ^2 $\quad R^2 = .9607$
St. 14 (average): flow = $1.3 + 1.08 *$ occ $- 0.013 *$ occ^2 $\quad R^2 = .8461$
St. 21 (median): flow = $1.2 + 1.42 *$ occ $- 0.034 *$ occ^2 $\quad R^2 = .6367$
St. 21 (average): flow = $1.2 + 1.39 *$ occ $- 0.020 *$ occ^2 $\quad R^2 = .8612$

All the estimated parameters were again significant at the 5 percent level, and the value of R^2 is in all four cases slightly higher than for the linear model. The fact that the quadratic term is significant led to the choice of the quadratic model over the linear model. As well, with the quadratic term, the value of the intercept dropped to 1 veh/30 sec.

For speed-flow data, the 30-sec observations showed high scatter, which made it difficult to predict a good model for this relationship. Nevertheless, three functions were tested to fit the data. The linear model showed significant values (at the 5 percent level) for the intercept and coefficient but a very low R^2, as follows:

St. 14 (median): speed = $114 - 0.36 *$ flow $\quad R^2 = .0679$
St. 14 (average): speed = $104 - 0.42 *$ flow $\quad R^2 = .0308$
St. 21 (median): speed = $100 - 0.48 *$ flow $\quad R^2 = .0623$
St. 21 (average): speed = $100 - 0.47 *$ flow $\quad R^2 = .0549$

The low values of R^2 can be attributed to the high scatter in the data. The fact that the R^2 is so low suggests that caution should be used when interpreting the coefficient on flow, even given that it is statistically significant. The flow used in the regressions is the actual 30-sec volume, which ranges from 0 to 25 vehicles. Hence, a coefficient of 0.4 would imply a speed drop of 10 km/hr over the full range. Because 25 veh/30 sec would translate to a flow rate of 3,000 veh/hr, it can be seen that there is very little speed drop over the range of flows. (Of course, the flow of 25 veh/30 sec is not sustained even for two consecutive 30-sec intervals, much less for a full 1 hr.)

A quadratic function was also estimated. It did not improve the R^2; it reduced the significance of the linear coefficient for the median lane data; and it did not result in significant coefficients for the three-lane average data. Therefore, the linear model was chosen over the quadratic model.

In addition, analysis was conducted to test the goodness of fit of a piecewise linear model. The break-point between the two segments of the model was identified by using the multiple regression technique with one dummy variable. The equation used was of the form

Speed = $A + B *$ flow + $C *$ dummy + $E *$ dummy $*$ flow

Five values of the breakpoint, from 12 to 18 veh/30 sec (1,440 to 2,160 vph), were tested. The R^2 values were all small (from .0680 to .0691). Compared with the value of R^2 for the linear model (.0679), there was not much gained by using a piecewise model. Therefore, in order to simplify the comparison between weather conditions, the simple linear model was used for the speed-flow relationship.

COMPARISON STUDY OF DIFFERENT WEATHER CONDITIONS

Three comparison tests were performed to study the effects of rainy and snowy weather on the underlying relationships: first, clear and rainy weather were compared to test the effect of rainy weather and investigate whether differences within the rainy weather were more important than differences between the clear and rainy weather; then clear and snowy weather conditions were compared in a similar manner, as were rainy and snowy conditions. All three analyses were conducted using the 30-sec data, which incorporates the greatest amount of variation. To test whether the results depended on the level of temporal aggregation, tests were also conducted with 5-min aggregated data. Finally, the effect of adverse weather on maximum observed flows is observed.

Variation Within Each Weather Condition

An important issue was whether to treat all days of the same weather condition as one data set. This issue was decided with a two-step analysis. First, a regression analysis was done for each day separately, and then the underlying functions for all days of each weather condition were plotted on the same graph to identify the upper and lower functions for each condition. Second, multiple regression analyses were conducted to test the differences between the highest and lowest functions for each weather condition.

The highest and lowest functions for clear days at both stations were found to be statistically different. For the flow-occupancy function, over the four data sets the dummy variable itself never entered, the dummy on slope was significant (at the 5 percent level) in two of the four analyses, and the dummy on curvature was significant three times. For the speed-flow relationship, the dummy entered three of four times (with the largest magnitude being 3 km/hr), as did the dummy on the slope. Hence there are differences between the upper and lower functions for clear days, but the nature of the differences is not the same across the four data sets.

Nevertheless, these results may reflect normal variation within a range of behavior that can be represented by an average equation for

all clear days. Hence, a test was conducted for the upper day, December 10, 1990, against the data of the other five clear days. There was a statistically significant difference in the curvature of the flow-occupancy function for the four data sets, but it was small in practical terms. For instance, the largest value of the difference in the quadratic term was found to be –0.005 (for median lane data at Station 14). For occupancy equal to 20 (the highest observed value), the difference in the flow would be 2 veh/30 sec.

For the speed-flow relationship, differences were absent in two data sets and minimal in the third. A noticeable difference was found for the slope and intercept only for median lane data at Station 14.

Although these analyses showed that there are some differences within the clear days, the differences are not consistent across the four data sets, nor are the differences for the most part of practical significance. Therefore, it was decided to consider the data for the 6 days as one data file representing the clear weather, which would have the benefit of retaining considerable variation within this part of the data.

Testing the differences within the rainy and snowy weather showed that there were significant differences in the intercept, slope, and curvature of the flow-occupancy relationship and in the intercept for the speed-flow relationship. The magnitude of these differences between the upper and lower functions for snowy days was much bigger than for the rainy days, especially for the slope of the flow-occupancy relationship and the intercept of the speed-flow function. Thus, it was decided not to treat the rainy or snowy days as one data file but to consider the highest and lowest functions in the comparison of weather conditions. The highest function was termed the "light" condition (rain or snow) and the lowest was termed the "heavy" condition. This means that although 5 and 6 days of data were selected for these two conditions, only December 21 and November 5, 1990, for the rainy days and December 4 and December 3, 1990, for the snowy days will be used in the ensuing analysis.

Comparison Between Clear and Rainy Weather

The comparison analysis used two dummy variables. The first tested the difference between clear and light rain (dummy1 = 0 for clear, 1 otherwise), and the second tested the difference between light and heavy rain (dummy2 = 1 for heavy rain, 0 otherwise). The results (Table 2) showed that the difference in slope of the flow-occupancy function within the rainy condition (ranging from –0.12 to –0.16) was more important than the difference between slopes for the clear and light rain (which vary from 0.07 to –0.04). The difference in the intercept of the speed-flow function within the rainy weather (ranging from –3 to –9 km/hr) was more important than the difference between the clear and rainy weather (which was only –1 or –2 km/hr).

Heavy rain caused a decrease in the slope of the flow-occupancy function (with a maximum value of –0.16 for the three-lane average data at Station 14). For the speed-flow function, there was a drop in the free-flow speed (with a maximum value of 10 km/hr for the median lane data at Station 14) and a change in the slope (with a maximum value of –0.56 for the three-lane average data at Station 14). Interestingly, the slope does not vary between light and heavy rain; only the intercept (free-flow speed) does.

Comparison Between Clear and Snowy Weather

A similar analysis was performed for snowy weather. The difference within the snowy weather was again greater than that between clear weather and light precipitation, as indicated in Table 3. Heavy

TABLE 2 Testing Difference Between Clear and Rainy Weather

Variable	Station 14		Station 21	
	Median lane	Average over 3 lanes	Median lane	Average over 3 lanes
The Flow-Occupancy Relationship				
Intercept	0.8	1.1	1.9	1.5
Occupancy	1.29	1.13	1.39	1.40
Occupancy^2	- 0.009	- 0.015	- 0.032	- 0.021
Dummy1				
Dummy2				- 0.7
Dummy1*Occupancy	-0.04	0.02		0.07
Dummy2*Occupancy	-0.12	- 0.14	- 0.16	
Dummy1*Occupancy^2				- 0.006
Dummy2*Occupancy^2				- 0.1
The Speed-Flow Relationship				
Intercept	114	105	101	102
Flow	- 0.36	- 0.49	- 0.53	- 0.64
Dummy1	- 1	- 2	- 2	
Dummy2	- 9	- 3	- 5	- 7
Dummy1*Flow	- 0.20	- 0.56	- 0.22	- 0.33
Dummy2*Flow				

Dummy1=0 for Clear
Dummy2=0 for Clear and Light Rain
Dummy1=1 for Light and heavy Rain
Dummy2=1 for Heavy Rain

TABLE 3 Testing Difference Between Clear and Snowy Weather

Variable	Station 14		Station 21	
	Median lane	Average over 3 lanes	Median lane	Average over 3 lanes
The Flow-Occupancy Relationship				
Intercept	0.8	1.3	1.8	1.3
Occupancy	1.29	1.08	1.42	1.39
Occupancy^2	- 0.009	- 0.012	- 0.034	- 0.020
Dummy1			- 0.3	
Dummy2	- 0.6	- 1.0	- 0.8	0.8
Dummy1*Occupancy				0.06
Dummy1*Occupancy^2	- 0.003	- 0.003	- 0.004	
Dummy2*Occupancy	- 0.46	- 0.36	- 0.51	- 0.47
Dummy2*Occupancy^2	- 0.007		0.017	
The Speed-Flow Relationship				
Intercept	114	105	101	102
Flow	- 0.37	- 0.43	- 0.54	- 0.64
Dummy1			- 3	- 1
Dummy2	- 50	- 41	-35	- 37
Dummy1*Flow	- 0.23	- 0.19		- 0.20
Dummy2*Flow		- 0.42		

Dummy1=0 for Clear Dummy1=1 for Light and heavy Snow
Dummy2=0 for Clear and Light Snow Dummy2=1 for Heavy Snow

snow affected traffic operation dramatically; it reduced the slope of the flow-occupancy function by a maximum value of 0.53 (three-lane average data at Station 21). It also caused a drop in free-flow speed between 35 and 50 km/hr and changed the slope of the speed-flow function three times out of four with a maximum value of –0.61 (three-lane average data at Station 14).

Comparison Between Snowy and Rainy Conditions

The differences between snowy and rainy weather were tested by using three dummy variables. The first distinguished between the snowy and rainy weather (dummy1 = 0 for light and heavy snow, 1 otherwise). The second involved differences within the snowy weather (Dummy2 = 1 for heavy snow, 0 otherwise), and the third tested differences within the rainy weather (dummy3 = 1 for heavy rain, 0 otherwise).

There was no significant difference between light rain and light snow, but there were great differences between heavy rain and heavy snow (Table 4). For the flow-occupancy function, the magnitude of difference in slope between the light snow and heavy snow (–0.42) was greater than the difference between the light rain and heavy rain (–0.12). For the speed-flow relationship there was a drop in the intercept values between and within each of the two adverse weather conditions. The greatest drop was within the snowy weather (–50 km/hr for median lane data at Station 14); next was that within the rainy weather (–9 km/hr); and the smallest drop was between the light snow and light rain (–1 km/hr).

In conclusion, heavy snow had a greater effect on traffic operations than heavy rain, whereas light rain and light snow have nearly the same effect on traffic operations. There are significant differences in the effects of the two adverse weather conditions depending on the degree of severity of each. Rainfall may affect traffic more than snow and vice versa, depending on the rate of fall, pavement wetness, and visibility.

Five-Minute Data Analysis

After the comparison study based on 30-sec data, median lane data at Station 14 were aggregated to 5-min intervals and similar analyses were repeated to test whether the findings depended on the level of temporal aggregation. Differences that were not significant when using 30-sec data may become significant when using 5-min data because of the lesser variability, since aggregation reduces the scatter of the data.

For the flow-occupancy function the quadratic term was not significant at the 5 percent level. Hence, the linear model appears to be appropriate for the 5-min uncongested data. For the speed-flow relationship, the linear model is still appropriate.

Testing differences within the same weather condition matched to a great extent the previous results based on the 30-sec data files except that there was a significant difference in the intercept of the flow-occupancy function within clear weather; it was practically minimal, however. Within the rainy condition the significant difference in the slope of the flow-occupancy function at the 30-sec data was not found for the 5-min data.

The results of the comparison study between different weather conditions using 5-min data matched the results of the analyses using the 30-sec data two-thirds of the time. For instance, differences between clear and rainy weather and between rainy and snowy weather matched the results for the 30-sec data.

Some results of the comparison between clear and snowy weather were different from those obtained for the 30-sec data. There was a

TABLE 4 Testing Difference Between Snowy and Rainy Weather

Variable	Station 14		Station 21	
	Median lane	Average over 3 lane	Median lane	Average over 3 lane
The Flow-Occupancy Relationship				
Intercept	0.9	0.4	2.0	0.6
Occupancy	1.25	1.25	1.29	1.51
Occupancy^2	- 0.009	- 0.023	- 0.031	- 0.033
Dummy1				
Dummy2	- 0.6		- 1.3	
Dummy3				
Dummy1*Occupancy			0.08	0.08
Dummy2*Occupancy	- 0.42	- 0.47	- 0.39	- 0.72
Dummy3*Occupancy	- 0.12	- 0.08	- 0.16	- 0.17
Dummy1*Occupancy^2		0.004		
Dummy2*Occupancy^2	- 0.010		0.010	0.018
Dummy3*Occupancy^2		- 0.005		
The Speed-Flow Relationship				
Intercept	114	106	98	102
Flow	- 0.58	- 0.76	- 0.56	- 0.93
Dummy1	- 1		- 2	
Dummy2	- 50	- 42	- 35	- 37
Dummy3	- 9	- 6		
Dummy1*Flow				
Dummy2*Flow				
Dummy3*Flow		- 0.28	- 0.44	- 6.76

Dummy1=0 for Light and Heavy Snow
Dummy2=0 for Light Snow, Light and Heavy Rain
Dummy3=0 for light and Heavy Snow and Light Rain
Dummy1=1 for Light and Heavy Rain
Dummy2=1 for Heavy Snow
Dummy3=1 for Heavy Rain

significant difference in the intercept between the clear and light snow for the flow-occupancy relationship, and there was a significant difference in the intercept between the clear and light snow for the speed-flow relationship.

The magnitude of the effect of adverse weather using the aggregated 5-min data was almost the same as that for the 30-sec data. For instance, heavy rain caused a reduction in the free-flow speed of 10 km/hr compared with 11 km/hr for the 30-sec data; heavy snow caused a reduction in free-flow speed of 60 km/hr compared with 50 km/hr for the 30-sec data.

Effect of Adverse Weather on Maximum Flows

The effect of adverse weather on capacity is also important. However, the data selected for the previous comparisons did not include operations at capacity: Station 14 lies upstream of three major bottleneck sections on the QEW freeway, and Station 21 lies just upstream of the major bottleneck section and thus is operating most of the time within a queue. Consequently, data on capacity operations were not available at those locations, and the effect of adverse weather on capacity cannot be investigated from the data for these two stations. However, some indication of the effect of adverse weather on flows can perhaps be obtained at these stations, simply by looking at the highest flow rates observed for clear days and for the worst days of rain and snow conditions.

Additional data, for the period from 6:00 to 10:00 a.m. to include data for the morning peak period (usually between 6:30 and 9:30 a.m.), were collected for these days to cover the highest flow rates at each site. The weather records were reviewed to ensure that only data during adverse weather were included.

The speed-flow median lane data at Station 14 during clear, heavy snow, and heavy rain conditions are shown in Figure 1. There is a

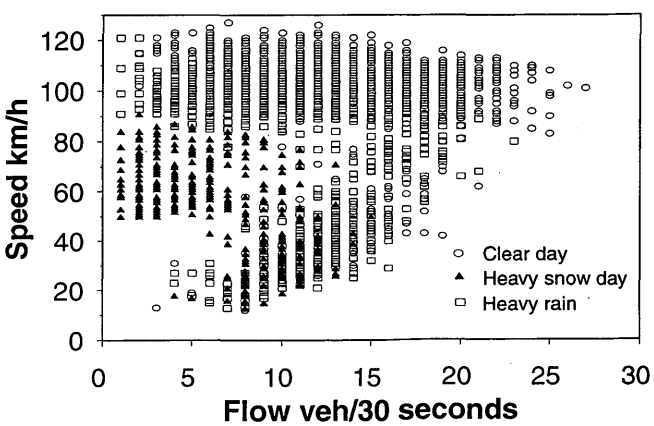


FIGURE 1 Speed-flow data for clear, heavy snow, and heavy rain at Station 14, median lane.

TABLE 5 Maximum Observed Flows for Clear, Heavy Rain, and Heavy Snow Conditions

	Station 14		Station 21	
	Median lane	Avg. over 3 lanes	Median lane	Avg. over 3 lanes
Clear Day				
Flow (vph)	3000	2160	2400	2400
Speed (km/h)	100	90	80	85
Occupancy (%)	25	19	20	18
Heavy Rain				
Flow (vph)	2400	1920	2160	2040
Speed (km/h)	90	75	70	70
Occupancy (%)	20	20	20	20
Heavy Snow				
Flow (vph)	1560	1200	1560	1680
Speed (km/h)	55	40	40	80
Occupancy (%)	30	30	27	13

Speed and occupancy are the averages at maximum flows.
The flows are hourly rates based on 30-second data.

drop in the maximum observed flows, with the drop increasing as the weather worsens. Table 5 presents a summary of the maximum attained values of flow (observed at least five times) and the average of observed occupancies and speeds at those maximum flows at Stations 14 and 21 for both median and three-lane average data.

At Station 14, heavy snow caused a reduction in the maximum flow rate of about 48 percent below the value for clear weather. Heavy rain caused a reduction in the maximum flow rate of about 20 percent. These reductions are slightly lower at Station 21 and for three-lane average data. It should be noted that these results are based on only a few hours of data on a few days at two locations on the freeway.

CONCLUSIONS

The comparison study, based on 68 data files of three weather conditions, confirmed that adverse weather conditions affect both the flow-occupancy and speed-flow relationships: the more severe the weather condition, the greater the effect on traffic operations.

Light rain and light snow caused minimal effects on both relationships. For instance, for the speed-flow function (Figure 2), light rain caused a drop in the free-flow speed of a maximum of only 2 km/hr (Table 2, Dummy 1) and an increase in slope varying from –0.20 to –0.56 km/hr/veh/30 sec. At a flow of 20 veh/30 sec (2,400 veh/hr), the combined effect of these two terms would be a drop in speed of about 13 km/hr relative to speeds during clear, dry weather. Light snow caused a drop in the free-flow speed of a maximum of 3 km/hr and a change in slope varying from –0.19 to –0.23. At 20 veh/30 sec, these coefficients imply a drop in speed of about 8 km/hr relative to dry weather. These changes due to light precipitation are statistically significant, but because of the high scatter of data, they may not be of practical importance.

Heavy precipitation, on the other hand, made a noticeable difference in both functions, with heavy snow having a much greater effect than that of heavy rain. The effect can be understood more easily with the speed-flow function results (Figure 3): heavy rain caused a drop in free-flow speeds varying from 5 to 10 km/hr (Table 2, Dummy 1 + Dummy 2) with a change in slope varying from –0.2 to –0.56; heavy snow caused a drop in free-flow speeds varying from 38 to 50 km/hr (Table 3, Dummy 1 + Dummy 2) with a change in the slope varying from –0.20 to –0.61. If flows of 20 veh/30 sec were still observed under heavy snow conditions, these coefficients suggest that the speeds would be reduced by more than 60 km/hr relative to dry weather. This effect stands out despite the variation in the data.

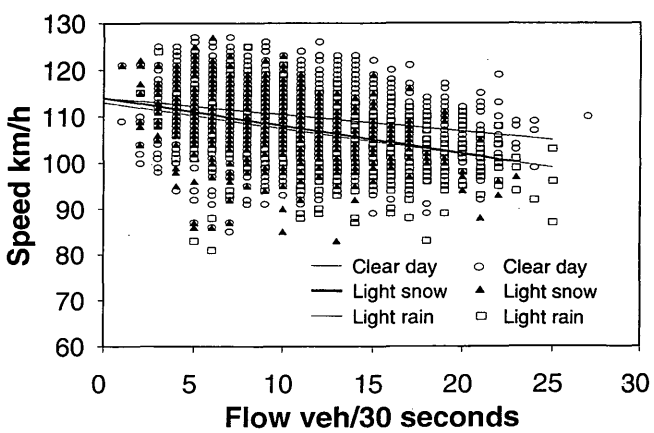


FIGURE 2 Speed-flow data (uncongested) and functions for clear, light snow, and light rain at Station 14, median lane.

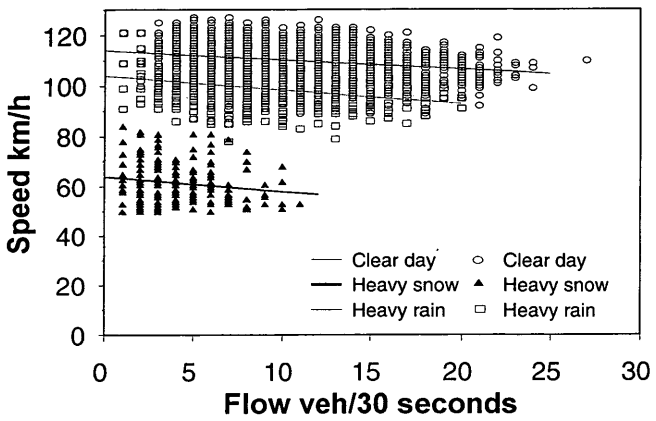


FIGURE 3 Speed-flow data (uncongested) and functions for clear, heavy snow, and heavy rain at Station 14, median lane.

Maximum observed flows were reduced during adverse weather. Heavy rain caused a reduction in the maximum flows of 10 to 20 percent, and heavy snow caused a reduction varying from 30 to 48 percent. Because of the nature of the sites selected, these flows are not capacity flows, but Figure 1 suggests that the flows are good approximations of capacity. These numbers come from only a small sample: two sites, a few hours of data, over a few days. Nevertheless, the changes in maximum observed flows may indicate the magnitude of changes in capacity flows under adverse conditions. Certainly the range of 10 to 20 percent reduction in maximum flows for rain is consistent with the average 14 percent reduction reported by Jones and Goolsby (*8,9*) and with the emphasis on the variability reported by Kleitsch and Cleveland (*10*).

In conclusion, adverse weather clearly affects both the flow-occupancy and speed-flow relationships. Other factors that may affect these relationships are a driver's familiarity with driving in rain and snow. One may expect a greater drop in speeds, for example, during snow in Washington, D.C., than in Ontario. The quality of drainage on the highway—whether it drains well or pools of water remain—can also affect the magnitude of the drop in speeds and flows due to adverse weather.

This paper has documented the range of effects quantitatively for both rainy and snowy conditions on the QEW highway in Ontario. Light precipitation, of either form, does not have a very large effect on any of free-flow speeds, maximum flows, or speed at maximum flows. Both heavy rain and snow can have great effects, such as a 50-km/hr reduction in free-flow speeds, and nearly a 50 percent reduction in maximum observed flows.

ACKNOWLEDGMENTS

The work reported in this paper has been taken from the master's thesis of the first author (*16*). The financial support of the Natural Sciences and Engineering Research Council of Canada for that work is gratefully acknowledged. The assistance of the Freeway Management Section, Ministry of Transportation of Ontario in providing the freeway data is much appreciated. This section also helped by making available the operators' log books from the FTMS control center in Mississauga. Thanks are also due to the Atmospheric Environment Service in Downsview for providing detailed weather data for Pearson International Airport, which helped to classify the weather conditions.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. Hall, F. L., and D. Barrow. Effect of Weather on the Relationship Between Flow and Occupancy on Freeways. In *Transportation Research Record 1194*, TRB, National Research Council, Washington, D.C., 1988, pp. 55–63.
3. Salonen, M., and T. Puttonen. *The Effect of Speed Limits and Weather to the Traffic Flow: Literature Study and Analysis of Traffic Flow on the Western Motorway of Helsinki* (in Finnish). Helsinki University, Finland, 1982.
4. Pursula, M., and T. Elolahde. Effect of Weather and Road Conditions on Motorway Traffic Flow (in Finnish). *Tie ja Liikenne*, Vol. 54, No. 7, 1984, pp. 278–280.
5. Maeki, S. *Effect of Weather and Road Conditions on Speeds* (in Finnish). Tielehti, Helsinki, Finland, 1972, pp. 353–360.
6. Situation-Dependent Control of Road Traffic on Urban Road Networks. PT., *Analysis of the Effects of Weather-Related Disturbances on the Flow of Traffic on Urban Road Networks*. Germany, 1986.
7. Andrey, J., and S. Yagar. A Temporal Analysis of Rain-Related Crash Risk. *Proc., 35th Annual Conference of the Association for the Advancement of Automotive Medicine*, Toronto, Ontario, Canada, 1991, pp. 486–483.
8. Jones, E. R., and M. E. Goolsby. The Environmental Influence of Rain on Freeway Capacity. In *Highway Research Record 321*, HRB, National Research Council, Washington, D.C., 1970.
9. Jones, E. R., and M. E. Goolsby. *Effect of Rain on Freeway Capacity.* Research Report 14-23. Texas Transportation Institute, Texas A&M University, College Station, Aug. 1969.
10. Kleitsch and Cleveland. *The Effect of Rainfall on Freeway Capacity.* Report Tr S-6. Highway Safety Research Institute, University of Michigan, Ann Arbor, 1971.
11. Drake, J. S., J. L. Schofer, and A. D. May. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154*, HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.
12. Cedar, A., and A. D. May. Further Evaluation of Single and Two-Regime Traffic Flow Models. In *Transportation Research Record 567*, TRB, National Research Council, Washington, D.C., 1976, pp. 1–15.
13. Easa, S. Selecting Two Regime Traffic-Flow Models. In *Transportation Research Record 869*, TRB, National Research Council, Washington, D.C., 1982, pp. 25–36.
14. Koshi, M., M. Iwasaki, and I. Ohkura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Flow Theory* (V. F. Hurdle, E. Hauer, and G. N. Stewart, eds.), University of Toronto Press, Ontario, Canada, 1981, pp. 403–426.
15. Hall, F. L., V. F. Hurdle, and J. H. Banks. Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992, pp. 12–18.
16. Ibrahim, A. T. *The Effect of Adverse Weather Conditions on Speed-Flow-Occupancy Relationships.* Master of Engineering thesis. Civil Engineering Department, McMaster University, Hamilton, Ontario, Canada, 1992.

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

Distribution-Free Model for Estimating Random Queues in Signalized Networks

ANDRZEJ TARKO AND NAGUI ROUPHAIL

A general-arrival, bulk service time queueing model is formulated for studying the distribution of random queues in signalized networks. The model is predicated on the occurrence of three traffic stream transformations: merging, splitting, and filtering. The model is applied to steady-state conditions (traffic intensity < 1.0) but can be eventually converted to a time-dependent form to account for oversaturation effects. A comparison of the results of the model with those of comparable models in the literature confirms that the use of random queue estimates derived from the assumption of a Poisson arrival process is inappropriate for networks. Marginal adjustments to the Poisson process by including a variance-to-mean ratio of the departure distribution improve the random queue estimate to a point. The results also confirm recent observations by Newell about the relationship of stochastic queues in an arterial network with their counterparts at isolated intersections. In general queue estimates for the network case are substantially smaller than those incurred at an isolated intersection with similar traffic intensity. The difference is attributable primarily to the process of traffic filtering.

Vehicle delays at signalized intersections contribute substantially to travel times on an urban street network. Delay is now the basic criterion for evaluating the level of service (LOS) at signalized intersections and a key ingredient for evaluating the LOS on arterials (*1*). The ability of a traffic analyst to estimate vehicle delay is critical in evaluating advanced traffic management systems (ATMS) as well as quantifying the environmental consequences of traffic decisions.

Average vehicle delay at a signalized intersection can be expressed as the sum of nonrandom and overflow delay components. Nonrandom delay refers to the average vehicle delay experienced with the assumption that traffic demand is uniform and averaged over all cycles during the analysis period. Overflow delay encompasses the additional delay caused by the randomness in arrival headways within each cycle and from one cycle to the next, in addition to that incurred when flow exceeds capacity for some period of time. Within-cycle random variations are usually negligible in terms of their impact on delay, an effect that is also not considered in this paper. Thus, the residual queue remaining at the end of the green phase (herein denoted as N_o) is considered the only source of overflow delay. The relationship between average overflow queue and average random delay d_o can be approximated for steady-state conditions as follows (*2*):

$$d_o = \frac{N_o}{q} \tag{1}$$

where

d_o = random delay (sec),
N_o = random queue (veh), and
q = arrival rate (veh/sec).

Since this relationship is straightforward and independent of the queueing model distributions, random delay is often investigated through the estimation of random queues.

BACKGROUND

Nonrandom delay formulas exist for both isolated and coordinated intersections (*1,3*). Estimating the second delay component for a signalized network is still a challenging research issue. Earlier theoretical work on queueing theory (*4–6*) hints at some major difficulties in obtaining delay formulas for general arrival and departure distributions. The most general steady-state delay models have been derived by Darroch (*5*), Newell (*2*), and McNeil (*7*), who incorporate the variance-to-mean ratio I_a in their models to include binomial or compound Poisson arrivals (Darroch's processes). Since these works did not deal directly with signalized networks, these questions remained: what actually are the arrival processes in signalized networks? and how is the value of I_a estimated if Darroch's processes are appropriate for signalized networks?

Van As addressed these issues using the Markov chain approach to model delays and arrivals at two closely spaced signals (*8*). He concludes that the Miller model improves random delay estimation for signalized networks in comparison with the Webster model. However, Van As's results also indicated that Miller's formula overestimated random delay in some cases. It is unclear whether that bias was caused by the non-Darroch's arrival process or by the coordinate transformation technique (*9*) used to obtain the time-dependent models investigated by Van As.

Tarko et al. have investigated the impact of an upstream signal on random delay using cycle-by-cycle macrosimulation (*10*). They found that in some cases the ratio I_a does not properly represent the non-Poisson arrival process and generally overestimates delay. The additional weakness of such models lies in the estimation of I_a. Although Van As worked out a straightforward formula for I_a, its dependence on the I_a calculated at an upstream signal creates the possibility of a systematic error propagation problem in the course of the calculations. To avoid that problem, Tarko et al. (*10*) proposed a random delay model that uses a function of the capacity differential between the critical upstream and subject signals instead of the I_a ratio to improve delay estimation in signalized networks. Their work also confirms that traffic platooning—that is, signal progression—in a signalized network operating on a common signal cycle has no effect on the cycle-to-cycle variation of the arrival

A. Tarko, Urban Transportation Center, University of Illinois at Chicago, 1033 West Van Buren, Suite 700 South, Chicago, Ill. 60607. N. Rouphail, Civil Engineering Department, North Carolina State University, P.O. Box 7908, Raleigh, N.C. 27695.

distribution process. In other words, signal offset does not affect random delays or queues.

The present paper can be seen as an extension of the author's previous work (*10*). A bulk service queueing model is presented that enables the description of the distribution of vehicle arrivals, departures, and random queues in a signalized network. The model is evaluated by comparing it with well-recognized random queue models for isolated intersections [Khintczine-Pollaczek, Newell (*2*), and Akçelik (*11*)] and for networks [modified Newell (*12*), Miller (*13*), and Tarko-Rouphail (*10*)]. Furthermore, Newell's hypothesis on the average random queue along a signalized arterial (*12*) is tested. Finally, a sensitivity analysis on the effect of secondary flows (midblock and turning movements) on random queues is presented and discussed.

ARRIVAL DISTRIBUTION IN A SIGNALIZED NETWORK

Consider an urban street network on which most intersections are signalized. An additional, and reasonably valid, assumption is that all these signals operate on a common signal cycle. The traffic stream moving through the network is subject to the following transformations: it can

- Merge with other traffic streams,
- Split into separate traffic streams, or
- Be filtered by traffic signals.

In such transformations, a traffic stream is represented by its arrival distribution in time periods that are equivalent to the common signal cycle. Arrival distributions are generated at locations where a given transformation takes place. For example, consider a traffic link connecting two signalized intersections (Figure 1). The link is modeled as a sequence of cross sections at which traffic streams are merged, split, or filtered. These arrival distribution transformations are modeled using the following processes:

- *Merging* produces a combined distribution of arrivals $P(x)$ from two independent traffic streams with arrival distributions $P_1(a)$ and $P_2(a)$ as follows:

$$P(x) = \sum_{a=0}^{x} P_1(a)P_2(x - a) \tag{2}$$

In cases of three or more streams, this formula is applied consecutively, so that $P_1(a)$ is the result from the previous application and $P_2(a)$ corresponds to the next stream to be combined.

- *Splitting* produces a distribution $P_s(x)$ of arrivals drawn with probability p from a traffic stream with known arrival distribution $P(a)$ according to

$$P_s(x) = \sum_{a=x}^{A} P(a) \frac{a!}{x!(a-x)!} p^x(1-p)^{a-x} \tag{3}$$

where A is the maximum number of arrivals considered to have a finite value. For entry links into the network, the Poisson distribution may usually be applied to estimate the number of arrivals. In this case the value of A is set sufficiently large to neglect the truncation error.

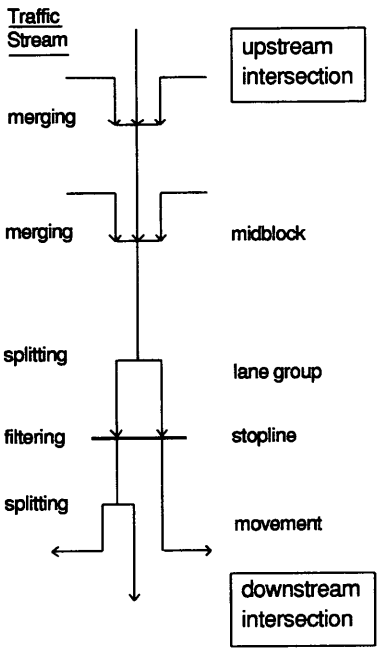


FIGURE 1 Traffic streams on a link in random queue modeling.

- Signal *filtering* transforms the arrival distribution $P(x)$ just upstream of an intersection into a departure distribution $P_d(x)$ just downstream of the intersection. The number of departures per cycle is equivalent to the sum of overflow queue (from previous cycles) in addition to all "new" arrivals in the subject cycle if that sum is less than the signal capacity. Otherwise the number of departures is set equal to the signal capacity

$$P_d(x) = \sum_{a=x}^{A} \Pi(a)P(x - a) \qquad \text{for } x < c \tag{4}$$

$$P_d(x) = 1 - \sum_{a=0}^{c-1} P_d(a) \qquad \text{for } x = c \tag{5}$$

where c is the fixed signal capacity per cycle and $\Pi(x)$ is the probability of x vehicles in queue at the end of green time in steady-state conditions.

Random queues on any link within the network can be handled by these transformations, first by modeling all the upstream links that feed into the subject link. For simplicity, a Poisson distribution may be assumed for arrivals on external links. Thus, the process of network link modeling is carried out in the following manner (Figure 1):

1. Combine the departure distributions from all exits at the upstream intersection.
2. Assume that the combined departure distribution in Step 1 constitutes the arrival distribution at the midblock unsignalized intersection (real or hypothetical). Combine this profile with the arrival profile from midblock traffic when applicable.
3. Assume, as in Step 2, that the combined profile at the midblock location constitutes the arrival profile at the downstream signal. Split the traffic stream in the segregation zone according to the prevailing lane assignment. For example, Figure 1 shows that through and

right-turning movements share common lanes. Thus, these movements are considered to form a single traffic stream. Left turners using exclusive lanes form a separate queue and are considered as a separate traffic stream that "splits" from the combined profile derived in Step 2.

4. Filter all separate traffic streams at the stopline. First, the random queue distribution is obtained from the arrival distribution and signal capacity. Next, the random queue distribution and the arrival distribution are used to produce the departure distribution. For example, the resulting departure profile for left turners is a final profile and may be used in modeling the appropriate downstream link.

5. Split shared traffic streams into individual movements (right turns from through traffic in Figure 1). The movements' departure distributions complete the requirements for processing the subject link.

6. Repeat Steps 1 through 5 for the downstream intersection.

This model assumes the arrival distribution to be identical to the upstream departure distribution. This assumption is valid if variations in vehicle speeds between the two intersections do not affect the number of arrivals in cycles at the downstream intersection. For long road sections, this assumption results in some underestimation of the random queues. The variations in the arrival distribution between cycles, which occur along road sections on which traffic is uninterrupted, require additional research. The arrival distribution considered here should not be confused with the average flow rate profiles used in models such as TRANSYT.

In the next section, the distribution of the random queue is modeled assuming a statistical equilibrium state. The random queue distribution is of main interest since it can be used to calculate expected random queues and delays. Splitting and merging transformations yield intermediate results that are required to either model random queues at the subject signal or continue modeling downstream links.

RANDOM QUEUE MODEL FOR GENERAL ARRIVAL DISTRIBUTIONS

A queueing system with a single server, random arrivals from a Poisson distribution, a deterministic bulk service, and queue discipline FIFO has been applied to random queue modeling at an isolated signalized intersection (*4*). The assumption of Poisson distribution is, however, too restrictive for signalized networks. Instead assume a general arrival distribution $P(x)$, where x is the number of arrivals in cycle, and overflow queue distribution in steady-state conditions $\Pi(k)$, where k is the number of vehicles in queue when green time ends. The signal capacity, expressed in number of vehicles that can be possibly served during cycle, is fixed and equal to c. An analysis of state transient probabilities under equilibrium conditions resulted in the system of balance equations for a general arrival distribution that is applicable to a signalized network:

$$\begin{aligned} &\sum_{i=0}^{\infty} P(x \le c - i)\Pi(i) - \Pi(k) = 0 \qquad \text{for } k = 0 \\ &\sum_{i=0}^{\infty} P(x = c - i + k)\Pi(i) - \Pi(k) = 0 \qquad \text{for } k = 1, \ldots, \infty \end{aligned} \tag{6}$$

To avoid the trivial and infeasible solution [$\Pi(i) = 0$ for each i], the first equation in the system (Equation 6) is substituted with the constraint on steady-state queue probabilities such that all probabilities $\Pi(i)$ sum to 1:

$$\sum_{i=1}^{n=\infty} \Pi(i) = 1 \tag{7}$$

The proposed queueing model gives an exact solution to the problem under steady-state conditions. However, to solve the equation numerically, the size of the problem must be limited to some finite and large numbers i and k, such that $i = k$. Finally, the system of linear equations in the standard form, convenient for many solution techniques, is

$$(\mathbf{a} - \mathbf{e}) \cdot \mathbf{\Pi} = \mathbf{b} \tag{8}$$

where

$\mathbf{a}$ = two-dimensional matrix ($M \times M$) with elements as follows:
$a_{1j} = 1$ for $j = 1, \ldots, M$
$a_{ij} = P(x = c + i - j)$ for $i = 1, \ldots, M$ and $j = 2, \ldots, M$
$\mathbf{e}$ = two-dimensional matrix ($M \times M$) with elements as follows:

$$e_{ij} = \begin{cases} 1 \text{ for } i > 1, j > 1, i = j \\ 0 \text{ otherwise} \end{cases}$$

$\mathbf{\Pi}$ = column vector of probabilities [$\Pi(k = 0)$, $\Pi(k = 1)$, . . . , $\Pi(k = M - 1)$];
$\mathbf{b}$ = column vector with M elements [1, 0, 0 . . .]; and
M = sufficiently large number such that truncation error is negligible for solution of system (8).

In the proposed model, signal capacity is assumed to have a fixed value. Olszewski (*14*) concluded that under reasonable capacity conditions, the use of a fixed value rather than a distribution is acceptable for unopposed traffic streams. The question arises whether this finding is applicable to a signalized network, since even small variations in the upstream signal capacity are propagated downstream when filtering takes place.

To answer this question, one should recognize that the principal source of capacity variations for unopposed streams is the cycle-to-cycle variations in traffic composition. However, the traffic composition at an upstream signal tends to be replicated downstream since the same vehicles arrive at both signals with some time lag. This means that the upstream and downstream signal capacities do not vary independently, which reduces the effect of capacity variations. Estimating the random queues for two cases—fixed and independently varying capacities—yields lower and upper bounds of random queue lengths. This is beyond the scope of this work.

COMPARISON WITH EXISTING MODELS

Single Intersection Models

The first step in the model evaluation is to compare its estimates with those of several well-known models: the Khintczine-Pollaczek (K-P) second term used by Webster (*3*) and Kimber and Hollis (*9*) in their formulas, Akçelik (*11*) random delay formulas, and the Newell model (*2*) modified by Cronje (*15*). These models converted into the random queue models according to Equation 1 are presented here:

- K-P for Poisson arrivals and deterministic departure processes:

$$N_o = \frac{X^2}{2(1 - X)} \tag{9}$$

- Akçelik:

$$N_o = \frac{1.5[X - (0.67 + c/600)]}{1 - X} \quad (10)$$

- Newell (with Cronje modification) for Poisson arrivals:

$$N_o = \frac{H(\mu)X}{2(1 - X)} \quad (11)$$

where

$$H(\mu) = \exp[-(1 - X)c^{0.5} - 0.5(1 - X)^2 c] \quad (12)$$

In these equations, X is the degree of saturation, and c is cycle capacity in vehicles per cycle.

These models are compared with the authors' results and with the results obtained by Olszewski (*14*). The comparison is presented in Figure 2 for degrees of saturation 0.90 and 0.95 and for capacities varying from 10 to 120 veh/cycle. Olszewski's model based on the Markov chain produces virtually identical results to the modified Newell model and, therefore, is omitted from the comparison. The results demonstrate the significant effect of cycle capacity even for a fixed degree of saturation. The K-P model highly overestimated random queues since it does not consider the bunching of serviced vehicles during the green signal. Excellent agreement is evident between the bulk service and Newell models. Observed discrepancy between the Akçelik model and other models is a result of linearity of the first one.

Network-Based Models

The second step in model evaluation is to compare the steady-state K-P, Miller (*13*), Tarko-Rouphail (*10*), and Newell (*2*) with Cronje modifications (*15*) models with the bulk service model estimates. All these models can be represented using the following generalization:

$$N_o = \frac{k \cdot (X - X_0)}{(1 - X)} \quad (13)$$

where

X = degree of saturation;
k = $0.5X$ in K-P formula;
= $0.52 \times I_a \times X$ in Miller formula;
= $I_a \times H(\mu)$ in modified formula [$H(\mu)$ is calculated according to Equation 11];
= $0.408[1 - e^{-0.5(c_u - c_d)}] \times X$, where c_u and c_d are the cycle capacities for the critical upstream and subject signals, respectively, in Tarko-Rouphail formula; and
X_0 = 0 in K-P, Newell, and Miller models
= $Q_d/100$ in Tarko-Rouphail formula.

In Miller's and Newell's models, I_a is meant to incorporate the effect of non-Poisson arrivals ($I_a < 1$), and $H(\mu)$ incorporates the bulk service in Newell's model. The Tarko-Rouphail formula includes an adjustment factor as a function of signal capacities.

To provide data for comparison, a system of three signals with no turning movements is considered. Capacities for the first and second intersections were allowed to vary from 30 to 40 veh/cycle in 1-veh increments from one computation to the next. The third signal had a fixed capacity of 30 veh/cycle. Traffic volume was also fixed at an average of 27.5 veh/cycle, resulting in a fixed degree of saturation of 0.92 at the third intersection. Poisson arrivals were assumed at the first (entry) signal. Here it is recognized that the upstream signals will substantially transform the arrival pattern at the downstream intersections. Comparative results are depicted in Figure 3 for the second and third signals and for cases in which the random queue is non-zero. As expected, the K-P model overestimates random queues. The addition of the I_a parameter to that model (Miller) improves its estimates. However, the lack of the bulk vehicle service property in both models still resulted in an overestimation of random queues. The modified Newell and Tarko-Rouphail formulas are comparable to the bulk service model.

Newell's Hypothesis

Newell recently discussed an interesting hypothesis. He suggested that from the standpoint of random queues, a signalized arterial can

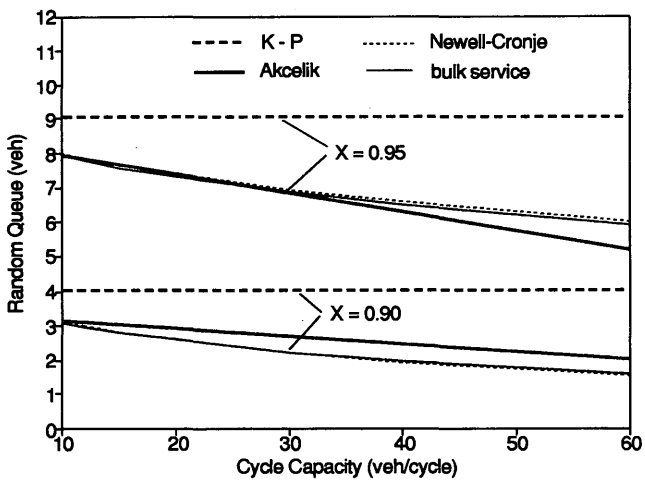


FIGURE 2 Comparison of random queue models for isolated signals.

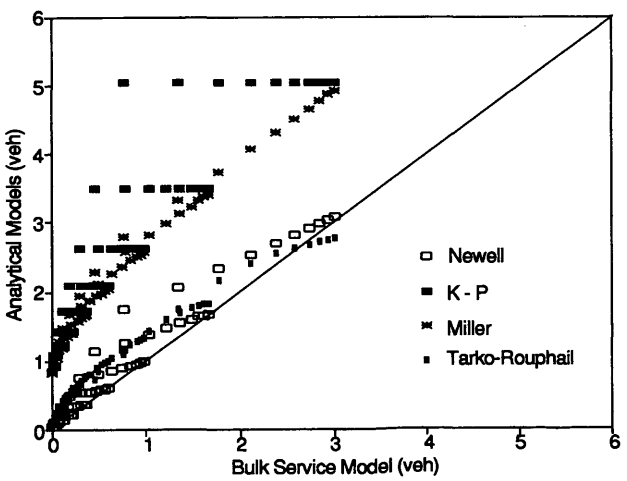


FIGURE 3 Comparison of random queue models for signalized networks.

be very easily considered as one system (*12*). He went on to state that the total random queue (and random delay) along *all* arterial signals is equivalent to the random queue that would be observed at the critical intersection were that intersection operating in isolation. The (limiting) assumption is made that there are no turning movements along all subject intersections. Approximate random delay formulas were also developed when turning movements are present, but only for modest levels.

Newell's hypothesis was tested using the same system of three signals described earlier. Figure 4 depicts the total random queue at the three signals as a function of the individual signal capacities. The results confirmed Newell's thesis that the total random queue along a signalized arterial is much lower than the total random queue if all intersections are treated as isolated (current state of the art). The results indicate an even stronger reduction than that hypothesized by Newell. It appears that Newell's estimate should be considered as an upper bound for total random queue, at least in cases in which turning movements are negligible. For practical purposes, however, his hypothesis provides much better random queue estimates than most of the formulas cited earlier.

ILLUSTRATIVE EXAMPLES

Two examples are provided to illustrate the model sensitivity to key traffic parameters and to highlight one or more stream transformations described earlier. In the first example, a single traffic stream is examined between two intersections (Figure 5, *top*). The arrival distribution at Intersection 1 is described by a Poisson process. Filtering at Intersection 1 causes a significant reduction in the variability of the departure process (Figure 5, *middle*). The resulting random queue distribution at the second signal is compared with the distribution when the upstream signal does not exist (Figure 5, *bottom*). In this case, the expected random queue at Intersection 2 is less than a third of the value computed assuming no filtering (2.60 versus 7.95). Furthermore, the total expected random queue in the system (1.82 + 2.60 = 4.42) falls far short of the expected queue length at Signal 2, assuming random arrivals (7.95). This confirms the results shown in Figure 4.

In the second example the sensitivity of the expected random queue at a downstream intersection to midblock flow levels is inves-

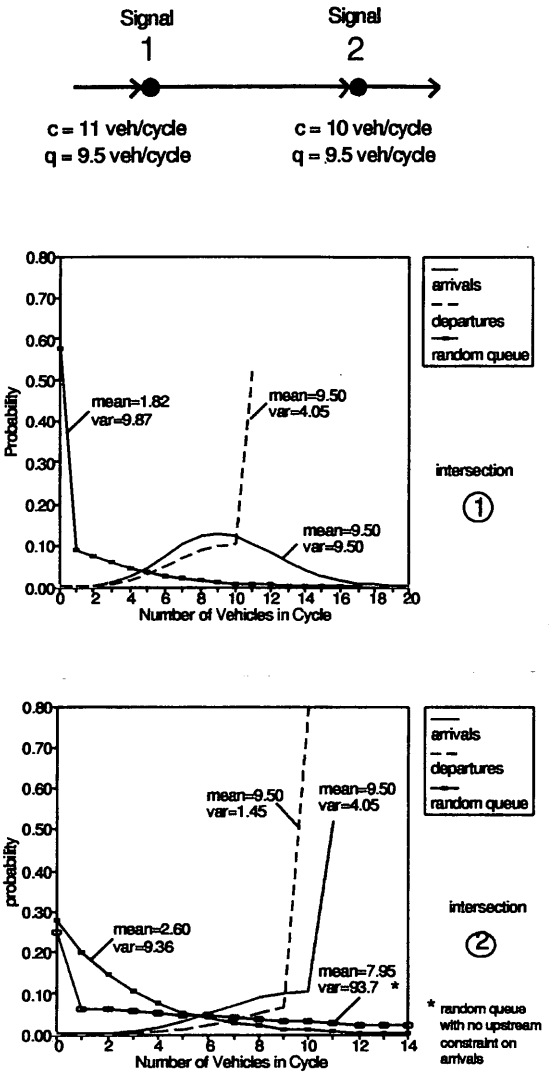


FIGURE 5 Single stream example.

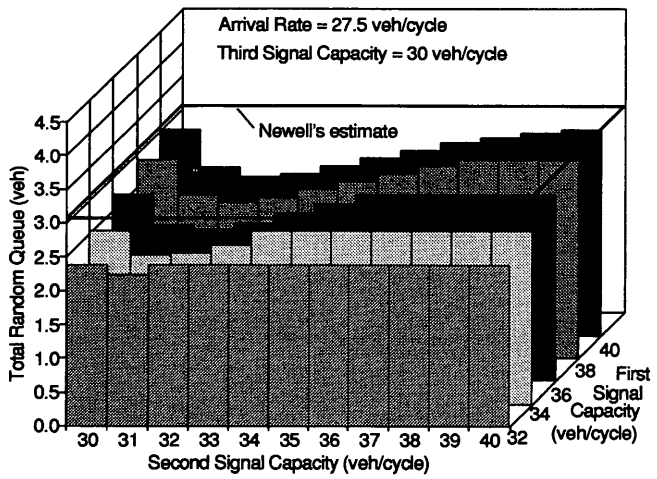


FIGURE 4 Newell's hypothesis evaluation.

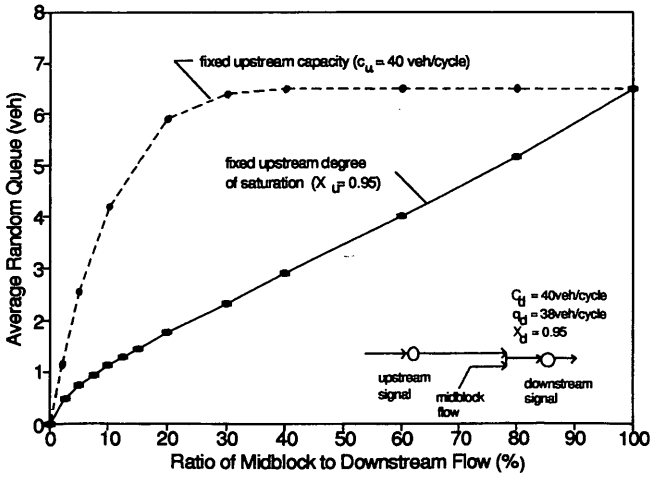


FIGURE 6 Midblock flow effect on random queue.

tigated (Figure 6). Here, stream merging and filtering effects are examined. Downstream conditions are kept fixed, including signal capacity, flow, and degree of saturation. For the upstream conditions, two scenarios are analyzed. In the first (dotted line in Figure 6), the upstream signal capacity is kept fixed while the midblock flow contribution to the total flow is allowed to increase. Consider the point at which the ratio of midblock to total flow is 30 percent. Here, the random queue reaches its maximum value. The upstream signal contribution is 70 percent of the total flow, or 26.6 veh/cycle, and its capacity is 40 veh/cycle, yielding a degree of saturation equal to 0.67. Consequently, the departure distribution is virtually unaffected by the capacity constraint (i.e., negligible filtering) and can be reasonably approximated by a Poisson process. Obviously, the combination of two Poisson processes (from signal and midblock) also produces a Poisson process with an equilibrium queue equivalent to the maximum value indicated in Figure 6.

In the second scenario (solid line), the midblock flow contribution to total flow is also increased, but the degree of saturation at the upstream intersection is maintained as fixed (0.95): for example, emulating the operation of an actuated or adaptive controller operation. Consequently, upstream filtering is active at all flow levels. The expected random queue at the downstream intersection varies almost proportionally to the portion of midblock flow in the total stream. The results clearly demonstrate the significance of midblock and turning (i.e., unfiltered) flows on random queue estimates and justify further research to incorporate that effect into analytical models of random delays and queues for signalized networks.

CONCLUSIONS

A general-arrival, bulk service time queueing model has been formulated for the study of random queues in signalized networks. The model is predicated on the occurrence of three traffic stream transformations in the network: merging, splitting, and filtering. The model is applied to steady-state conditions (traffic intensity < 1.0) but can be eventually converted to a time-dependent form to account for the effects of oversaturation. The study yielded the following conclusions:

1. Models for random queues (or delay) that are based on the Poisson arrival process (e.g., isolated intersections) are not generally transferable to networks because of the filtering effect of upstream signals.
2. Filtering tends to reduce the size of random queues. Although this finding is consistent with earlier observations by Newell (*12*), the observed reductions were even higher than Newell's estimates.
3. When two or more traffic streams merge, the resulting downstream random queue is dependent on the level of filtering that has taken place before merging. If streams are unfiltered (e.g., midblock flows, or signal departures at very low volume-to-capacity ratios), the random queue will be similar to that expected at an isolated intersection. Highly filtered streams, however, can substantially reduce random queues.
4. There is a need to consider incorporating the model results into network signal timing software, since many control strategies use a minimum delay or queue criterion for signal optimization. If methods for estimating random queues in networks must be revisited, then signal strategies that rely on such estimates should be examined.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. Newell, G. F. Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. *SIAM Review*, Vol. 7, 1965.
3. Webster, F. V. *Traffic Signal Settings.* Road Research Laboratory Technical Paper 39. Her Majesty's Stationery Office, London, England, 1958.
4. Haight, F. A. *Mathematical Theories of Traffic Flow.* Academic Press, New York, 1963.
5. Darroch, J. N. On the Traffic-Light Queue. *Annals of Mathematical Statistics*, No. 35, 1964, pp. 380–388.
6. Gazis, D. C. *Traffic Science.* Wiley-Interscience, 1974, pp. 148–151.
7. McNeil, D. R. A Solution to the Fixed-Cycle Traffic Light Problem for Compound Poisson Arrivals. *Journal of Applied Probability*, No. 5, 1968, pp. 624–635.
8. Van As, S. C. Overflow Delay at Signalized Networks. *Transportation Research*, Vol. 25A, No. 1, 1991, pp. 1–7.
9. Kimber, R., and E. Hollis. *Traffic Queues and Delays at Road Junctions.* Report 909. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1979.
10. Tarko, A., N. Rouphail, and R. Akçelik. Overflow Delay at a Signalized Intersection Approach Influenced by an Upstream Signal: An Analytical Investigation. In *Transportation Research Record 1398*, TRB, National Research Council, Washington, D.C., 1993, pp. 82–89.
11. Akçelik, R. *Traffic Signals: Capacity and Timing Analysis.* Research Report 123. Australian Road Research Board, Nunawading, 1981.
12. Newell, G. F. Stochastic Delays on Signalized Arterial Highways. *Proc., 11th International Symposium on Transportation and Traffic Flow Theory*, Elsevier, New York, 1990, pp. 589–598.
13. Miller, A. J. *Australian Road Capacity Guide—Provisional Introduction and Signalized Intersections.* Bulletin 4. Australian Road Research Board, Nunawading, 1968 (superseded by Research Report 123, Australian Road Research Board, 1981).
14. Olszewski, P. Modelling of Queue Probability Distribution at Traffic Signals. *Proc., 11th International Symposium on Transportation and Traffic Flow Theory*, Elsevier, New York, 1990, pp. 569–588.
15. Cronje, W. B. Analysis of Existing Formulas for Delay, Overflow, and Stops. In *Transportation Research Record 905*, TRB, National Research Council, Washington, D.C., 1983, pp. 89–93.

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

Variability Analysis of Traffic Simulation Outputs: Practical Approach for TRAF-NETSIM

RAHIM F. BENEKOHAL AND GHASSAN ABU-LEBDEH

Stochastic traffic simulation models, such as TRAF-NETSIM, use random number seeds to generate variables to describe driver, roadway, and traffic characteristics. In analyzing outputs from these models, one should consider the variability of the responses. The variability of NETSIM's output using the methods of replication and batch means was explored. For the batch means method, it is proposed to compute the measures of effectiveness (MOEs) for intermediate time intervals using a proposed interval calculation (PIC) procedure. The MOEs were evaluated at the network, intersection, and link levels of aggregation. Depending on the MOE and level of aggregation, the two methods yielded significantly different results. Hence, depending on the study objective, outputs may need to be examined at different levels of aggregation to obtain meaningful results. The practical implications of the variability are also discussed, and statistical approaches are proposed to deal with output variability. Auto- and cross-correlations must be examined explicitly, particularly when dealing with link MOEs resulting from very short simulation time. Ignoring positive cross-correlation is not detrimental but leads to more conservative confidence intervals. Either the batch means with PIC method or replication method must be used to build confidence intervals. NETSIM's direct output for intermediate time intervals should not be used to build a confidence interval unless an autocorrelation analysis is done. Not using proper statistical procedures can lead to erroneous and misleading conclusions.

Computer simulation models have been used, as a decision tool, to evaluate the effects of alternative traffic control strategies. Simulation results, however, will vary when either the sequence of the random numbers used (internal variables) or the input variables (external variables) are changed. Stochastic traffic simulation models, such as TRAF-NETSIM (henceforth NETSIM) (*1*), use random number seeds (RNSs) to generate random variates to describe roadway, traffic, and driver characteristics. Using a different RNS changes the outcome of the simulation model.

Misleading and erroneous conclusions may be obtained if the variability in NETSIM's output is not considered, especially when alternative measures are being evaluated or the effects of certain traffic-related changes are being quantified. There is, however, no clear guideline on ways to handle the output variability, the length and number of runs needed, and the magnitude of the effects of internal variables on the simulation results. This paper presents a practical approach to deal with these issues. It also examines auto-correlation and cross-correlation issues in NETSIM.

R. F. Benekohal, Civil Engineering Department, University of Illinois at Urbana-Champaign, 205 North Mathews Avenue, Urbana, Ill. 61801. G. Abu-Lebdeh, Champaign-Urbana Urbanized Transportation Study, 1303 North Cunningham Avenue, Urbana, Ill. 61801.

PROBLEM STATEMENT

Consider a situation in which one wants to compare two conditions, such as assessing the traffic impact of a new office park on an existing network, using NETSIM. When running a stochastic simulation model such as NETSIM, a few options are available. The easiest and the most widely used option is running NETSIM for conditions "with" and "without" the office park traffic and comparing the results. However, one may get misleading results using this approach, as illustrated by the following example. Assume that the measure of effectiveness (MOE) used is the average delay per vehicle. Assume that the delay for the base condition is X and that when the office park traffic is added to the network the delay is Y. Does $(Y - X)$ provide enough information to assess the impact of office park traffic? If it does, how large should $(Y - X)$ be to be considered a significant impact?

These questions appear to apply when other software, such as Highway Capacity Software (HCS) (*2,3*), is used for impact assessment. However, a major difference between NETSIM and models such as HCS is that the former is a stochastic model and the latter is a deterministic model. Thus, for a given traffic and roadway condition, HCS results do not vary when the sequence of vehicle arrivals is changed. However, NETSIM's results do vary when the sequence of random events (e.g., sequence of arrival of vehicle) is changed even though traffic and roadway conditions remain unchanged.

In NETSIM the characteristics of a vehicle-driver unit are assigned randomly upon arrival of that vehicle into the system. Assume that using an RNS, the sequence of arrival of vehicles on a given approach is red car, blue car, and white car. Using a different RNS (while keeping traffic and roadway conditions the same) may result in the sequence white car, red car, and blue car. Running NETSIM with these two sequences of arrivals yields different results. It should be noted that in both runs, all input data remain the same except the cars' order of arrival.

Considering that NETSIM's outputs vary due to changes in the internal variables, one should not rely on the difference between Y and X without knowing the variability caused by the change in the sequence of random events. One cannot assess the impact of the office park correctly by getting two delay values from two long runs of NETSIM. Another important question is whether the impact should be assessed at the network level, intersection level, or approach level. If link and intersection data are used, which link or intersection should be used to represent the impact of the office park traffic? If network data are used, how large should the network be?

There are several approaches to dealing with output variability (*4*); the two most widely used are replication and batch means. For several well-known queueing and inventory systems, Law com-

pared the batch means and replication methods and concluded that batch means was superior (*5*). However, Law's findings are not directly applicable to traffic simulation models because traffic flow in urban networks does not necessarily resemble queue behavior.

Gafarian and Halati indicated that developing a confidence interval on the basis of a single run of NETSIM (batch means method) is extremely complex because it involves estimating auto- and cross-correlations of numerator and denominator variables (*6*). This is so because certain NETSIM MOEs, such as average speed, are estimated on the basis of ratios of sample means of observations that are auto- and cross-correlated. They analyzed the output directly produced by NETSIM (henceforth BM/direct) for a single intersection and recommended not using the batch means method. They suggested using the replication method and considering the covariance of the numerator and denominator variables. (It should be noted that the direct NETSIM outputs are cumulative statistics that are inherently correlated.)

Chang and Kanaan applied NETSIM to a congested isolated intersection to assess the variability of NETSIM's output (*7*). They used replication and BM/direct approaches and suggested using the approach given by Fishman (*8*) to find the appropriate batch size. The procedure is complex because it needs an autoregressive analysis. The use of the direct method could explain, at least partly, the correlation that existed among the data that Chang and Kanaan used. Thus, it is not clear when and how the batch means or replication method may be used.

BACKGROUND

The applications of NETSIM have been numerous and diverse, ranging from simulation of complex conditions (*9*) to simulation of simple systems such as alternative control strategies at single intersections (*10*). NETSIM proved to be a very powerful and flexible tool for the entire range of applications. Several case studies involving NETSIM have demonstrated clearly that the program can be used effectively to simulate unconventional settings in which traffic other than automobiles, buses, or trucks is involved (*11–13*).

The NETSIM applications can be grouped into three general categories:

1. Studies evaluating traffic control and geometric alternatives.
2. Studies assessing alternatives as well as NETSIM itself.
3. Studies focusing on NETSIM. Studies in this category can be divided into three general subcategories: those aimed at
   - –Addressing the degree of accuracy of NETSIM by either comparing its results with results from other software or validating them in the field,
   - –Dealing with the issue of variability of NETSIM's output, and
   - –Exploring NETSIM's potential and flexibility as well as its strengths and weaknesses.

This classification is subjective and some studies may fit into more than one group.

Category 1 includes studies that used NETSIM to evaluate different geometric alternatives (*14*) and different signal control strategies and timing plans to optimize travel time and fuel consumption (*11,15–19*). Yauch et al. (*12*) used NETSIM to assess the impact of drawbridges, and Luedtke (*13*) used it to simulate the impact of light rail transit on signal operations. Others used it to compare the results from different software with those of NETSIM (*20,21*). Rathi and Lieberman used NETSIM to evaluate the potential effects of restricting traffic flow on approaches to a congested urban street network (*9*). Papacostas and Willey used NETSIM for analyzing the traffic impact of a real estate development (*22*). Only in limited cases were field observations taken to verify the results. For the most part, however, little or nothing was done to address or account for the issue of variability of the program's output.

Category 2 includes studies that had focuses similar to those of Category 1, except NETSIM itself was of equal importance as some efforts were made to verify or validate the simulation results. Hurley and Radwan used NETSIM to study different aspects of traffic flow and to estimate the impact of various traffic control settings on fuel consumption and delay (*23*). Wong compared field observations with capacity and level of service estimates from NETSIM and HCS (*24*). Radwan and Hatton used NETSIM to simulate traffic operations at conventional and single-point diamond interchanges (*25*). Ten replications were used to minimize the effect of variation. Kim and Messer used NETSIM to evaluate different control strategies for saturated signalized diamond interchanges (*26*). To account for output variability, the same RNSs were used for the paired simulation trials. Torres et al. (*27*) used NETSIM to estimate the impact of lane obstruction on arterial streets. To account for output variability, three replications were used to quantify the significance of any particular combination of factors.

The first subcategory includes those studies that aimed at evaluating NETSIM by comparing its results with those of other programs or with field data. Yagar and Case addressed the issue of sensitivity of NETSIM/UTCS-1 to aggregation of traffic flows and to the RNS (*28*). Davis and Ryan used NETSIM to estimate delay and queue length at an isolated intersection (*10*).

The second subcategory includes those studies that dealt specifically with variability of the NETSIM's output. This issue has come up in several of the studies that used the model, but for the most part the users did not address it. Those who did mostly used multiple runs (*19,21,25–27*). However, the degree of success in neutralizing the effect of variability was hardly addressed. Recent acceptance and widespread use of NETSIM, along with improved features of the program gave rise to new efforts aimed at reducing or otherwise dealing with such variability through variance reduction techniques.

Variance reduction techniques were applied to NETSIM to assess their effectiveness. Rathi and Santiago used the common RNS (*29*), and Rathi and Venigalla applied antithetic variates techniques (*30*). Regardless of the variance reduction issue, the authors stressed that statistical analysis must be used to interpret the simulation output data properly. It should be noted that in applying the variance reduction techniques to TRAF-NETSIM, they assumed that the desired correlations (e.g., synchronization) are attainable with the way TRAF-NETSIM generates random variables.

The last subcategory includes studies that aimed at exploring NETSIM's potential and flexibility as well as drawing attention to its strengths and weaknesses. Wong used the detailed simulation capabilities and graphics of NETSIM to estimate capacity and level of service (*31*). Very few guidelines, however, were given on ways to run TRAF-NETSIM and to deal with the output variabilities.

STUDY APPROACH

Analysis Approach

Batch means and replication methods were used to assess the variability in NETSIM's MOEs due to changes in the internal variables.

These methods were applied to two cases: the base condition (Case 1) and the base condition with volume added (Case 2). For each case, the MOEs for the network, a typical link, and a typical intersection were analyzed using the two methods.

The results for the replication method were obtained by running NETSIM 24 times each for 10 min. The independent replications were achieved by using 24 RNSs from a random number table (*32*). The selected seeds satisfied the NETSIM's requirements for RNS. The results from the batch means were obtained by running NETSIM once for 4 hr, with intermediate results computed every 10 min (batch size = 10 min) using the proposed interval calculation (PIC) procedure.

Proposed Interval Calculation Method

If the true variability of the different MOEs among the intermediate intervals of a long run is to be determined, the statistics must be calculated for the individual batches. The PIC method computes such MOEs. In the PIC method, vehicle trips and phase failures are computed by finding the differences between successive batches. Delays and speeds are computed as described in the following.

To compute the average delay per vehicle for each batch, the total time—which is the sum of the link travel times for all vehicles in that batch—is divided by the number of vehicle trips in that batch. To find the number of vehicle trips or the total travel time for a specific batch, the previous intermediate output values are subtracted from the current output values.

To find the average speed of vehicles during each batch, the total number of miles driven within that batch is calculated and then divided by the total time the vehicles spent during that batch to complete those miles. For illustration, batch statistics for the average delay time and speed are calculated here using direct NETSIM statistics given in Table 1.

PIC delay = (44.54 − 22.07) * 60/(2,227 − 1,110) = 1.207 min/veh-trip
PIC speed = (1,053.03 − 529.36)/(79.64 − 39.71) = 13.11 mph

NETSIM's direct results (BM/direct) were not used. The BM/direct method gives the mean MOEs for the entire simulation run up to that time, not the mean MOEs for each batch. Since confidence intervals and statistical tests cannot be constructed for the BM/direct method, it is dropped out of any further discussion. In some graphs the results from the BM/direct method are shown for comparison.

Description of Network

The batch means and replication methods were implemented on a nine-intersection network in downtown Champaign, Illinois (Figure 1). Dummy nodes were introduced to collect statistics at entry links. Traffic signals had cycle lengths of 60 sec, and the overall network was not congested.

MOEs Used

Three MOEs will be examined: average delay, average speed, and vehicle trips. In NETSIM, delay time is defined as the difference between the actual time that a vehicle spends in the system and the ideal amount of time based on free-flow speed. Speed is defined as the ratio of the total distance traveled by all vehicles in the system to their total travel time. A vehicle completes a trip on a link when it passes the stopline.

TABLE 1 Direct NETSIM Statistics Used To Calculate Average Delay Time and Speed

Elapsed Time	Vehicle-mi	Trips	Delay Time (veh-hr)	Total Time (veh-hr)
0:10:00	529.36	1,110	22.07	39.71
0:20:00	1,053.03	2,227	44.54	79.64

Statistical Methods

Batch Means Method

The batch means method is performed by running the simulation model for one long run and then dividing it into smaller time intervals (batches). For each batch, statistics are collected and variability among batches is used to build a confidence interval on the simulation output. If the batches are long enough, the means from the batches may be uncorrelated. Increasing the length of the batches may reduce their autocorrelation. The advantage of batch means over the replication method is in having only one initialization period. However, the correlation problem between batches, the length of each batch, and the number of batches must be determined carefully.

Replication Method

The replication method is performed by running the simulation for a number of independent runs. The independent simulation runs are made for the same roadway and traffic conditions. Each run will have an initialization time until the system reaches equilibrium condition. After the warm-up time, statistics on system performance are collected. The advantage of this method is that the autocorrelation is eliminated. However, one must know the length of each run and the number of replications. These will depend on the range of variability of the responses. One obvious disadvantage of this method is that each run needs its own initialization period. However, with the speed of today's computers, the initialization for each run may not be prohibitive.

Correlation Among Batches

Two methods are proposed for dealing with the correlation issues among batches. One is using time series analysis, and the other is checking the correlation coefficient. When a long run of NETSIM is divided into batches, statistics for these batches may be treated as stationary time series data. A time series is considered stationary when statistical properties (e.g., mean and variance) of the time series are essentially constant over time. Plot of the mean value for each batch against time will help to determine visually whether the time series is stationary. Analytical techniques can also be used to determine whether a time series is stationary.

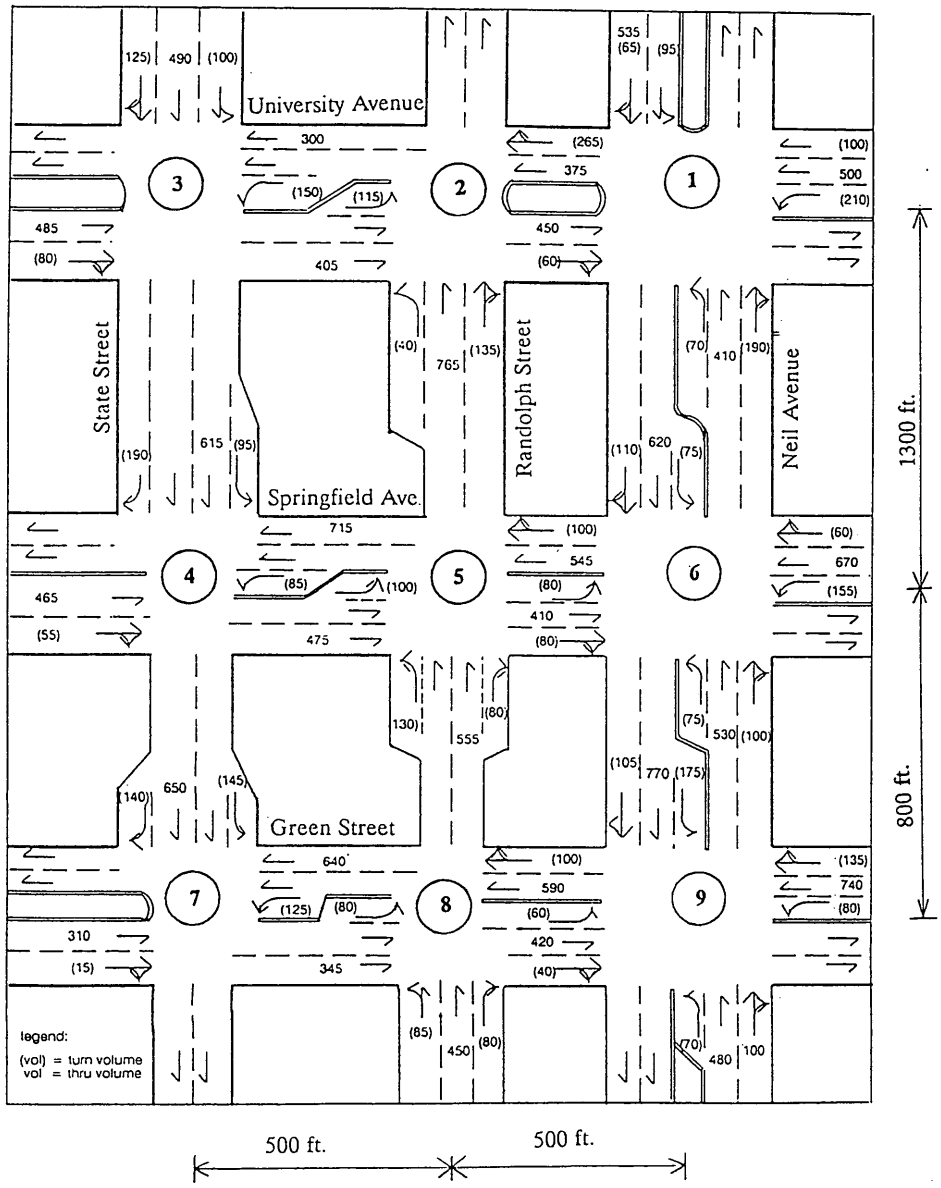


FIGURE 1 Network geometry and traffic volumes.

For a stationary time series of $Z_b, Z_{b+1}, \ldots, Z_n$, the sample autocorrelation at lag k, denoted by r_k, is computed from the following equation (*32*):

$$r_k = \frac{\sum_{t=b}^{n-k} (z_t - \bar{z})(z_{t+k} - \bar{z})}{\sum_{t=b}^{n} (z_t - \bar{z})^2}$$

where

$$\bar{z} = \frac{\sum_{t=b}^{n} z_t}{(n - b + 1)}$$

r_k measures the linear relationship between time series observations separated by a lag of k time units. The value of r_k will be between -1 and $+1$. When r_k is close to -1 or $+1$, the observations separated by a lag of k time units have a strong tendency to move together in a linear fashion (*33*).

The standard error of r_k is

$$S_{rk} = \frac{\left[1 + 2\sum_{j=1}^{k-1} (r_j)^2\right]^{1/2}}{(n - b + 1)^{1/2}}$$

The t_{rk} statistic is

$$t_{rk} = \frac{r_k}{S_{rk}}$$

TABLE 2 Autocorrelations for Average Delay for Lags 1–12

Network

Lag	r_k	$S_{r\,k}$	Lag	r_k	$S_{r\,k}$
1	.00939	.20412	2	-.08439	.20414
3	.12182	.20559	4	.06076	.20858
5	-.15485	.20931	6	-.14545	.21403
7	-.00879	.21811	8	-.01030	.21813
9	-.26227	.21815	10	-.05152	.23091
11	.17652	.23139	12	-.14364	.23694

Intersection 6

Lag	r_k	$S_{r\,k}$	Lag	r_k	$S_{r\,k}$
1	-.14666	.20412	2	-.29162	.20847
3	.25560	.22482	4	-.05741	.23662
5	-.03321	.23720	6	-.07393	.23740
7	-.21173	.23835	8	-.00162	.24607
9	.08054	.24607	10	-.05383	.24716
11	.02738	.24765	12	-.07025	.24778

Link 9-6

Lag	r_k	$S_{r\,k}$	Lag	r_k	$S_{r\,k}$
1	.15073	.20412	2	-.20663	.20871
3	-.18384	.21707	4	.14179	.22346
5	-.05707	.22718	6	-.36020	.22777
7	-.18585	.25039	8	.19916	.25607
9	.30415	.26244	10	-.01755	.27674
11	-.24191	.27679	12	-.05680	.28546

For the network and link delay r_k and s_{rk} values are given in Table 2. The values indicate that there was not a strong correlation at any lag. Lag 1 particularly is of interest to us, because it would indicate how strongly the adjacent batches are correlated. Furthermore, the t_{rk} values are smaller than 1.6. The t_{rk} values greater than 1.6 are considered to be statistically large for lags of 1, 2, and perhaps 3 (*33*). This indicates that there are no spikes at any lags in the data.

Another way of checking correlation between adjacent batches is looking at the correlation coefficient. A procedure that is much simpler than the time series analysis is suggested to examine the correlations among batch means. Correlation coefficient and Lag 1 autocorrelation would provide the same results. Thus, one may compute correlation coefficient (r) if dependency between adjacent batches are considered. The r-values will be between -1 and $+1$. Finding r and interpreting it is much easier than autocorrelation analysis.

DISCUSSION OF RESULTS

Table 3 contains a summary of the differences between the two methods or cases at the three levels of aggregation. The Yes or No entries indicate whether or not the differences between the values shown were statistically significant. Figures 2 and 3 depict the results graphically for the network and Link 9-6, respectively. Only discussion is provided for Intersection 6.

Case 1: Base Network

Average Delay

The average delays from the BM/PIC method were significantly higher than those from the replication method. Delay from the BM/direct method appears to converge to a constant value as the duration of the simulation increases. Some authors incorrectly considered this to be a sign of the stability of NETSIM's output (*30,34*). The convergence does not indicate that a "stable" condition is reached. This convergence is not a real phenomenon in the simulation model or real-world traffic; it occurs because the delay is computed from cumulative statistics. In fact, it is more realistic to have fluctuation in delay than the convergence. The delays computed by the BM/PIC method clearly show that NETSIM does not converge to a value. The cumulative statistics used in the direct method conceal this fluctuation.

Average Speed

The mean speeds estimated by the PIC method exhibited wider range and larger variance. The average speeds estimated by the PIC method were significantly lower than those from the replication method. Similar to delay, speeds from the direct method do not show the true speed fluctuation among batches. It is incorrect to assume that speed reaches a constant value when the duration of simulation is long. The relative speed differential between the two methods is much more pronounced at the link level than at the network level.

Vehicle Trips

The vehicle trips data are important because they are used to compute many of the NETSIM MOEs. At the network level, vehicle trips estimated by the replication method were significantly lower than those estimated by the BM/PIC method. At the link level, the estimated vehicle trips from the BM/PIC and replication methods did not show a statistically significant difference. At the intersection level, the mean vehicle trips estimated by the BM/PIC method are slightly higher than the replication method, but the difference is not statistically significant.

Case 2: Base Network with 120 Through Vehicles Added

The results for Case 1 indicated that the two methods, for the most part, give different results, even though the differences may seem small for practical purposes. The absolute difference between the two methods is not as important as the relative difference, which shows how much more traffic should be added to the network to increase the delay by the amount equal to the difference between the two methods. In fact, one needs to find out how much additional traffic would cause changes similar to those noted between the two methods, and how much added traffic can be "handled" within the internal variability of NETSIM.

In practical terms, one can ask whether NETSIM is sensitive enough to be used for traffic impact studies. Of particular importance is whether the impact assessment should be measured at the

TABLE 3 MOEs for Three Levels of Aggregation and Results of Comparisons Between Different Methods or Cases

Method/Case	Delay			Speed			Vehicle Trips		
	Network (sec./v.trip)	Link 9-6 (sec/v.trip)	Intersection 6 (sec/v.trip)	Network (mph)	Link 9-6 (mph)	Intersection 6 (mph)	Network (v.trips)	Link 9-6 (v.trips)	Intersection 6 (v.trips)
Replication	73.2	29.66	37.33	12.9	11.5	9.9	1110	473	108
Replication Compared to BM/PIC	Yes[1]	Yes	Yes	Yes	Yes	Yes	Yes	No[2]	No
BM/PIC	75.0	32.09	42.82	12.7	11.1	9.1	1123	475	109
BM/PIC Compared to BM/PIC- Added	Yes	Yes	No	Yes	Yes	No	Yes	Yes	Yes
BM/PIC-Added	83.4	35.83	43.06	12.0	10.7	9.0	1143	496	129

Yes[1]: Difference is statistically significant.
No[2]: Difference is not statistically significant.

network, link, or intersection level. To answer these questions, 120 through vehicles were added to all northbound links composing Neil Street. The number was selected such that all of the added vehicles will go through the links without oversaturating them. Output for network, Link 9-6, and Intersection 6 with the 120 vehicles added (henceforth BM/PIC) were used for comparison purposes.

At the network level, adding 120 vehicles greatly affects average delay, average speed, and number of vehicle trips. For Link 9-6, however, contrary to expectations, the changes in delay and speed are not significant, whereas changes in vehicle trips and phase failures are. At the intersection level, adding 120 vehicles has a significant impact on the three MOEs. Further examination of the other links indicated that the added volume affects the MOEs at links that did not receive additional traffic. This unrealistic effect is due mainly to the effects of external variables on the internal variables in NETSIM.

It should be noted that for any pair of comparisons, the MOEs either decreased or increased at all three levels. However, the amount of increase or decrease was markedly different. Although the changes in delay and speed for Link 9–6 due to the 120 added vehicles were not significant, the same changes attributed to the change of procedure (replication versus BM/PIC) were significant. It is conceivable that the reserved capacity of Link 9-6 may have partly concealed the impact of the 120 vehicles.

Practical Implications and Proposed Approach

The results presented in this paper have significant implications to the users of NETSIM. Such issues as the method to use (replication, BM/direct, or BM/PIC), the length and number of runs, and the amount of error to be tolerated were shown to be relevant. Given the stochastic nature of NETSIM, it may not be possible, or even necessary, to provide ready-to-use answers to all of these questions, but there should at least be some guidelines. The following section provides one possible approach to dealing with such issues.

The first question is that of which method to use: batch means (direct and PIC) or replication. Since statistical tests and confidence intervals cannot be performed easily with the BM/direct method, the question becomes whether to use the replication or BM/PIC method.

Other factors to be considered in selecting a method include human resources, computer time (both initialization and simulation), size of system, and previous experience with NETSIM. For large networks in which the initialization period is likely to be longer, it would be inconvenient and time-consuming to use the replication method. Replication will require more runs and may need more human resources, although this can be overcome with some programming. Currently, both methods require considerable time to perform statistical analyses. For the typical uncongested traffic system, neither method offers a clear advantage. The number of technical considerations could be the deciding factor.

Replication Method

The replication method has the advantage in that autocorrelation is eliminated as independence of observations is ensured through the use of different RNSs. The questions, then, are on the number and length of the runs—which are related characteristics. Answering the first question will automatically answer the second.

To start with, users can make X (say 10) runs of reasonable length. The length of each replication run would depend mainly on the size of the network and traffic conditions. In general, it should

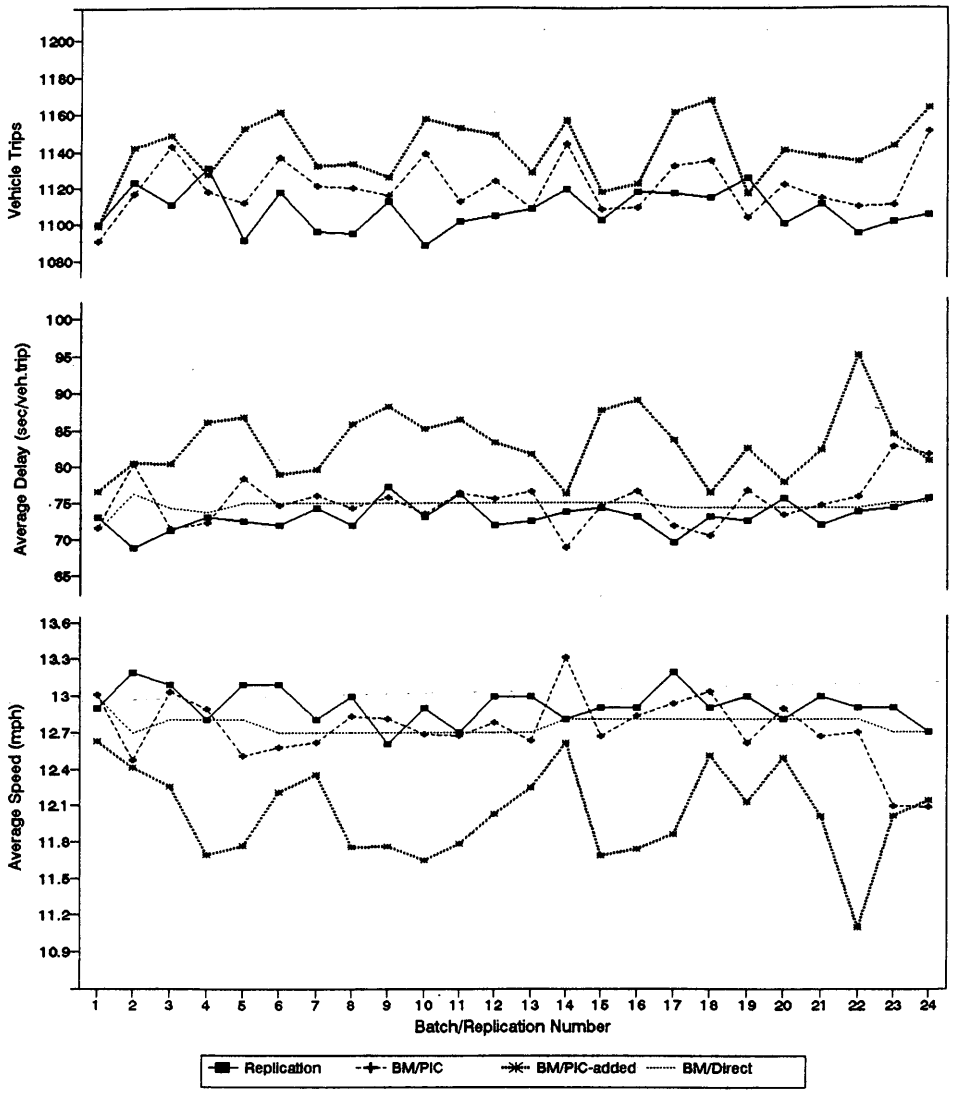


FIGURE 2 Comparison of delay, speed, and vehicle trips at network level.

be long enough that the user believes that it reflects real-world conditions. It is proposed to use a simulation time that is at least 10 signal cycle lengths for small networks. Another proposed way of estimating duration is to make it at least as long as the initialization time given by NETSIM.

Next, find the mean and variance for those observations (runs). Select a confidence level (usually 90 or 95 percent), then calculate the number of observations needed (n), which is a function of the standard deviation and the tolerable error, using Equation 1.

$$n = (ts)^2/e^2 \quad (1)$$

where

s = standard deviation of observations,
e = tolerable error, and
t = critical t-value from a t-distribution table for $(n - 1)$ degrees of freedom and selected confidence level.

If n is greater than X, make a few more runs and repeat the procedure. Once the computed n is less than X, stop. Construct confidence intervals using the results from X runs.

The tolerable error depends on the accuracy that the user desires, which depends on the distribution of data around the mean. For widely dispersed data, the tolerance can be high, but if the data are concentrated around the mean, the tolerance will be low. Thus, it should be a function of the range and dispersion of the data. It is suggested to use 5 to 15 percent of the range when a data set does not have extreme values. When a data set contains extreme values, discard them first and then find the range.

BM/PIC Method

If a user decides to use the BM/PIC method, the length of the simulation run, T, must be determined. T should be divided into X intervals (e.g., start with 10 intervals), each being T/X. The value of T/X

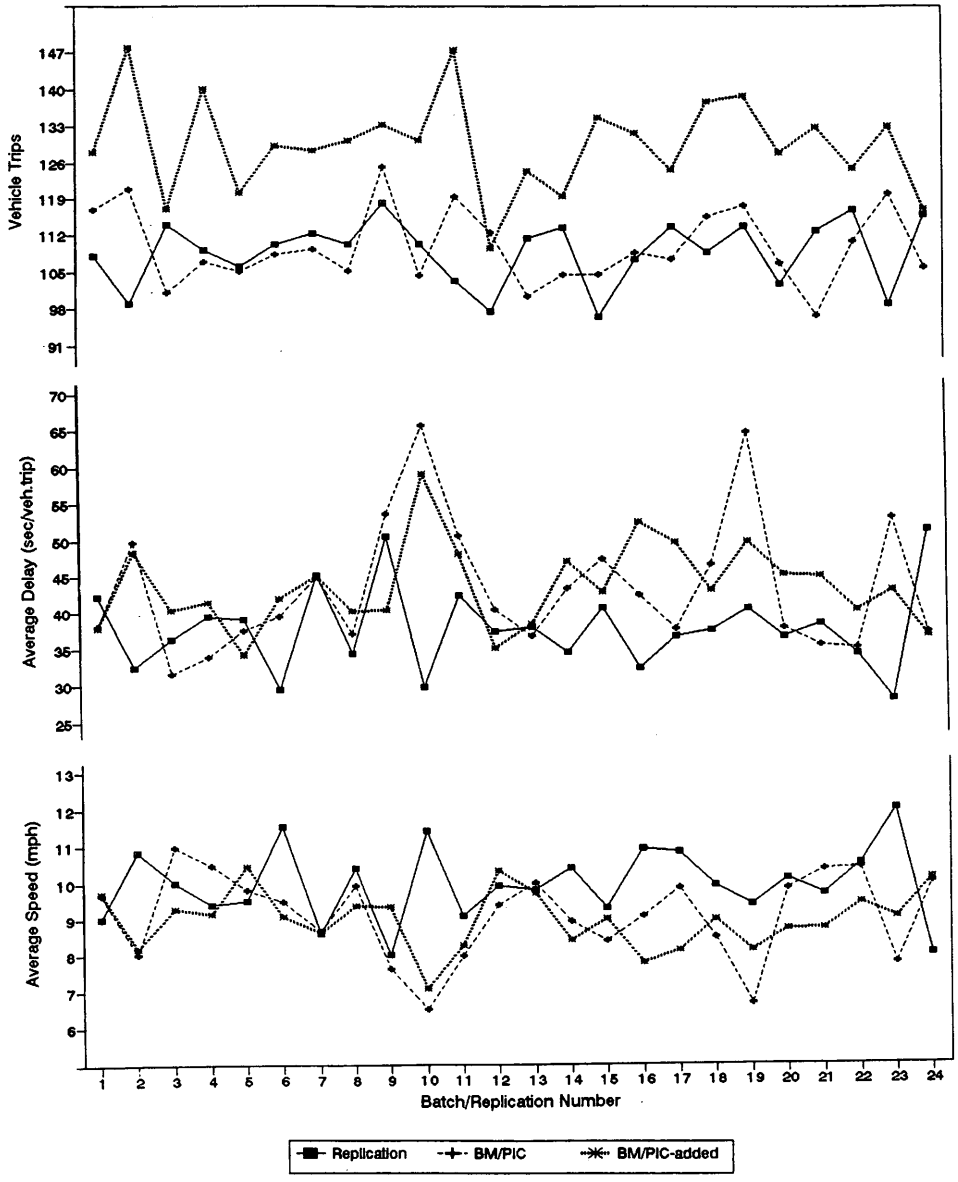


FIGURE 3 Comparison of delay, speed, and vehicle trips at link level.

depends on the size of the network, anticipated correlation among batches, traffic conditions, and other factors. It is suggested that the starting value for *T/X* be equal to an integer value of the cycle length, say, 10 average cycle lengths (if cycle length is 60 sec, starting *T/X* would be 10 min) or at least to the initialization time. Knowing the starting value of *T/X* and *X*, one can determine *T*. Run the simulation and compute MOEs for the *X* batches. Then compute the mean and variance for these batches.

Now, one needs to examine the correlation among batches using the procedures outlined before. If correlation is not significant, one may use this batch size to find the number of batches (shortening the batch size until the correlation becomes significant is optional at this point). To determine the number of batches, use the procedure described in the replication section. If the correlation is a problem, the batch length should be increased. Fishman suggested doubling the batch size to expeditiously arrive at an appropriate batch length (*8*). In the context of this study, it corresponds to doubling the batch length. However, one may consider increasing batch size at a slower rate, perhaps by 50 percent at a time.

Using the new batch size, run the simulation and examine the correlation again. If the correlation is no longer significant, then use the procedure described before to find the number of batches. Knowing the batch size and number of batch, one can easily find the length of the simulation run. However, if the correlation problem persists and the batch length becomes unreasonably long, consider using the replication method.

Choosing a Method

If tests did not reveal any serious correlation problems (with the BM/PIC results), the user is basically free to use either method.

However, users need to be aware of the following characteristics exhibited by output from each method:

- *The vehicle trips estimated by the replication method are consistently lower than the batch means.* The delays estimated by the replication method are consistently lower than the batch means (which is partly due to the lower estimate of vehicle trips).
- *The variability of the BM/PIC method output tends to be higher than that of the replication.* As indicated earlier, statistically there is no clear-cut advantage of using one method over the other for traffic simulation.

CONCLUSIONS AND RECOMMENDATIONS

This study explored the variability of NETSIM's output when batch means and replication methods are used. For purposes of statistical analysis, it is proposed to compute the MOEs for intermediate time intervals using the PIC procedure. The proposed PIC procedure is used with the batch means method.

Three MOEs were evaluated at three levels of aggregation; network, intersection, and link. The BM/PIC method resulted in average delays that are significantly higher than those of the replication method at network, intersection, and link levels. The opposite trend was true for the average speeds. Depending on which MOE is being examined and at what level of aggregation, it was apparent that the two methods can issue very different results.

The practical implications of variability were discussed, and approaches were proposed to help users utilize NETSIM and properly account for its variability. Suggestions also were made with regard to batch length, batch size, and number of runs.

The batch means (when the MOEs are computed by the PIC method) or replication methods should be used to compute confidence intervals for the MOEs. Depending on the objective of the study, the responses may need to be viewed at different levels of aggregation in order to properly assess the magnitude of the phenomena being studied. Not using a proper statistical procedure to make inferences about the results can lead to erroneous results. Output obtained directly from NETSIM for intermediate time intervals should not be used to build confidence intervals because the responses are autocorrelated, and as such complicated statistical procedures are necessary to construct proper confidence intervals.

Auto- and cross-correlation may exist for MOEs collected for a link for a very short period (e.g., 1 min). However, the correlations are likely to become weaker when data are collected at the intersection or network level and for a longer period. Cross-correlation and its magnitude should be examined, particularly when a significant negative correlation exists. Ignoring a positive correlation results in more conservative confidence interval.

Finally, Gafarian and Halati ruled out the use of a single long run (batch means method) to build confidence intervals on the MOEs because the direct output from NETSIM is autocorrelated and may be cross-correlated (*6*). They recommended using the replication method and including covariance of the variables to construct confidence intervals for the MOEs that are cross-correlated. When significant negative cross-correlation exists, the authors concur and recommend including a covariance term regardless of which method is used. However, when cross-correlation is not significant or is positive (which it is in most cases), not including the covariance term would result in a more conservative estimation of confidence intervals, which is desirable. Note that cross-correlation exists only when the MOE of interest is actually a ratio of the means. The BM/PIC method should be built into NETSIM's output structure. A comprehensive study dealing with the sensitivity and output variability of TRAF-NETSIM is recommended.

REFERENCES

1. *TRAF-NETSIM User's Manual.* FHWA, U.S. Department of Transportation, 1992.
2. *Highway Capacity Software.* FHWA, U.S. Department of Transportation, 1987.
3. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
4. Law, A. M., and W. D. Kelton. *Simulation Modeling and Analysis*, 2nd ed. McGraw-Hill, Inc., 1991.
5. Law, A. M. Confidence Intervals in Discrete Event Simulation: A Comparison of Replication and Batch Means. *Naval Research Logistics Quarterly*, Vol. 24, No. 4, Dec. 1977, pp. 667–678.
6. Gafarian, A. V., and A. Halati. Statistical Analysis of Output Ratios in Traffic Simulation. In *Transportation Research Record 1091*, TRB, National Research Council, Washington, D.C., 1986.
7. Chang, G. L., and A. Kanaan. Variability Assessment for TRAF-NETSIM. *Journal of Transportation Engineering*, Vol. 116, No. 5, Sept.–Oct. 1990, pp. 636–657.
8. Fishman, G. S. *Principles of Discrete Event Simulation.* Wiley and Sons, New York, 1978.
9. Rathi, A. K., and E. B. Lieberman. Effectiveness of Traffic Restraint for a Congested Urban Network: A Simulation Study. In *Transportation Research Record 1232*, TRB, National Research Council, Washington, D.C., 1989, pp. 95–102.
10. Davis, C. F., and T. A. Ryan. Comparison of NETSIM Results with Field Observations and Webster Predictions for Isolated Intersections. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 91–95.
11. Labrum, W. D. Application of NETSIM Computer Simulation Model to Traffic Control Problems. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 42–50.
12. Yauch, P. J., J. C. Gray, and W. A. Lewis. Using NETSIM To Evaluate the Effects of Drawbridge Operations on Adjacent Signalized Intersections. *ITE Journal*, Vol. 58, No. 5, May 1988, pp. 35–39.
13. Luedtke, P., S. Smith, H. Lieu, and A. Kanaan. Simulating DART's North Central Light Rail Line Using TRAF-NETSIM. Presented at ITE Meeting, the Hague, the Netherlands, Sept. 1993.
14. Hagerty, B. R., and T. L. Maleck. NETSIM: A User Perspective. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 40–42.
15. Maki, R. E., and D. R. Branch. Signal Timing Optimization and Evaluation: Route M-53, Macomb County. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 80–83.
16. Maki, R. E., and J. J. Saller. System Timing Optimization and Evaluation of US-12, Detroit. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 83–86.
17. Schafer, B. F. Comparison of Alternative Traffic Control Strategies at a T-Intersection. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 87–89.
18. Slee, K. L. Signal System Modernization and Time Optimization: Ludington Street, Escanaba. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, p. 86.
19. Bruce, E. L., and J. E. Hummer. Delay Alleviated by Left-Turn Bypass Lanes. In *Transportation Research Record 1299*, TRB, National Research Council, Washington, D.C., 1991, pp. 1–8.
20. Nemeth, Z. A., and J. R. Mekemson. Comparison of SOAP and NETSIM: Pretimed and Actuated Signal Control. In *Transportation Research Record 904*, TRB, National Research Council, Washington, D.C., 1983, pp. 84–89.

21. Sadegh, A., and A. E. Radwan. Comparative Assessment of 1985 HCM Delay Model. *Journal of Transportation Engineering*, Vol. 114, No. 2, March 1988, pp. 194–208.
22. Papacostas, C. S., and M. Willey. Use of TRAF-NETSIM to Estimate the Traffic Impacts of an Urban-Resort Area Development. In *Microcomputers in Transportation* (J. Chow, ed.), ASCE, New York, 1992.
23. Hurley, J. W., and A. E. Radwan. Traffic Flow Simulation: User Experience in Research. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 50–54.
24. Wong, J. Y. Comparing Capacities and Delays Estimated by Highway Capacity Software and TRAF-NETSIM to Field Results. *Compendium of Technical Papers*, ITE, 1990, pp. 224–227.
25. Radwan, A. E., and R. L. Hatton. Evaluation Tool of Urban Interchange Design and Operation. In *Transportation Research Record 1280*, TRB, National Research Council, Washington, D.C., 1990, pp. 148–155.
26. Kim, Y., and C. J. Messer. *Traffic Signal Timing Models for Oversaturated Signalized Interchanges*. Research Report 1148-2. Texas Transportation Institute, Texas A&M University, College Station, Jan. 1992.
27. Torres, J. F., A. Halati, and M. Danesh. *Impact of Lane Obstruction*, Vol. 2. Research Report. FHWA Contract DTFH61-84-C-00064. JFT Associates; FHWA, U.S. Department of Transportation, Feb. 1986.
28. Yagar, S., and E. R. Case. Summary Evaluation of UTCS-1/NETSIM in Toronto. In *Special Report 194: Application of Traffic Simulation Models*, TRB, National Research Council, Washington, D.C., 1981, pp. 95–99.
29. Rathi, A. K., and A. J. Santiago. Identical Traffic Stream in the TRAF-NETSIM Simulation Program. *Traffic Engineering and Control*, Vol. 31, No. 6, June 1990, pp. 351–355.
30. Rathi, A. K., and M. M. Venigalla. Variance Reduction Applied to Urban Network Traffic Simulation. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992, pp. 133–143.
31. Wong, S. Y. Capacity and Level of Service by Simulation: A Case Study of TRAF-NETSIM. *Proc., International Symposium on Highway Capacity*, Karlsruhe, Germany, July 1991, pp. 467–483.
32. Ott, L. *An Introduction to Statistical Methods and Data Analysis*, 3rd ed. PWS-Kent Publishing Company, Boston, Mass., 1988.
33. Bowerman, B., and L. O'Connell. *Time Series Forecasting Unified Concepts and Computer Implementation*. Duxbury Press, 1987.
34. Wong, S. Y. Discussion: Variance Reduction Applied to Urban Network Traffic Simulation. In *Transportation Research Record 1365*, TRB, National Research Council, Washington, D.C., 1992, pp. 143–146.

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

Calibration of INTRAS for Simulation of 30-sec Loop Detector Output

RUEY L. CHEU, WILFRED W. RECKER, AND STEPHEN G. RITCHIE

Since its inception in the early 1980s, the Integrated Traffic Simulation (INTRAS) model has been used in many studies involving freeway corridor traffic simulation. The model was originally calibrated with data collected in the 1970s in Los Angeles. In view of changing traffic conditions during the past decade, the validity of the parameter values as calibrated in the original setting is questionable. In several recent studies that used INTRAS as an evaluation tool, the model has been recalibrated with recent data. However, because of the different applications of the INTRAS model in these studies, the calibrations were made with output averaging at longer time intervals and for different output variables. To simulate traffic operation on Southern California freeways consistent with surveillance data currently being collected by the California Department of Transportation in traffic operations centers, INTRAS has been calibrated with respect to loop detector data at 30-sec intervals. The calibration process involved traffic during conditions with and without incidents, based on data collected along a 5-mi section of a major freeway in Orange County. Key parameters calibrated in this study include car-following sensitivity constants, minimum car-following distance, vehicle lengths, effective detector lengths, and the INTRAS "rubbernecking factor." The calibrated model has been used to simulate detector data for evaluating incident detection algorithms and for training artificial neural network models to detect freeway incidents.

INTRAS is a microscopic freeway traffic simulation model designed for freeway corridor traffic simulations. It is structured to facilitate evaluation of different incident detection algorithms and ramp metering strategies (*1*). During program development, detector output in INTRAS was calibrated with data collected in the 1970s at a freeway in Los Angeles (*2*). The validity of the parameter values as calibrated in the initial setting are questionable in view of changing traffic conditions and vehicle performance characteristics that have occurred during the past decade. In several recent studies using INTRAS as an evaluation tool (*3–5*), the model has been recalibrated with more recent data. However, these evaluations were made with model output averaging at longer time intervals, and with variables of different interest, than required for INTRAS to produce meaningful detector output consistent with typical 30-sec field data collected by traffic operations centers (TOCs) in Southern California. For this application, several parameters in the model should be calibrated in more detail.

In this paper, the authors describe the process of calibrating 30-sec station average volume and occupancy at loop detector stations located on a 5-mi section of the westbound SR-91 Riverside Freeway in Orange County, between the SR-57 and Interstate 5 freeways. The calibration process consisted of two parts. First, parameters related to car-following and loop detector operations were calibrated against incident-free data. Once the appropriate combination of the nonincident parameters had been found, incident-related parameters were calibrated against incident data sets. The objective of the calibration was to adjust the input parameters of INTRAS to produce volume and occupancy consistent with field data while maintaining the integrity of the engineering bases of the parameters, which are fundamental to the simulation model. It is hoped that the calibration process discussed in this paper may serve as a useful basis for future calibration with INTRAS or similar simulation models in order to produce detector output at relatively short intervals.

In the following section, the calibration procedure is described in more detail. Important deterministic inputs to the simulation model are featured including network coding, input volume, and free-flow speed. The next section discusses the calibration of nonincident parameters with an incident-free data set. This section also includes validation of the calibrated parameters with an independent data set. The calibration of incident-related parameters, in particular the "rubbernecking factor," is presented next. The results are discussed, and the calibration study is summarized.

Institute of Transportation Studies and Department of Civil Engineering, University of California, Irvine, Calif. 92717.

CALIBRATION PROCEDURE

For this calibration, 3 days of field data collected by Caltrans were used. The first day of data (Data Set 1) was collected on June 8, 1987 (Monday), from 7:00 a.m. to 7:00 p.m.; it contained incident-free loop data. This was the only data set available for calibration when the study began. Two subsequent data sets (Data Sets 2 and 3) were later obtained from Caltrans for the completion of this work. These data sets contained volume counts and occupancy values at each loop detector on freeway lanes aggregated at 30-sec intervals. Loop detector data at the same counting station were aggregated to station average values in this study.

The latest version of INTRAS source code was obtained from FHWA and made operational for a Sun Sparcstation. The section of program code that processes loop detector data was modified to simulate the actual data accumulation process occurring in the field. Deterministic input data such as network geometry and input volume were coded into the input file. Key parameters thought to influence vehicle travel were adjusted. At each adjustment, the volume and occupancy values at each INTRAS station were compared with the actual field data. The optimal combinations were ascertained after repeated trials.

The adjustment of parameters was divided into two parts. In the first part, parameters that influence vehicle movement during incident-free conditions were calibrated with the incident-free data set (i.e., Data Set 1). After necessary adjustments had been made, the adjusted parameters were validated against an incident-free portion of Data Set 3. In the second part of the calibration, the rubbernecking factor was calibrated against the incident portions of Data Sets 2 and 3.

SIMULATION INPUT

Network Coding

The entire study section of the SR-91 Freeway in the westbound direction was coded into an INTRAS input file, following the striping plan supplied by the local Caltrans district. The striping plan provided information on mileposts of on- and off-ramps, lane configurations, and detector locations. The coded nodes and links, along with a schematic of the study section, are displayed in Figure 1. The entire section was also videorecorded from a vehicle moving at a constant speed. The video recording helped to both validate the information provided by the striping plan and provide additional information on the location of exit signs, length of acceleration lanes, and other information input to the INTRAS simulation model.

There are eight detector stations in the study section, and each station has three inductance loops. The detector locations and their post miles are shown in Figure 1. These detectors are either circular loops 6 ft in diameter or squares 6 by 6 ft. Hence, an initial effective length of 6 ft was assumed for all the loops.

Traffic Volume

To ensure that nonincident parameters were calibrated for a variety of flow conditions, simulations were performed for three different flow levels (i.e., at low, moderate, and heavy flows). The field data was divided into 15-min intervals. For this study section, the daytime average 15-min volume ranges from 1,200 to 1,960 vehicles per hour per lane (vphpl). In Data Set 1, the 15-min periods beginning at 6:15 p.m., 8:45 a.m., and 4:45 p.m. were selected to represent typical low, moderate, and heavy flow levels in the day. The corresponding average volumes were 1,443, 1,681 and 1,858 vphpl, respectively. For incident simulation, the input volume was governed by the 15-min volume measured immediately before the occurrence of the incident.

In INTRAS, input volumes at all on-ramps must be specified in the data file. The off-ramp volumes are specified as the percentage of vehicles turning off from the freeway links. Within the study section, most of the detector stations are located near major cross streets, and there is always an on-ramp and an off-ramp between the two stations (see Figure 1). The data files provided by Caltrans did not include ramp volumes. Traffic volumes at the on-ramps were thus deduced from Traffic Accident Surveillance and Analysis System (TASAS) data base (*6*). The off-ramp volumes were computed using the principle of continuity of flow, based on the volume count obtained at the detector stations and the ramps.

The vehicle composition in the simulation runs was as follows:

Vehicle Type	*Percentage*
Low-performance passenger cars	46
High-performance passenger cars	47
Buses	0
Single-unit trucks	2
Truck trailers	5

A truck percentage of 6.8 to 7.1 percent in total bidirectional flow was estimated at Milepost 3.26 (near Harbor Boulevard) in 1982 (*7*). For the simulation, a value of 7 percent was used, assuming the percentage of truck in total traffic has remained relatively constant over the years. This figure was divided further into single-unit trucks and truck trailers. The distribution of truck type has been reported in a separate study (*8*) at several freeway-to-freeway connectors in the Los Angeles area. These figures were rounded to 2 percent for single-unit trucks and 5 percent for truck trailers. Since information on the percentage of buses was lacking, it was assumed that the proportion of buses was insignificant compared with the total volume, and a value of 0 percent was used. The remaining 93 percent of the traffic was arbitrarily split equally into the two passenger car categories.

Free-Flow Speed

The free-flow speed on the freeway was deduced from a speed-density relationship fitted to field data. The 30-sec volume and occupancy data from individual loop detectors were used to established the speed-density relationship. Assuming that there was a uniform

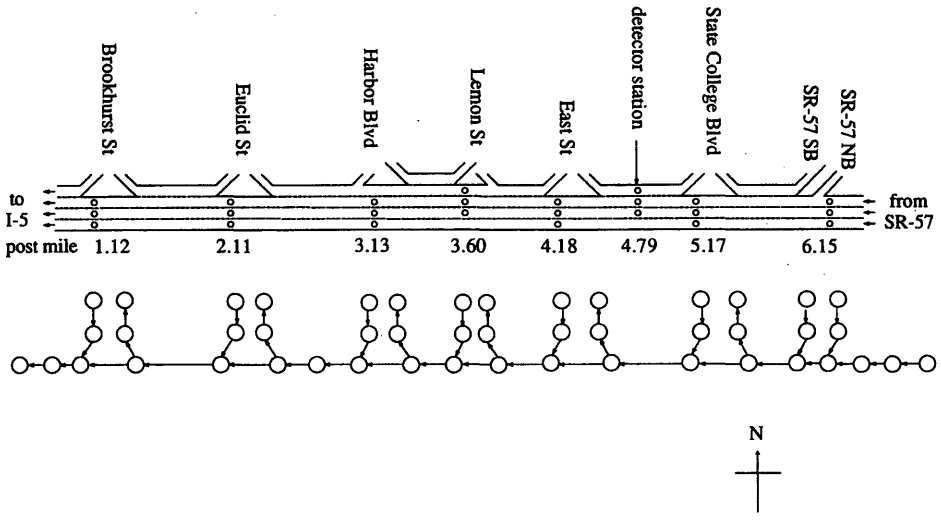


FIGURE 1 Schematic of freeway site and INTRAS nodes and links.

speed-density function represented at all locations within the study section, 100 loop-specific data points were randomly selected from Data Set 1. The average speed and density were deduced from volume and occupancy, using an average vehicle length of 19 ft computed from the aforementioned vehicle composition and the default vehicle lengths in INTRAS. An effective loop length of 6 ft was used in the computation. The data points followed the form of the Greenshields model. The free-flow speed, estimated from linear regression, was 81 mph. It should be noted that the lowest daytime freeway volume at the study section was about 1,200 vphpl (under incident-free conditions). At this volume the fitted Greenshields model gives a space-mean speed of 69 mph, which is close to the actual driving speed on the freeways in Southern California.

Output Data

The objective in calibrating INTRAS included making it produce output similar to the Caltrans 30-sec station average detector data. Correspondingly, the detector output interval was set at 30 sec. For incident-free conditions, a 15-min simulation was conducted for each volume level. The detector output at all stations during simulation runs at low, moderate, and heavy flow levels were combined and plotted in two graphs for evaluation: INTRAS output volume versus Caltrans field volume (volume plot) and INTRAS-measured occupancy versus Caltrans field occupancy (occupancy plot). For illustration, the volume and occupancy plots for INTRAS runs with default parameters are shown in Figures 2 and 3, respectively. The correlation coefficients (r-values) and slopes of fitted straight lines that pass through the origin derived from these plots were used as performance measures. In addition, the speed-density, volume-density, and volume-speed plots between field data and INTRAS output were compared.

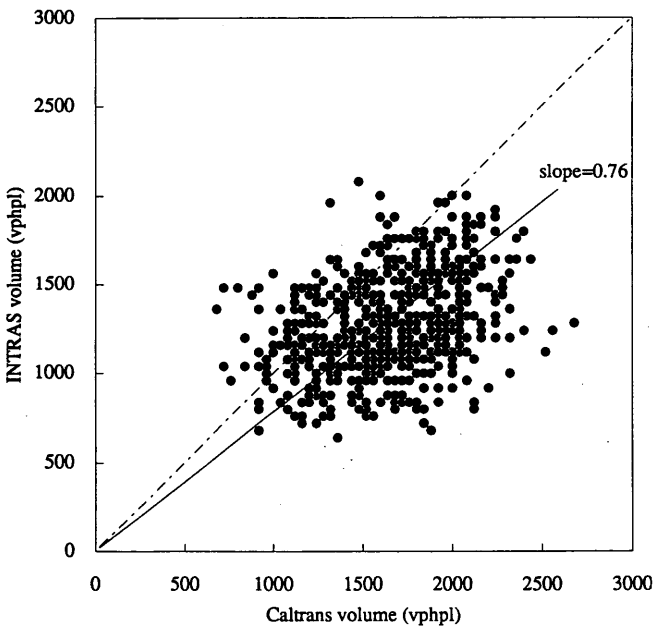


FIGURE 2 Volume plot of INTRAS runs with default parameters.

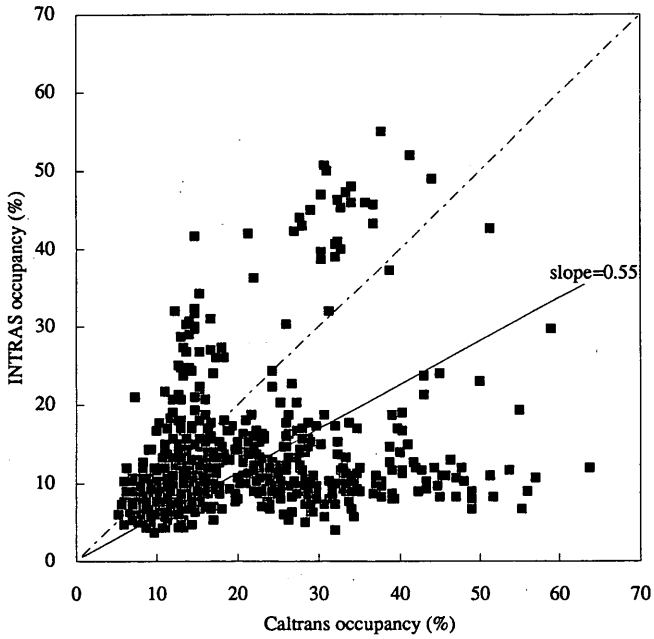


FIGURE 3 Occupancy plot of INTRAS runs with default parameters.

CALIBRATION OF NONINCIDENT PARAMETERS

Parameters thought to affect vehicle movement and detector operation during incident-free traffic simulation were calibrated as described in this section. Important car-following parameters were first identified: (a) car-following sensitivity constant, (b) minimum car-following distance, and (c) vehicle lengths. To reduce the number of possible combinations of parameter values in this calibration, these parameters were calibrated sequentially (i.e., at any time, only the value of one parameter was varied); while the optimum value of any particular parameter was being calibrated, the remaining parameters were treated as constants. For this part of the calibration, the parameters were adjusted to bring the volume and occupancy produced by INTRAS closer to the field data, with emphasis placed on volume count. Initially, the effective length of all loop detectors was set at 6 ft. Although the loop length does not affect volume count, it does affect the occupancy value. After the car-following sensitivity constants, minimum car-following distance and vehicle lengths were calibrated, the effective loop length could still be adjusted to fit occupancy close to actual data. The calibration of incident-free parameters made use of Data Set 1. The model with the calibrated parameters was then validated using the incident-free portion of Data Set 3.

Car-Following Sensitivity Constant

The movement of individual vehicles in INTRAS is governed by the car-following equation (*1,9*):

$$h(t) = L + m + kv(t) + bk[u(t) - v(t)]^2 \quad (1)$$

where

$h(t)$ = spacing headway at time t (ft);

L = length of lead vehicle (ft);
$v(t)$ = speed of following vehicle at time t (ft/sec);
$u(t)$ = speed of leading vehicle at time t (ft/sec);
k = sensitivity constant;
b = relative sensitivity constant, which is set to 1 if $u(t) > v(t)$ and 0 otherwise; and
m = 10 ft of minimum spacing.

The car-following sensitivity constant (k) in Equation 1 was first calibrated. The default values of k are from 10 to 19 at increments of 1, each corresponding to a particular type of driver. To check the sensitivity of INTRAS output with different k-values, simulation runs with three different series of k-values were carried out. The first series contained the default values of k. Series 2 and 3 consisted of k-values from 5 to 14 and 15 to 24, respectively.

A simulation run with each series of k-values was repeated three times, each with a different random number seed. For each set of simulation results with a random number seed, the slopes of the fitted straight lines and r-values of the volume and occupancy plots were computed. The average slopes and r-values obtained from the three random number seeds were examined.

With the reduction of sensitivity in Series 2, drivers followed each other at at higher speed, keeping the same distance. The average slope of the fitted lines and r-values in the volume plots remained approximately the same as those obtained with Series 1. The average slope of the fitted lines in the occupancy plots fell from 0.45 to 0.38. The low occupancy resulted in a very small average r-value of 0.04 in Series 2, not significantly different from 0 at $\alpha = 0.01$, based on Fisher's r-to-z transformation test (*10*).

The Series 3 of the k-values corresponded to more sensitive (or more conservative) car-following behavior. The resulting volume plots had an average slope of only 0.65 (compared with 0.79 obtained with Series 1), while that for the occupancy plots had a higher value of 0.74. Sensitive drivers tend to have greater following distance; therefore, the volume count was lower than the field measurement. They also tend to slow down more with increasing volume, giving rise to higher occupancy. This type of car-following behavior tends to produce unstable conditions that belong to the right-hand side of the volume-density plot.

From the results, it was obvious that efforts to increase the slope of the volume plot caused a decrease in the slope of the occupancy plot and vice versa. Since INTRAS input and output volume at the ramps were computed from actual data, the simulation model should produce a volume count close to the actual value. Among the three series, the default values gave the highest average r-value and slope in the volume plots. Consequently, the default k-values in Series 1 were retained as a good compromise among the three sets of attempted values.

Minimum Car-Following Distance

The m-value in Equation 1 sets the default minimum car-following distance at 10 ft. This distance may be too large, considering that drivers are observed to queue up bumper to bumper when traffic comes to a complete stop. An alternative of 0-ft minimum following distance was tested to study the effect of shortening this value. Simulation runs were performed with $m = 0$ combined with the three series of car-following sensitivity constants.

Series 1 and 3 had high slopes and r-values in volume plots but very low r-values in occupancy plots. The r-values of the occupancy plots for Series 1 and 3 were not significantly different from 0 at $\alpha = 0.01$. Series 2 had relatively higher r-values for occupancy plots to compromise for lower r-values in volume plots. None of the results here was superior to that obtained with Series 1 of sensitivity constants and with 10 ft of minimum following distance. The default minimum following distance of 10 ft therefore was kept unchanged.

Vehicle Lengths

The next step of the calibration involved changing the default vehicle lengths. The default vehicle lengths in INTRAS are as follows:

Vehicle Type	*Length (ft)*
Low-performance passenger cars	17
High-performance passenger cars	17
Buses	40
Single-unit trucks	23
Truck trailers	50

To test the sensitivity of volume and occupancy plots with different vehicle lengths, three series of simulations were carried out: with (a) the default vehicle lengths, (b) the default length plus 5 ft, and (c) the default length minus 5 ft.

By putting shorter vehicles into the simulation model, it is possible to increase the volume in INTRAS or to reduce the occupancy. The results showed that the average slope of the volume plots remained the same while there was a reduction in the average slope of occupancy plots from 0.45 to 0.37. Setting the vehicle lengths to 5 ft longer than the default values brought the slope of the occupancy plot closer to unity. However, data points in the volume plot scattered in a circular region rather than showing a trend of a straight band. Using the default vehicle lengths gave a better match between the simulation results and field data. The vehicle lengths therefore were not adjusted.

Effective Length of Loop Detectors

With the default car-following constants, in order to keep the same INTRAS detector volume but increase the occupancy value, it was necessary to increase the effective length of the detectors. The lane width of the freeway is 12 ft. Assuming 1 ft of minimum clearance on both sides of the lane striping, the maximum size of a square or circular loop is 10 ft. INTRAS simulations at low, moderate, and high volume were made with all the detector lengths set at 10 ft. This brought the average slope of the fitted straight lines in the occupancy plots from 0.45 to 0.49. Since this step involved changing the vertical values of data points in the occupancy plots, the r-values remained the same.

Although the physical size of the loop is 6 ft and its effective zone may be slightly larger, practically it should not be as large as 10 ft. However, to bring the INTRAS occupancy values closer to the field data without changing the form of the car-following model, it was decided to make the numerical adjustment here.

Adjustment of Free-Flow Speed

The free-flow speed of 81 mph on the freeway links was estimated using an effective loop length of 6 ft. Since the effective loop length was increased to 10 ft, it was necessary to reestimate the free-flow

speed using the new loop size. The same 100 data points (lane-specific volume and occupancy) used earlier were used to recompute the density and speed, assuming a 10-ft loop length. The fitted Greenshields free-flow speed was 95 mph. The speed of 95 mph was then set for all the freeway links in another set of INTRAS runs, and the results were compared with those obtained with a free-flow speed of 81 mph and loop length of 10 ft.

The average slope of the volume plots, with a free-flow speed of 95 mph, was 0.77. This was closer to the 0.79 obtained with the free-flow speed of 81 mph. But the data points with the higher free-flow speed were more scattered, as reflected in the reduction in *r*-value, from 0.34 to 0.28. The occupancy plots had an average slope of 0.61, which was an improvement on the 0.49 obtained with the free-flow speed of 81 mph. The average *r*-value of the occupancy plot increased from 0.15 to 0.31.

None of the free-flow speeds was distinctly superior. The free-flow speed of 81 mph was closer to the actual driving speed on the freeway and gave better *r*-values in the volume plots. The free-flow speed of 95 mph produced better matched data in the occupancy plots. For free-flowing traffic, it is more important to match volume than occupancy, especially when the actual field volume was used as part of the simulation input. The free-flow speed of 81 mph was thus retained.

Validation of Calibrated Parameters for Incident-Free Conditions

The calibration process changed the freeway free-flow speed to 81 mph and the effective loop length to 10 ft. Another set of detector data (Data Set 3), collected on December 12, 1990 (Wednesday), from 5:00 a.m. to 10:45 p.m. was acquired to validate these adjusted parameters. After excluding time segments encompassing incidents and the period after which detector data appeared to be influenced by them, the remaining segments were divided into 15-min intervals. Three 15-min incident-free periods starting at 10:30 a.m., 5:30 p.m., and 7:00 p.m. were selected to represent moderate-, high-, and low-volume conditions. The average station volumes at these periods were 1,527, 1,771, and 1,264 vphpl, respectively.

This validation data set gave average slopes of 0.87 and 0.54 for volume and occupancy plots, respectively. The corresponding average *r*-values were 0.46 and 0.39. These values were higher than their respective values obtained with Data Set 1. The volume-density curve of field data and that obtained by INTRAS simulations were inspected, and they matched very closely. The calibrated parameters were thus retained.

CALIBRATION OF INTRAS RUBBERNECKING FACTOR

After the nonincident parameters were calibrated and validated, the rubbernecking factor in INTRAS was calibrated with incident Data Sets 2 and 3, collected on February 4, 1991, and December 12, 1990, respectively.

Calibration of Rubbernecking Factor with Data Set 2

Data Set 2 included an incident that occurred at 6:21 a.m. between Harbor Boulevard and Euclid Street, and lasted for 810 sec. This incident resulted in Lane 2 being blocked by a single vehicle, but the exact location within the 1-mil section between the two cross streets was unknown. Repeated simulation runs were performed, placing the incident at different locations between the two stations. In each simulation run, the 15-min actual average volume before the occurrence of the incident was used as the input volume in INTRAS. Five minutes of free-flowing traffic were simulated before the incident. By comparing the time at which a sharp increase occurred in the occupancy of the upstream station at the onset of the incident for the different runs against the trend of increments in actual data, the location of the incident was deduced to be 300 ft downstream from the on-ramp at Harbor Boulevard.

The INTRAS user's manual (*9*) recommends that a rubbernecking factor of 10 be used with each incident simulation. In the calibration, values of 5, 10, 20, 30, and 40 were considered in five simulation runs. Initially, the rubbernecking effect was set at the incident location for the rest of the lanes that were not blocked as well as for all the lanes immediately downstream of the incident as recommended. A second set of five simulation runs was performed with these factors, without the rubbernecking effect downstream of the incidents. Another run with no rubbernecking was made, bringing the total number of simulations to 11.

For each simulation run, the upstream and downstream occupancy and volume at the incident location were plotted against time. Among the four variables, the upstream occupancy was found to be most sensitive at onset, during, and at termination of the incident. Since the upstream occupancy during and after the incident was affected by the rubbernecking factor, the root mean square (*RMS*) error between INTRAS occupancy and actual occupancy during this period was used as a performance measure. The 11 simulation runs were repeated three times, each with a different random number seed. The average *RMS* error of the runs without rubbernecking was computed to be 9.69 percent, and the *RMS* errors from all the remaining runs with rubbernecking were all greater than 18 percent. From the results, apparently no rubbernecking factor is necessary for incident specification. The fluctuation of upstream occupancy with time for the field data and from the simulation run with the default random number seed is shown in Figure 4. INTRAS was capable of generating occupancy values comparable to the field data.

To test (a) the stability of INTRAS in producing occupancy values that were closely matched with the field data, and (b) the hypothesis that the data points produced by INTRAS, such as those in Figure 4, were not significantly different from the actual values collected in the field, the following experiment was conducted. First, 30 simulation runs (without the rubbernecking effect) were performed with different random number seeds. At the end of each 30-sec interval, the occupancy values extracted from the these simulations were taken to compute the mean and standard deviation of the simulated occupancy of that interval. Assuming that the simulated occupancy value at a particular time interval fluctuates about the sample mean and follows a normal distribution, the 95 percent confidence interval was constructed. Successive confidence intervals were plotted against simulation time to form a confidence envelope and superimposed with actual data obtained from the field. All the actual data points fell within the 95 percent confidence envelope, indicating that the actual data points were not significantly different from the values generated by INTRAS.

Calibration of Rubbernecking Factor with Data Set 3

Data Set 3 had an incident that occurred at 10:56 a.m. between Euclid Street and Brookhurst Street. This incident, caused by a six-vehicle collision, resulted in Lane 1 (the leftmost lane) being

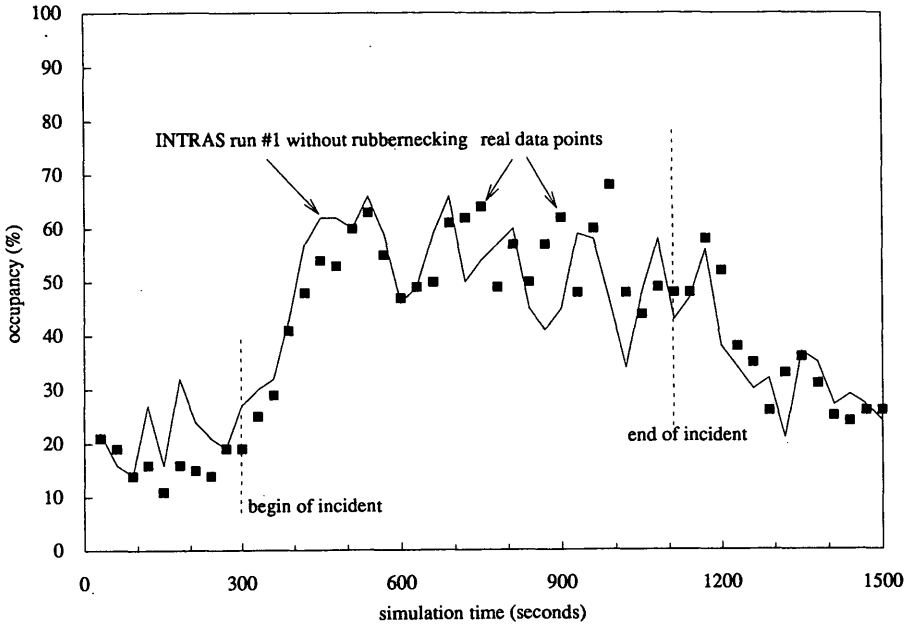


FIGURE 4 Upstream occupancy without rubbernecking factor for incident in Data Set 2.

blocked from 10:56 to 11:33 a.m. From 11:33 to 11:36 a.m., the entire freeway section was closed for the incident management team to move the vehicles involved in the collision from the left lane to the shoulder, after which all the lanes were opened to traffic.

The following 1-h incident scenario was simulated in INTRAS: 5 min of incident-free traffic, incident with Lane 1 blocked for 37 min, followed by a full blockage of 4 min and 14 min of clearance time after the removal of the blockage. Since this incident had been split into two parts and INTRAS permits only two blockage specifications per simulation run, no rubbernecking factor was assigned downstream of the incident. The length of incident was set at 140 ft (for six vehicles) according to the guideline provided in the INTRAS user's manual. By means of trial and error, the upstream end of the incident was placed 500 ft downstream from the on-ramp at Euclid Street.

The same rubbernecking factors used for the Data Set 2 incident were tested here. The average RMS error of upstream occupancy was found to be 14.32 percent without rubbernecking, compared with at least 34 percent with the rubbernecking factors. Similar to the earlier finding, no rubbernecking was required to produce a closer match between INTRAS output and field data. The minimum average RMS error of 14.32 percent was of higher magnitude than the 9.69 percent found in the Data Set 2 incident. The larger difference is caused by the magnitude of the random fluctuation of the field data during the incident as well as by consistent bias in INTRAS output after the incident. For illustrative purposes, the upstream occupancies from INTRAS Run 1 with no rubbernecking and with the default random number seed are plotted against field data in Figure 5.

Simulation runs without the rubbernecking effect were repeated 30 times, each with a different random number seed to construct the 95 percent confidence envelope. Twenty-three of 120 actual data points fell outside the 95 percent confidence envelope. It should be noted that 17 of the 23 outliers occur 4 min after the removal of the incident. INTRAS is good in simulating queueing situations during incidents, but it may underestimate the occupancy during free-flow conditions as well as the recovery periods after incidents. These phenomena were reflected in all the occupancy plots (see Figure 3). Except for this apparent shortcoming, INTRAS is capable of simulating incidents and producing reasonably accurate detector output.

DISCUSSION OF RESULTS

The car-following equation in Equation 1, if used to derive a macroscopic traffic stream model, results in a speed-density function that slopes downward. This function corresponds to the high-density region (rightside) of the commonly used bell-shaped volume-density curve. This is understandable because car-following occurs only when traffic density has reached a certain level. The left side of the volume-density curve is constrained by the free-flow speed imposed on the freeway links.

In general, the car-following equation gives satisfactory results. The only apparent drawback is that it fails to simultaneously produce volume and occupancy high enough to match with the actual data collected in the study section. There may be some combinations of input parameters that can give better results but have not been tested in this limited study. Alternatively, there may be different forms of car-following models that can better represent the behavior of drivers in the study area. Since human driving behavior is complex, one should not expect that the same form of equation would apply to all drivers. One suggestion for improvement may be to replace the car-following model with a series of artificial neural network models, one for each type of driver. The primary advantage of neural network models is that no car-following rules and associated parameter values need to be specified explicitly. Such a neural network model could receive input such as speed, relative speed, vehicle spacing, and vehicle length and produce output indicating acceleration and lane changing responses. An initial attempt to use an artificial neural network model to mimic elements of driver behavior has been performed with laboratory-simulated data in a

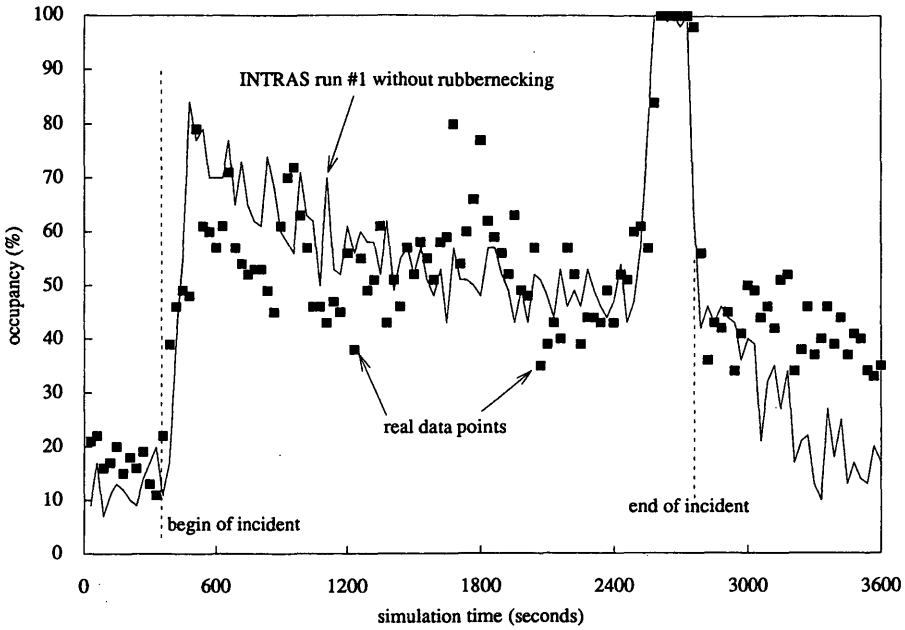


FIGURE 5 Upstream occupancy without rubbernecking factor for incident in Data Set 3.

simplified driving environment (*11*), and the result has demonstrated the potential of such an application.

Two principal limitations in INTRAS could affect data generation for incidents:

1. INTRAS does not permit placement of loop detectors at on- and off-ramps. If such provision existed, one may be able to use ramp volume and occupancy to simulate real-time ramp metering as well as to provide additional inputs to incident detection algorithms.
2. INTRAS permits coding of the rubbernecking factor only in the three rightmost lanes on a freeway (excluding any auxiliary lane). This may not affect simulation runs, as the calibration results have shown that it is not necessary to assign a rubbernecking factor at any incident location. However, if the calibration results at other freeway sites necessitate the use of a rubbernecking factor and the freeway section has more than three lanes, appropriate subroutines in the INTRAS program code would have to be modified.

Despite these limitations, INTRAS is still the most widely used and most readily available microscopic freeway simulation model that can produce detector data at 30-sec intervals. The calibration procedure described in this paper should also apply to the successor of INTRAS, namely, the FRESIM model currently under development, which is believed to have the same fundamental structure as INTRAS. The INTRAS model, with the calibrated parameters, has been used to simulate hundreds of incidents in the study section. The 30-sec station average volume and occupancy output has been used to train artificial neural network models for detection of incidents on the freeway(*12*).

SUMMARY

On the basis of the data sets available, the following input values were used in the INTRAS data file to produce simulation output that closely matched the calibration data:

1. The free-flow speed of the freeway study section was estimated at 81 mph.
2. The default car-following constants were the best among the attempted values in describing driver behavior. With these car-following constants, INTRAS was capable of "moving" vehicles at volumes very close to the actual count made on the freeway. However, the loop detector occupancies were always lower than the values collected in the field. To artificially increase occupancy value, it was necessary to increase the effective loop length to 10 ft.
3. To simulate traffic operation during the incidents considered, it was not necessary to assign any rubbernecking factor in the incident specification. Putting only the actual lane blockage at the incident location on the freeway was enough to produce an occupancy pattern closely resembling the actual traffic operations during incidents.

ACKNOWLEDGMENTS

The research reported in this paper was supported by the U.S. Department of Transportation/University of California Transportation Center, the Caltrans PATH (Partners for Advanced Transit and Highways) program, and the National Science Foundation.

The authors would like to thank Stephen Cohen of FHWA for providing them with the working version of INTRAS and Susan King from the District 7 Traffic Operations Center of Caltrans for her assistance in providing field data.

REFERENCES

1. Wicks, D. A., and E. B. Lieberman. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model, Vol. 1: Program Design, Parameter Calibration and Freeway Dynamics Component Development.* Report FHWA/RD-80/106. FHWA, U.S. Department of Transportation, 1980.

2. Goldblatt, R. B. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model, Vol. 3: Validation and Application*. Report FHWA/RD-80/108. FHWA, U.S. Department of Transportation, 1980.
3. Hamad, A. I. *Evaluation of Ramp Metering Strategies at Local On-Ramps and Freeway-to-Freeway Interchanges Using Computer Simulation Modelling Approach*. Ph.D. dissertation. Michigan State University, East Lansing, 1987.
4. Skabardonis, A., M. Cassidy, A. D. May, and S. Cohen. Application of Simulation to Evaluate the Operation of Major Freeway Weaving Sections. In *Transportation Research Record 1225*, TRB, National Research Council, Washington, D.C., 1989, pp. 91–98.
5. Fazio, J. *Modelling Safety and Traffic Operations at Freeway Weaving Sections*. Ph.D. dissertation. University of Illinois at Chicago, 1990.
6. *Manual of Traffic Accident Surveillance and Analysis System*. Office of Traffic Engineering, California Department of Transportation, Sacramento, 1987.
7. *1983 Annual Average Daily Truck Traffic on the California State Highway System*. Office of Traffic Engineering, California Department of Transportation, Sacramento, 1984.
8. Leonard, J. D., and W. W. Recker. *Analysis of Large Truck Crashes on Freeway-to-Freeway Connectors*. Draft Final Report. Contract RTA-556916. Institute of Transportation Studies, University of California, Irvine, 1991.
9. Wicks, D. A., and B. J. Andrews. *Development and Testing of INTRAS, a Microscopic Freeway Simulation Model, Vol. 2: User's Manual*. Report FHWA/RD-80/107. FHA, U.S. Department of Transportation, 1980.
10. Hays, W. L. *Statistics for the Social Science*, 2nd ed. Holt Rinehart and Winston, Inc., 1973.
11. Shepanski, J. F., and S. A. Macy. Teaching Artificial Neural Systems to Drive: Manual Training Techniques for Automated Systems. *Proc. 1987 Neural Information Processing Systems Conference*, American Institute of Physics, New York, 1987, pp. 693–700.
12. Ritchie, S. G., and R. L. Cheu. Freeway Incident Detection Using Artificial Neural Networks. *Transportation Research*, Vol. 1C No. 3, Pergamon Press, 1993, pp. 203–217.

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