TRANSPORTATION RESEARCH

RECORD

No. 1358

Highway Operations, Capacity, and Traffic Control

Vehicle Routing, Traveler ADIS, Network Modeling, and Advanced Control Systems

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1992

Transportation Research Record 1358
Price: $22.00

Subscriber Category
IVA highway operations, capacity, and traffic control

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

Printed in the United States of America

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

Vehicle routing, traveler ADIS, network modeling, and advanced control systems.
p. cm.—(Transportation research record, ISSN 0361-1981 ; no. 1358)
"A peer-reviewed publication of the Transportation Research Board."
Includes bibliographical references.
ISBN 0-309-05224-6
1. Motor vehicles—Automatic control. 2. Traffic engineering. 3. Transportation, Automotive—Communication systems. I. National Research Council (U.S.). Transportation Research Board. II. Series: Transportation research record ; 1358.
TL240.V4 1992
388.3' 12—dc20 92-21411
CIP

Sponsorship of Transportation Research Record 1358

GROUP 1—TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Chairman: Ronald F. Kirby, Metro Washington Council of Governments

Transportation Forecasting, Data, and Economics Section
Chairman: Mary Lynn Tischer, Virginia Department of Transportation

Committee on Transportation Supply Analysis
Chairman, Hani S. Mahmassani, University of Texas at Austin
David E. Boyce, Yupo Chan, Carlos F. Daganzo, Mark S. Daskin, Michel Gendreau, Theodore S. Glickman, Ali E. Haghani, Randolf W. Hall, Rudi Hamerslag, Bruce N. Janson, Haris N. Koutsopoulos, Chryssi Malandraki, Eric J. Miller, Anna Negurney, Earl R. Ruiter, K. Nabil A. Safwat, Mark A. Turnquist

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

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

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

Committee on Freeway Operations
Chairman: Ronald C. Sonntag, Wisconsin Department of Transportation
Secretary: Jeffrey A. Lindley, Federal Highway Administration
B. Beukers, Peter M. Briglia, Jr., Donald G. Capelle, Glen C. Carlson, Robert F. Dale, Conrad L. Dudek, Walter M. Dunn, Jr., Jack L. Kay, Peter R. Korpal, Walter H. Kraft, Steven Z. Levine, Louis E. Lipp, Robert E. Maki, Joseph M. McDermott, Nancy L. Nihan, Colin A. Rayman, James R. Robinson, David H. Roper, James F. Shea, Henry B. Wall III, Thomas C. Werner, Sam Yagar, Michael J. Zezeski, Philip Zove

Committee on Expert Systems
Chairman: Michael J. Demestsky, University of Virginia
Hamid Aougab, Anselmo Osvaldo Braun, Edmond Chin-Ping Chang, Louis F. Cohn, David J. Elton, Ardeshir Faghri, Jon D. Fricker, Geoffrey D. Gosling, Jerry J. Hajek, Roswell A. Harris, Chris T. Hendrickson, Andrew B. Levy, Hani S. Mahmassani, Olin W. Mintzer, Prahlad D. Pant, Ajay K. Rathi, Stephen G. Ritchie, K. Nabil A. Safwat, Gary S. Spring, John R. Stone, James A. Wentworth, Marcus Ramsay Wigan, Robert A. Wolfe, Charles A. Wright

Task Force on Advanced Vehicle and Highway Technologies
Chairman: Daniel Brand, Charles River Associates Inc.
David E. Boyce, Richard P. Braun, Peter Davies, Shaun S. Devlin, Robert D. Ervin, Michael M. Finkelstein, Robert L. French, Franz K. Gimmler, Thomas A. Griebel, William J. Harris, Jr., Bernard F. Heinrich, Jack L. Kay, Bill M. McCall, Lawrence G. O'Connell, Donald E. Orne, Robert E. Parsons, Stephen Edwin Rowe, Lyle Saxton, Joseph L. Schofer, William M. Spreitzer, Burton W. Stephens, Joseph M. Sussman, John Vostrez, David K. Willis

James A. Scott and Richard A. Cunard, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1991.

Transportation Research Record 1358

Contents

Foreword

The papers in this Record cover a wide range of transportation topics. Areas of interest include transportation network analysis, the relation of street network geometrics and control features to quality of traffic service, transit network structure for information service of direct and connecting trips, and real-time network traffic simulation. Other areas are intelligent vehicle-highway systems and their standards, and system technologies, driver information systems, vehicle control systems, system safety, and vehicle direction. Expert system applications are related to congestion management and application of artificial neural networks in transportation engineering, and validation procedures are given for a prototype expert system for traffic control through highway work zones.

Efficient Search Algorithms for Route Information Services of Direct and Connecting Transit Trips

ANTHONY F. HAN AND CHIEN-HUA HWANG

Easy-to-access and efficient route information service is essential to encourage the ridership of a transit system. Given the origin (bus) stop and destination stop of a transit trip, it is not difficult to find the direct bus lines connecting the given origin-destination stops. However, when the transit network structure is involved with many transfer trips or characterized with many overlapping bus lines, as are many transit systems in major cities outside North America, the problem of finding bus lines connecting through transfer points becomes very complicated. A simple yet efficient algorithm to find direct lines and a search algorithm using the hash-table data structure techniques for quickly finding routing information for connecting trips with one transfer are presented. The search algorithms have been successfully implemented on a microcomputer-based transit information service system in Taipei, Taiwan, since January 1991.

Easy-to-access network information is essential to encourage the ridership of a transit system. The more people who are aware of the transit network structure and routing information, the more people who can use the transit system. For this reason, most transit authorities in big cities worldwide are providing some form of information service to assist passengers in determining the best use of public transport for specific trips (*1–4*). Fruin described in detail the type of passenger information including visual and oral communication, distributed information, and automatic passenger interactive means (*5*).

The importance of using computerized management information technologies to improve the productivity and performance of a transit system has long been recognized. Recently, microcomputer-based systems have been widely applied to enhance the transit planning and operational capabilities in areas such as traffic and data management (*6*), fleet management (*7*), maintenance management (*8*), and performance monitoring (*9,10*). Cutler studied the impact of information technologies on the efficiency of telephone information services provided by 15 transit authorities in the United States (*11*). The success of such a telephone on-line transit routing information system relies on, among other factors, efficient search algorithms that can provide quick responses and accurate routing information. However, literature focused on the routing information search for direct and connecting transit trips appears to be rare. Vanigrok developed an algorithm for selecting desirable public transport connection from the time tables (*3*).

In this paper, we are concerned with the problem of finding all the bus lines connecting two bus stops with or without transfers in a transit network. For simplicity and pragmatic considerations, the following discussions will be limited to transit trips involving no more than one transfer. Given the origin (O) bus stop and the destination (D) stop of an inquired trip, the problem of finding the bus lines that directly connect the stops is not difficult if such lines exist. However, the problem becomes difficult when the bus transfer is inevitable. In particular, when the transit network is characterized with overlapping bus lines, as many transit systems in major cities outside North America are, the problem can be very complicated. For example, for the origin stop A and the destination stop B, shown in Figure 1, no direct bus lines connect A and B; the rider has seven transfer options: from Line 1 to Line 2, transferring at one of the stops $\{a, b, c, d\}$, or from Line 3 to Line 4, transferring at one of the stops $\{e, f, g\}$. Note that the situation is even more complicated when any one of the single lines (1, 2, 3, or 4) can be a set of overlapping bus lines running on the same street segment. As to the transit network characterized with heavily overlapping bus routes, Han and Wilson proposed a heuristic method for optimal bus allocation for the Cairo transit system (*12*), Han assessed the transfer penalty to bus riders in Taipei (*13*), and Kwan analyzed how to coordinate the joint headway for overlapping bus routes in the United Kingdom (*14*).

SEARCH ALGORITHM FOR DIRECT CONNECTING LINES

To provide accurate routing information, the system must distinguish the bus lines with different attributes such as direction and zonal or express service. As a result, almost every bus stop is covered by more than one bus line, say, 32 eastbound and 32 westbound. Consequently, when the transit network is characterized with heavily overlapping bus routes, finding direct connecting bus lines is much more complicated than expected. For example, in the Taipei Transit System there are approximately 490 bus lines and 1,200 bus stops, and more than 50 bus stops are served by more than 30 (overlapping) bus lines; the maximal overlapping stop is the Taipei Train Station, which is covered by 63 bus lines (*15*).

Given two bus stops, A and B, we now present a search algorithm to find all bus lines connecting directly from A to

Department of Transportation Engineering and Management, National Chiao Tung University, Hsinchu 30049, Taiwan, Republic of China.

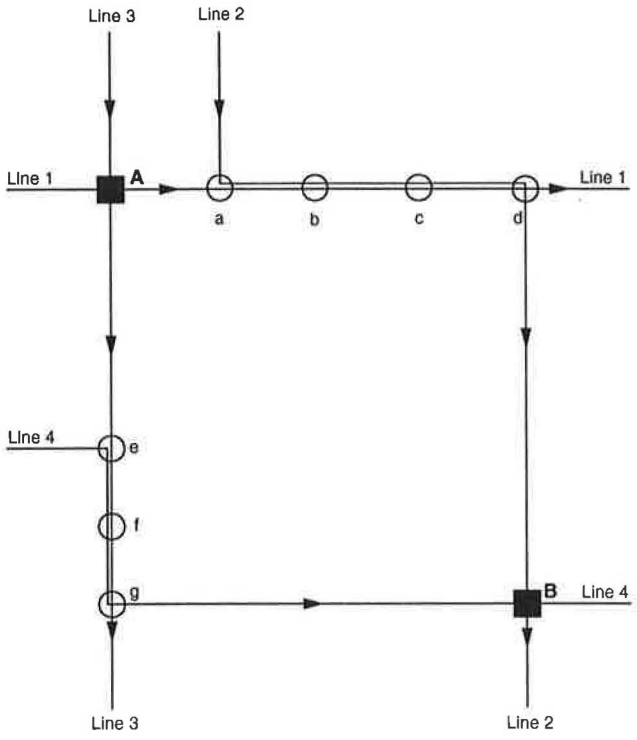

FIGURE 1 Connecting trip with multiple transfer stops.

B. The algorithm has two phases: connection search and direction check. The algorithm for direct transit trips is summarized in the following.

Algorithm A1: Connection Search Phase

1. Initialization: $S = \phi$, S is the set of direct connecting lines.
2. For both stops, A and B, list and sort all bus lines passing each of them. We get two arrays each presenting the set of bus lines passing the corresponding bus stop, and the numbers (codes) in both arrays are in ascending order.
3. Pick up from each of the two arrays the first number and do a pairwise comparison.
   - If the two numbers are the same—say, they are both N_1—then we have found a direct connecting line. Let $S = S \cup \{N_1\}$, and remove N_1 from both arrays.
   - If the two numbers are not the same, remove the smaller one from its corresponding array.
4. Repeat Step 3 until both arrays are exhausted. If $S = \phi$, stop—there are no direct connecting lines. Otherwise, go to Step 5 in the next phase.

Algorithm A1: Direction Check Phase

5. For each element in S, check if Stop A is upstream of Stop B. If not, remove it from S; otherwise, continue.
6. The solution is found as S. If $S = \phi$, there are no direct lines connecting from A to B.

Algorithm $A1$ is simple and efficient. The computational time of the sort operation in Step 2 is to the order of $L(\log L)$, where L is the number of overlapping bus lines passing a bus stop. In practice, the set of bus lines passing A or B can be sorted and filed before the on-line implementation of the algorithm. Therefore, the algorithm $A1$ can be easily implemented as a linear-time algorithm with order of $O(L)$, and it can serve as an efficient building block for other related search algorithms.

CONNECTION TRIPS WITH ONE TRANSIT TRANSFER

When no direct lines connect two bus stops, transit transfer activities are inevitable. For simplicity, connection trips with two or more transfers are not considered in this paper. We are concerned with finding all possible connection lines, with one transit transfer between two bus stops. This is more difficult than finding direct connecting lines, mainly because the transfer stop is not necessarily unique and the multiple transfer stops are hard to identify. A scenario of a connecting trip with seven possible transfer stops is shown in Figure 1.

Finding the transfer points between two stops A and B is complex because every downstream stop of A along any bus line passing A as well as every upstream stop of B along any bus line passing B is a potential transfer point and must be traced and checked. Figure 2 illustrates such a situation, in which every one of the stops of a_is and b_js $i, j = 1, 2, \ldots, 7$, is a potential transfer stop. Consequently if one uses the complete-enumeration type of pairwise comparison to search for the transfer stops, it takes approximately the order of L^2S^2 iterations simply to find the potential transfer stops, where L is the number of overlapping bus lines passing one stop and S is the average number of bus stops per line.

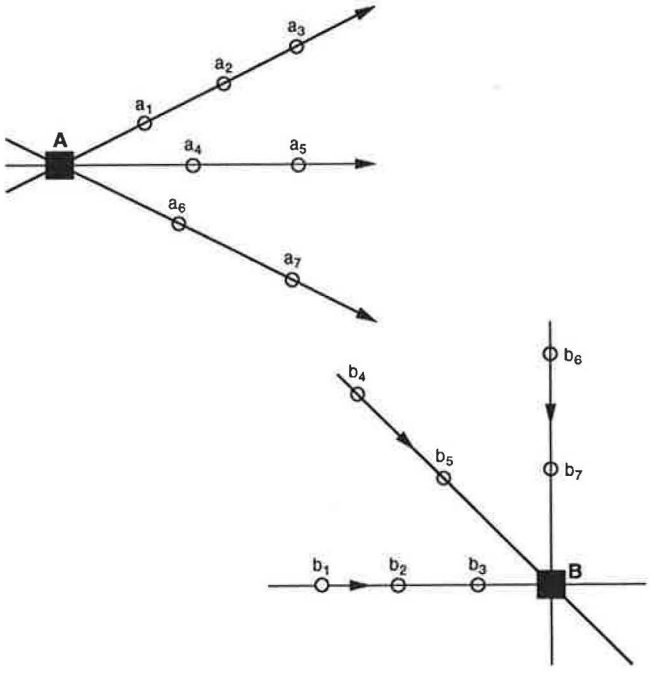

FIGURE 2 Potential transfer points of connecting trip.

The aforementioned complete-enumeration algorithm is a polynomial-time algorithm of $O(L^2S^2)$. It is still efficient in terms of combinational optimization. However, for practical on-line information service applications, the response time of such an algorithm may be too slow to accept. According to our implementation experience in Taipei, this algorithm when implemented by a PC 286/16 AT microcomputer took 20 to 90 sec to find the routing information for connecting trips. Such a performance was considered unacceptable because nobody is willing to hold the phone and wait for 20 sec or longer. This also led to the development of a much more efficient search algorithm.

SEARCH ALGORITHM USING HASH-TABLE TECHNIQUES

The full routing information of connection trips from A to B with one transit transfer can be considered as path information represented as $A-X-B$, where X is the transfer point. However, from a microcomputer it is difficult to find such a piece of full information at once. Our solution is to decompose the problem into subproblems and then combine the partial information obtained to form the final solution. Generally speaking, four subproblems are involved here:

- ($S1$)—Identification of all possible transfer points, the Xs;
- ($S2$)—Identification of bus lines connecting from A to B, the $X-B$ part of routing information;
- ($S3$)—Identification of bus lines connection A to X, the $A-X$ part of routing information; and
- ($S4$)—Combination of $A-X$ and $X-B$ to form the full information of $A-X-B$.

The subproblems can be solved efficiently by the use of hash-table, or hashing, techniques. Hashing is one of the data structure techniques commonly used to handle symbol tables in computer science (*16*). Considering the limited RAM storage space on a microcomputer, we used a single hash table of 32 × 32 bytes to keep track of all the potential transfer stops. The hash table is partitioned into 32 buckets (in bytes) and each bucket is capable of holding 256 records (in bits). A hashing function $f(X)$ maps the identifier X—that is, the code of a bus stop—to the address of X in the hash table. The typical code of a bus stop is $X = NNyyy$, where NN is the zonal number, $NN = 1, 2, \ldots, 32$, and yyy is the stop number, $yyy = 1, 2, \ldots, 256$. We defined the following hashing function to set up the hash table used in our algorithm:

$$f(NNyyy) = (NN - 1) \times 256 + yyy \quad (1)$$

The hash table can be considered as 32 × 256 matrix. Each element in the matrix is a binary digit that indicates the unique address of its corresponding bus stop. The address is well defined by the hashing function of Equation 1.

Our Algorithm A2, designed for the search of full routing information of $A-X-B$, has five major steps. The framework of the algorithm design is shown in Figure 3. A detailed description of the algorithm A2 (for connecting transit trips) is given as follows:

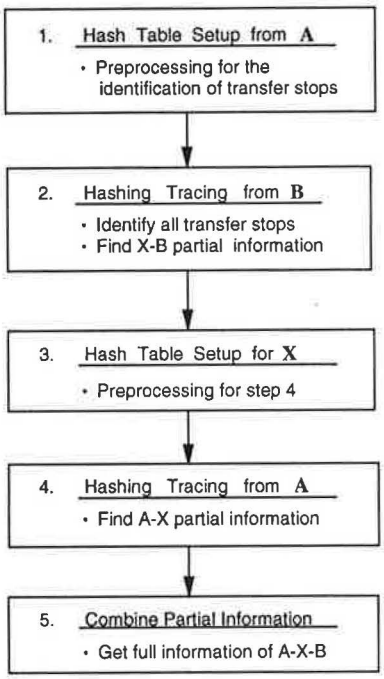

FIGURE 3 Design framework of Algorithm A2.

1. Use Equation 1 to define the hash table for all the downstream stops of A along all the bus lines passing A. All the corresponding addresses of the bus stops such as $a_1, a_2, \ldots, a_7$ (Figure 2) in the hash table are turned on.

2. Select any one line that passes B and for every upstream bus stop of B along that bus line, use the hashing function to map the stop to its corresponding address in the hash table, and check if the address is on. If the corresponding address is on, it is a transfer stop, X; otherwise, continue to check the next bus stop.

Repeat the previous step until all bus lines passing B are traced. Now we have obtained the set of all possible transfer stops and the corresponding bus lines leading to the destination stop B, that is, the routing information about the $X-B$ part.

3. Clear the hash table developed in Step 1, and use the hashing function (Equation 1) to rebuild a hash table for all the transfer stops found in Step 2.

4. Pick one of the bus lines passing stop A, and for every downstream bus stop of A along that line, use the hashing function to check if its corresponding address in the new hash table is on. If yes, keep the record of the transfer stop X as well as the line connecting A to X; otherwise, continue to check the next stop.

Repeat the previous step until all bus lines passing A are traced. Now we have obtained the routing information about the $A-X$ part.

5. Combine partial information of $A-X$ and $X-B$ by using the sort-and-compare techniques similar to those described in Algorithm A1: sort the two ways of Xs in ascending order and do pairwise comparison to combine $A-X$ and $X-B$ for every common transfer stop.

6. Postprocess the route information to rank the alternative connecting paths in preference order based on shortest path or other criteria such as maximal transfer combinations (optional).

Algorithm A2 is much more efficient than the complete-enumeration type of algorithm mentioned earlier. Without using hashing functions, the aforementioned algorithm requires $O(L^2S^2)$ time just to solve the subproblem $S1$. And by using hash-table structures, the computational time of the algorithm for solving the whole problem can be reduced to the order on LS. It is because in each of Steps 1 through 4 of the algorithm, the setup and tracing of a hash table requires only a constant amount of computational time for each bus stop processed, and the number of all potential transfer stops is $O(LS)$. Step 5 involves sort-and-compare procedures, thus it requires time comparable to the order of $S_x[\log(S_x)]$, where S_x is the number of transfer stops of the connecting trip A–X–B. For practical applications in transit systems, the number of all potential transfer stops, LS, usually is much greater than $S_x[\log(S_x)]$. Therefore, the computational time of Algorithm A2 is $O(LS)$.

Note that the $O(LS)$ algorithm is much more efficient than the $O(L^2S^2)$ algorithm described in the previous section. For the Taipei Transit System, the order of LS ranges from 10 to 10^3. This explains why the on-line implementation of Algorithm A2 can run hundreds of times faster than that of the complete-enumeration algorithm and reduce the computational time from minutes to seconds.

Algorithm A2 is primarily designed for microcomputer applications. The hash table takes only about 1K, that is, 32 × 32 = 1,024 bytes, of RAM space, making it very efficient for applications in MS-DOS. With other computing facilities, one can use hash tables bigger than 32 × 32 bytes to store the data of both transfer stops and connecting lines, and to reduce the number of steps as proposed in our algorithm. However, the use of a bigger hash table may not improve the implementation time of the algorithm.

IMPLEMENTATION RESULTS AND CONCLUSIONS

Algorithms A1 and A2 have been successfully running on a microcomputer-based transit information system in Taipei since January 1991. Taipei is the capital city of Taiwan, the Republic of China (ROC). The city of Taipei has an area of 272 km² within its administrative boundaries and a population of about 2.7 million. Bus transit services in the Taipei city are primarily provided by 10 major bus companies that have joined to form the United Operating Center (UOC) of the Taipei Transit System. As of August 1991, the UOC operates more than 250 routes with 3,174 buses, carrying approximately 2.1 million passenger trips a day (*17*).

The transit routing information service system in Taipei is a telephone on-line service system. Two shifts of trained operators answer the telephone inquiries with the help of computerized route information system. The computerized system is programmed in Turbo C and implemented on two IBM PC 386/20 AT microcomputers with math coprocessors and other peripherals. Specifically, the system operates with two telephone lines, (02)321-2000 and (02)341-2000, and two independent computerized route information systems in order to provide maximal on-line services to the public. On average, the system receives and answers about 105 calls each day.

Because of the complicated transit network structure, a significant portion of transit trips in Taipei involve transfers. It was estimated that approximately 1 million passenger trips a day are made through bus transfers (*13*). Therefore, most callers ask for routing information of transfer or connecting trips. The system would first check if there were direct connecting lines. If not, it would automatically search for the bus line connecting through transfer stops and list on the monitor screen the connecting trips in order of shortest distance or maximal transfer combinations, depending on the choice of the rider.

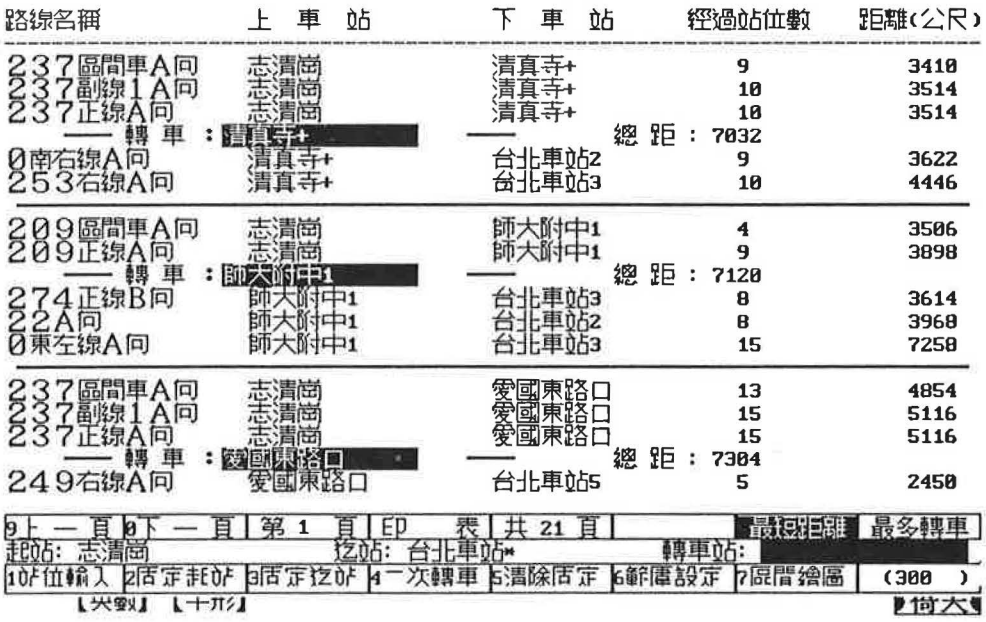

FIGURE 4 Typical screen output of connecting trip information in Taipei Transit System.

A scenario of a connecting trip with seven possible transfer stops was shown earlier in Figure 1. For the dense transit network operated by 10 bus companies in Taipei, it is not uncommon to find connecting trips with more than 10 transfer stops. The full routing information of $A-X-B$ thus can be as long as 20 to 30 (screen) pages. Figure 4 shows a hard copy of the first-page screen output of a 21-page full output file. The routing information listed on the top of the screen tells a person aboard at Chih-Chin Ridge can take Line 237 (zonal, express, or regular) to Mosque and make a transfer from there to Line 0–south or Line 253R to get to the destination stop, the Taipei Train Station. This is the shortest-distance path of the person's total of 38 connecting path alternatives. Algorithm A2 works fine for the system. Taking only 2 or 3 sec to find the 21-page output file and show the first-page output on the monitor screen.

The algorithms proposed in this paper should be useful for microcomputer-based routing information systems for an integrated public transportation system. The techniques of using a small hash table to implement the search algorithm efficiently on a microcomputer appear to be useful for other microcomputer applications as well. With modifications or extensions, Algorithm A2 might be applied to routing information systems for airlines or the track-and-trace systems for courier service companies. Such potential applications need further investigation.

ACKNOWLEDGMENT

This research was jointly supported by the National Science Council, ROC and the Taipei Transit System UOC. An earlier preliminary study of this research was supported by the Institute of Transportation (IOT), Ministry of Transportation and Communications, ROC. The authors are grateful to Chia-Juch Chang, Director General of IOT, for his initiative support.

REFERENCES

1. R. G. P. Tebb. Travel Information Research at TRRL. Presented at the 11th Annual Seminar on Public Transport Operations Research, University of Leeds, England, July 1979.
2. C. O. Tong and A. J. Richardson. Computer General Travel Information for Urban Transit Networks. *Proc., 10th Australia Transport Research Forum*, Melbourne, Vol. 2, May 1985, pp. 197–213.
3. C. J. Vanigrok. *Designing an Integrated Travel Information System*. Monograph, Institunt TNO Wiskunde Infomatieverwerking en Statistiek, Delft, the Netherlands, Sept. 1982.
4. M. W. Pickett. The Production, Dissemination and Costs of an Integrated Public Transport Travel Information System. Report SR657. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1981.
5. J. J. Fruin. *NCTRP Synthesis of Transit Practice 7: Passenger Information Systems for Transit Transfer Facilities*. TRB, National Research Council, Washington, D.C., 1985.
6. K. G. Baass, B. Allard, and R. Chapleau. Traffic and Transport Data Management by Micro-Computer. *Proc., 2nd North American Conference on Microcomputer Applications in Transportation*, ASCE, 1987, pp. 536–545.
7. D. Knight and B. Albee. SCT's Experience with Automated Fleet Management. *Proc., 2nd North American Conference on Microcomputer Applications in Transportation*, ASCE, 1987, pp. 474–483.
8. J. McCarthy. Using Microcomputers for Maintenance Management in a Small Transit Agency. *Proc., 2nd North American Conference on Microcomputer Applications in Transportation*, ASCE, 1987, pp. 387–394.
9. D. J. Wahl. Application of Data Processing Techniques to the Monitoring of Transit System Performance in San Diego. In *Transportation Research Record 1050*, TRB, National Research Council, Washington, D.C., 1985, pp. 53–56.
10. J. M. Ward and M. J. Demetsky. Computerization of Transit Performance Evaluation. *Proc., 2nd North American Conference on Microcomputer Applications in Transportation*, ASCE, 1987, pp. 336–347.
11. M. R. Cutler. Impact of Technology and Labor Management Strategies on the Efficiency of Telephone Information Services. In *Transportation Research Record 1039*, TRB, National Research Council, Washington, D.C., 1985, pp. 1–9.
12. A. Han and N. Wilson. The Allocation of Buses in Heavily Utilized Networks with Overlapping Routes. *Transportation Research*, Vol. 16B, 1982, pp. 221–232.
13. A. F. Han. Assessment of Transfer Penalty to Bus Riders in Taipei: A Disaggregate Demand Modeling Approach. In *Transportation Research Record 1139*, TRB, National Research Council, Washington, D.C., 1988, pp. 8–14.
14. R. Kwan. Coordination of Joint Headway. In *Computer-Aided Transit Scheduling* (J. Daduna and A. Wren, eds.). Springer-Verlag, New York, N.Y., 1988, pp. 304–314.
15. A. Han and H. L. Chang. *Route Structure Analysis of the Taipei Transit Network* (in Chinese). Research report. Department of Transportation Engineering and Management, National Chiao Tung University, Taiwan, 1987.
16. E. Horowitz and S. Sahni. Hashing. In *Fundamentals of Data Structure in PASCAL*, 3rd ed. Computer Science Press, New York, N.Y., 1982, Chapter 9.
17. *The Statistical Abstract of Taipei Traffic* (in Chinese). Taipei Municipal Government Office, Taiwan, ROC, Nov. 1991.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Influence of Urban Network Features on Quality of Traffic Service

SIAMAK A. ARDEKANI, JAMES C. WILLIAMS, AND SUDARSHANA BHAT

The relation between street network geometric and control features and the network quality of traffic service is investigated. The goal is to quantify macroscopically the degree of improvements made when an urban street network undergoes modifications in its control or geometric features. The quality of traffic service in several city networks is assessed through field calibration of the two-fluid model. The model parameters are then correlated to 10 geometric and control features for each city network. It is found that under low traffic concentrations in the network, the average speed limit and the degree of signal progression are the most influential features. On the other hand, for high-concentration conditions, the fraction of one-way streets, the average number of lanes, and the fraction of signals actuated most affect the service quality. The density of signalized intersections affects peak and off-peak traffic. The signal density is beneficial during peak periods but detrimental during off-peak periods.

In traffic engineering practice, when an urban street network such as a downtown street system undergoes modifications in its geometric configuration or its control strategy, it is often desirable to obtain some quantitative measure of the resulting improvements. Similarly, when the need arises to improve the quality of traffic service in a street network, it is necessary to be able to predict the level of improvements to be attained as a result of any specific modifications in the network geometry or control features. This permits engineers to identify those strategies that yield the greatest level of improvement per unit cost of implementation.

The work reported herein presents a methodology to assess the degree of influence of various geometric and control features on the quality of service in an urban street network. From an initial list of some 20 potential factors, 10 key geometric and control attributes have been considered. They include

- Block length,
- Extent of one-way streets,
- Number of lanes per street,
- Intersection density,
- Signal density,
- Speed limit,
- Cycle length,
- Extent of on-street parking,
- Degree of signal actuation, and
- Degree of signal progression.

Department of Civil Engineering, P.O. Box 19308, University of Texas, Arlington, Tex. 76019.

Data on each of these features are collected for 19 urban street networks. The quality of traffic service across the networks studied is compared in light of the values of the geometric and control variables in each respective network. Statistical analyses are performed to identify those variables that most affect the traffic service quality as well as their respective degrees of impact.

The quality of service in each of the 19 networks is macroscopically quantified using the two-fluid model methodology. The model correlates the average travel time per unit distance to the stopped delay per unit distance in a network. Simply stated, under similar traffic loading conditions (same average concentration), a network with better quality of service will yield mile-long trips with shorter trip time and stop time values. Although the amount of traffic load on a facility is generally expressed in terms of volume, previous field and simulation studies (*1–3*) show a strong correlation between the averages of flow and concentration across a network, as is the case along a single roadway. As such, average concentration has been used in this study as a measure of traffic demand in a network since it is considerably easier to measure than the networkwide average flow.

The two-fluid model, formulated on the basis of this principle, has been used to quantify the quality of traffic service in several cities around the world (*4*). Subsequent visits to some of those cities has shown the model to be robust and accurate (*5*). The model parameters are hence used to characterize the service quality in the networks in this study. A more detailed description of the model is therefore merited.

TWO-FLUID MODEL

As discussed previously by Herman and Prigogine (*6*), the concept of a two-fluid model appeared in the kinetic theory of multilane highway traffic when the transition to the so-called collective flow regime was made at sufficiently high vehicular concentrations. For highway traffic, the speed distribution for the cars splits into two parts at the collective transition: one part corresponds to the moving vehicles and the other to the vehicles that are stopped as a result of local conditions such as traffic jams. Likewise, the traffic in a city network can be considered to consist of two traffic fluids: one part composed of the moving cars and the other of cars that are stopped as a result of congestion, traffic signals, stop signs, other traffic control devices, and obstructions resulting from construction, accidents, and such—but not parked vehicles. The parked cars are ignored since they are not a component

of the traffic but instead form a part of the geometric configuration of the street.

In the two-fluid model the ideas in the kinetic theory of traffic are followed by assuming that the average speed of the moving cars, v_r, depends on the fraction of the cars that are moving, f_r, in the following form:

$$v_r = v f_r^{-1} = v_m f_r^{\mathbf{n}} = v_m (1 - f_s)^{\mathbf{n}} \tag{1}$$

where

f_s = average fraction of vehicles stopped,
v_m = average maximum running speed in the network system,
v = average speed of the traffic, and
n = parameter whose significance will be discussed later.

The boundary conditions are reasonably satisfied because for $f_s = 0$ and 1, the running speeds are v_m and 0, respectively. The following identities should also be noted:

$$f_r + f_s = 1 \tag{2}$$

$$v_m = 1/T_m \tag{3}$$

$$v_r = 1/T_r \tag{4}$$

$$v = 1/T \tag{5}$$

where

T_m = parameter representing the average minimum trip time per unit distance,
T_r = average running time per unit distance, and
T = trip time per unit distance.

If, in addition, the stop time per unit distance is denoted by T_s, it follows that

$$T = T_s + T_r \tag{6}$$

In the model it is also assumed that the fraction of the time stopped for an individual vehicle circulating in a network, (T_s/T), is equal to the average fraction of vehicles stopped in the system, f_s, over the same time period, namely,

$$f_s = (T_s/T) \tag{7}$$

It is important to remember with regard to the second assumption that, if the concentration varies widely—that is, fluctuates rapidly during the time of a trip—the condition stated in Equation 7 may not be satisfied (*4*). The concentration must vary slowly over the time scale during which (T_s/T) and f_s are measured.

The assumptions stated in Equations 1 and 7 lead to the two-fluid model relation between the trip time per unit distance, T, and the running time per unit distance, T_r, namely,

$$T_r = T_m^{1/(\mathbf{n}+1)} T^{\mathbf{n}/(\mathbf{n}+1)} \tag{8}$$

yielding the final result

$$T_s = T - T_m^{1/(\mathbf{n}+1)} T^{\mathbf{n}/(\mathbf{n}+1)} \tag{9}$$

It is emphasized that in the two-fluid theory the variables are always meant to be averages taken over the entire system.

It follows from Equation 8 that

$$\log T_r = [1/(\mathbf{n} + 1)] \log T_m + [\mathbf{n}/(\mathbf{n} + 1)] \log T \tag{10}$$

or

$$\log T_r = A + B \log T \tag{11}$$

with

$$\mathbf{n} = B/(1 - B) \tag{12}$$

and

$$\log T_m = A/(1 - B) \tag{13}$$

The parameters **n** and T_m associated with a traffic network can be obtained from Equations 12 and 13 by collecting trip time versus stop time data for a test vehicle circulating in that traffic network.

On a trip time–stop time diagram, the two-fluid model represented by Equation 9 plots as a slightly concave down curve. Figure 1 shows the curves for two hypothetical networks with the same value of T_m = 1.5 min/mi but different **n**-values of 1 and 3, respectively. Also shown in Figure 1 is a line representing a fraction of vehicles stopped, $f_s = T_s/T$, of 0.25. It can be seen that for the same fraction of vehicles stopped, representing the same traffic loading conditions, the network with a smaller value of **n** = 1 yields considerably shorter trip time and stop time values. Previous studies have shown that the average fraction of vehicles stopped during a time period is a function of the average concentration in the network during that time (*1*,*4*). Furthermore, f_s has been shown, through the use of aerial photographs (*1*), to not vary greatly from one city to another for a given level of concentration.

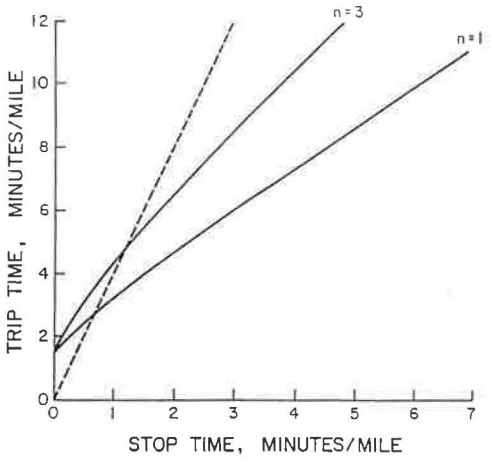

FIGURE 1 Two-fluid trends for two hypothetical networks with same T_m-value of 1.5 min/mi but different n-values of 1 and 3, respectively; dashed line corresponds to 0.25 fraction of vehicles stopped.

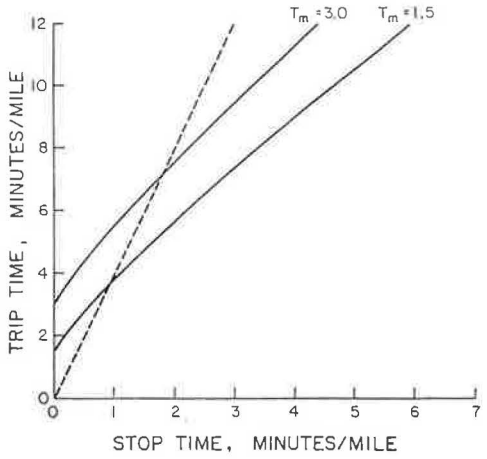

FIGURE 2 Two-fluid trends for two hypothetical networks with same n-value of 2 but T_m values of 1.5 and 3.0 min/mi, respectively; dashed line corresponds to 0.25 fraction of vehicles stopped.

A similar case can be made for the parameter T_m. Figure 2 shows two networks with the same **n**-value (**n** = 2) but different T_m-values of 1.5 and 3.0 min/mi, respectively. Again, the network with a lower T_m-value yields much lower trip time and stop time values for a given fraction of vehicles stopped.

A street network with a better quality of traffic service can be said to be one that yields smaller trip time and stop time per unit distance for a given level of traffic demand. Through the use of aerial photographs, it has been shown that for a level of traffic concentration (demand) a network with smaller values of T_m- and **n**-parameters offers lower trip time and stop time per unit distance (*1*). Parameters T_m and **n** can therefore be considered to be meaningful and reliable indicators of the quality of traffic service in an urban street network. Thus, the study described sets out to determine the degree of influence of the various network features on the parameters T_m and **n** as measures of network service quality.

DATA COLLECTION

The data collection phase of the study was performed in two steps. A number of networks for which the two-fluid model had already been calibrated were selected. These included downtown networks of Albuquerque, New Mexico (1983); Austin, Texas (1984); Dallas, Texas (1983); Lubbock, Texas (1984); Houston, Texas (1983); San Antonio, Texas (1984); Mexico City, Mexico (1983); and Matamoros, Mexico (1983). With the exception of the two cities in Mexico, all other networks were revisited in 1990 and the two-fluid model was recalibrated for each network. Trip time–stop time data were also collected in Texas cities of Arlington and Fort Worth in 1990. Figure 3 shows the trip time–stop time data and the resulting two-fluid trend for the Fort Worth central business district (CBD). Each point represents a 1-mi-long trip while circulating in the CBD network using a chase-car technique (*4*). Observations were made, as in all other cities, during various times of day and traffic conditions.

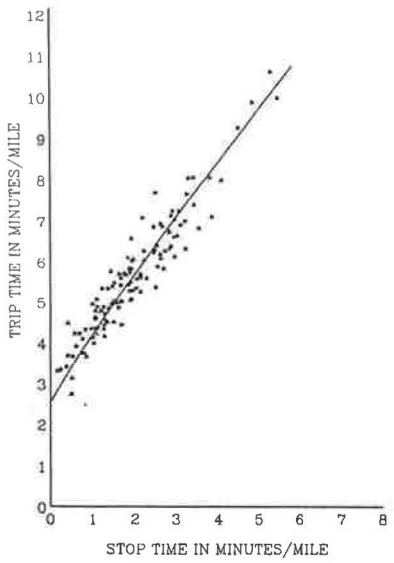

FIGURE 3 Trip time–stop time data and resulting two-fluid trend for 1990 Fort Worth CBD; T_m = 2.52 min/mi, n = 0.88.

Additionally, Arlington, Dallas, and San Antonio underwent major changes in their control or street geometry in their downtown systems during the 1990–1991 period and were restudied in 1991. In Arlington, Cooper Street, a major arterial in the CBD, was reopened to traffic following a multiyear widening and reconstruction project. In Dallas, a portion of the previously studied network was closed to traffic for several weeks during filming of the motion picture "JFK." Early in 1991, a major signal retiming plan was implemented in the San Antonio CBD, and a number of street construction projects were concluded, improving the network geometry.

In all, 19 CBD networks were calibrated. The number of miles of data collected in each city varied from 80 to 120 mi depending on the network size. In each city, data were collected under a wide range of traffic demand conditions. In Texas cities in which concentration was measured, the average concentration varied from 5 vehicles per lane mile during off-peak to 35 vehicles per lane mile during peak (*1*). Table 1 summarizes the two-fluid model parameters obtained in each of the networks under study. As shown in this table, T_m-values range from 2.98 min/mi for Matamoros to 1.72 min/mi for Mexico City; **n**-values range from 2.10 for Matamoros to 0.61 for the 1991 Arlington network.

The parameter values reported in Table 1 have been tested for serial correlation. The test is necessary because the trip time–stop time data used in estimating the parameters are obtained from consecutive 1-mi microtrips. It is therefore reasonable to expect that the random error terms in the least-squares estimation technique violate the normal independence assumption and exhibit serial correlation. Consequently, a formal test for serial correlation was performed (*7*). Six of the field data sets showed first-order serial correlation and required slight corrections using a procedure detailed elsewhere (*7*).

TABLE 1 Two-Fluid Parameters for Downtown Networks Under Study

Downtown NETWORK	T_m* (minutes/mile)	n*
Arlington 1 (90)	1.98	1.06
Arlington 2 (91)	1.95	0.61
Fort Worth (90)	2.52	0.88
Dallas 1 (83)	2.12	1.36
Dallas 2 (90)	2.79	0.77
Dallas 2 (91)	2.77	0.80
Austin (84)	1.95	1.58
Austin (90)	2.14	1.46
Lubbock (84)	2.03	0.97
Lubbock (90)	1.78	1.27
Houston (83)	2.70	0.80
Houston (90)	2.24	1.11
San Antonio (84)	1.99	1.33
San Antonio (90)	2.52	1.14
San Antonio (91)	2.40	1.05
Albuquerque (83)	1.93	1.62
Albuquerque (90)	2.32	0.94
Matamoros (83)	2.98	2.10
Mexico City (83)	1.72	1.63

* Adjusted for serial correlation

The T_m- and **n**-values represent a wide variation in the traffic qualities from one network to another. A preliminary list of network features that could help explain such variations in traffic service quality was devised. From this list, 10 geometric and control features that were potentially influential yet not difficult to quantitatively measure were selected. They include

- X_1 (average block length). Calculated as the ratio of the total route miles to the total number of blocks in the network.
- X_2 (fraction of one-way streets). Calculated as the ratio of the total length of one-way streets to the total route miles in the network.
- X_3 (average number of lanes per street). Calculated by dividing the total number of lane miles open to traffic during the p.m. peak hour by the total route miles in the network.
- X_4 (intersection density). Calculated as the total number of intersections in the network divided by the total network land area (in square miles).
- X_5 (signal density). Calculated as the total number of signalized intersections in the network divided by the total network land area (in square miles).
- X_6 (average speed limit). Calculated by weighting according to the lengths of streets for which the speed limit (in miles per hour) is posted.
- X_7 (average cycle length). Calculated as the average of signal cycle lengths (in seconds) in the network during the p.m. peak hour. No weight based on approach volumes was used.
- X_8 (fraction of curb miles with parking allowed). Calculated as the ratio of the total number of curb miles on which parking was allowed during the p.m. peak period to the total curb miles in the network.
- X_9 (fraction of signals actuated). Calculated as the ratio of the total number of signal actuated intersections to the total number of signalized intersections in the network.
- X_{10} (fraction of approaches with signal progression). Calculated as the ratio of the number of intersection approaches in the network that are part of a progression scheme to the total number of signalized intersection approaches in the network.

The values of X_1 through X_{10} were determined for each of the networks under study. City maps, city transportation departments, and field surveys were the major data collection sources. Fairly accurate city maps were used to determine the average block length (X_1), the fraction of one-way streets (X_2), and the intersection density (X_4). Data related to signalization (X_5, X_7, X_9, X_{10}) were obtained through the traffic engineering offices in each city. Field surveys were performed to determine the values for lanes per street (X_3), speed limit (X_6), and curb parking (X_8). The data for the latter variables had already been recorded for the networks studied in 1983 and 1984 (*8*). Table 2 summarizes the data obtained.

DATA ANALYSIS AND RESULTS

Stepwise regression analyses were performed to identify those network geometric and control variables that most affected the quality of service parameters T_m and **n**. The data in Tables 1 and 2 were used in the analyses.

Stepwise regression is a procedure to select among a number of potential variables those that are best suited for inclusion in the regression model. Stepwise regression examines all candidate variables for the one that best explains the variation in the dependent variable. This variable enters the model. The significance of each variable entering the model is assessed by the *F*-statistic at a level of significance specified by the modeler. The process is repeated to find the next most influential variable to enter. At each step, however, the correlations between pairs of independent variables already in the model are examined. For strong correlations between independent variables, the least significant of those variables exits the model. Exiting the model is also determined from the *F*-statistic at a user-specified level of significance. Stepwise terminates and a model is output when either of these cases is met: (a) none of the variables outside the model has an *F*-value higher than specified by the user for entry into the model, and every variable in the model has an *F*-value higher than the user-specified value for leaving the model, and (b) the variable outside the model, which is significant enough to enter the model, is one that was discarded at the last step.

Two sets of stepwise regression analyses were conducted, with T_m and **n** as dependent (prediction) variables. In each case, independent variables X_1 through X_{10} (Table 2) were considered for entry into the model. A level of significance

TABLE 2 Network Geometric and Control Features

Downtown Network	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Arlington 1 (90)	496	0	2.77	108	17	30.6	94.0	0.375	1.000	0.537
Arlington 2 (91)	479	0	2.92	124	25	30.3	91.0	0.315	1.000	0.557
Fort Worth (90)	300	0.71	3.0	355	268	30.0	75.0	0.360	0.110	0.50
Dallas 1 (83)	350	0.40	2.6	160	80	30.0	80.0	0.218	0.324	0.328
Dallas 2 (90)	338	0.65	3.2	282	224	30.0	80.0	0.131	0	0.488
Dallas 2 (91)	340	0.67	3.1	270	215	30.0	84.0	0.130	0	0.290
Austin (84)	430	0.52	3.0	209	82	30.3	64.0	0.834	0	0.32
Austin (90)	409	0.43	2.9	175	82	30.2	68.0	0.834	0	0.32
Lubbock (84)	380	0.30	3.1	187	45	31.1	81.0	0.602	0.01	0.27
Lubbock (90)	446	0.15	3.05	153	30	31.1	97.1	0.602	0.02	0.37
Houston (83)	324	0.76	4.3	309	213	30.0	80.0	0.366	0	0.77
Houston (90)	324	0.76	4.3	309	213	30.0	80.0	0.366	0	0.77
San Antonio (84)	365	0.45	2.7	244	144	30.0	100.0	0.280	0.008	0.327
San Antonio (90)	365	0.41	2.6	240	150	29.8	89.3	0.280	0.008	0.395
San Antonio (91)	365	0.40	2.82	252	147	30.0	61.8	0.269	0.008	0.451
Albuquerque (83)	381	0.47	2.7	222	111	25.5	61.0	0.441	0.171	0.52
Albuquerque (90)	381	0.46	2.7	221	113	25.5	62.2	0.428	0.169	0.53
Matamoros (83)	380	0.67	1.2	269	22	12.9	70.0	0.850	0	0
Mexico City (83)	359	0.72	3.9	225	47	31.2	120.0	0.110	0	0.613

X_1: Avg Block Length (ft)

of 30 percent was used as the criterion for including or removing variables from the model. A sensitivity analysis on the level of significance indicated that the model structure in terms of variables retained varied a great deal for significance levels below 30 percent while at this level and higher stable T_m- and **n**-models were obtained.

At the 30 percent level of significance, T_m was shown to be most influenced by the signal density (X_5), average speed limit (X_6), and fraction of approaches in the network with signal progression (X_{10}). The parameter **n** was most influenced by four variables at the 30 percent level of significance or higher. They included the fraction of one-way streets (X_2), the average number of lanes per street (X_3), the signal density (X_5), and the fraction of signals actuated (X_9). The resulting models for T_m and **n** are as follows:

$$T_m = 3.93 + 0.0035X_5 - 0.047X_6 - 0.433X_{10}$$

$$R^2 = .72 \qquad (14)$$

$$\mathbf{n} = 1.73 + 1.124X_2 - 0.180X_3 - 0.0042X_5 - 0.271X_9$$

$$R^2 = .75 \qquad (15)$$

The parameter T_m is an estimate of the average minimum travel time per unit distance or the reciprocal of the average maximum speed in the network. T_m is, therefore, expected to be highly influenced by the speed limit, with a higher speed limit yielding a lower T_m-value. It is therefore expected that the variable X_6 appears in the T_m-model and has a negative coefficient, as is the case in Equation 14. Equation 14 also indicates that the greater the signal density in an area, the greater the value of T_m, that is, the lower the quality of traffic service. This is also somewhat intuitive, because a higher number of signals per unit area generally results in greater delay to off-peak traffic, hence a greater T_m-value. On the other hand, a greater number of approaches in signal progression will result in a lower T_m-value (negative X_{10} coefficient), as is to be expected.

The influence of speed limit (X_6) on the value of T_m and its consequent degree of impact on the service quality are graphically illustrated in Figure 4. Figure 4 shows the two-fluid trend for the current network of downtown Fort Worth. Also shown is the line $f_s = T_s/T = 0.36$ corresponding to the current peak-period conditions in the Fort Worth CBD. Under the scenario examined, the average speed limit in Fort Worth is to be increased by 30 percent. By rewriting Equation 14 in the form

$$\Delta T_m = 0.0035\Delta X_5 - 0.047\Delta X_6 - 0.433\Delta X_{10} \qquad (16)$$

the expected T_m-value for a 30 percent increase in speed limit can be calculated.

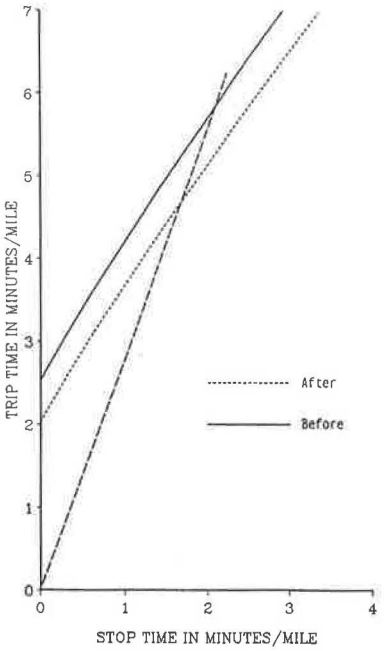

FIGURE 4 Two-fluid trends before and after 30 percent increase in average speed limit (X_6) in network; dashed line corresponds to fraction of vehicles stopped (0.36) in network during p.m. peak period.

The dashed curve in Figure 4 represents the expected two-fluid trend should such a change be implemented. The new curve is obtained by a decrease in T_m of 0.42 min/mi from the current 2.52 to 2.10 min/mi. As can be seen, for the current peak-period fraction of vehicles stopped in Forth Worth, the peak-period trip time and stop time per unit distance would be decreased on the average by about 0.97 and 0.35 min/mi vehicle, respectively—a 16.7 percent reduction in each. By the same token, a 30 percent increase in networkwide signal progression (X_{10}) would save 0.65 and 0.23 min/mi in peak-period trip time and stop time, respectively.

Whereas T_m is mostly influenced by the network control features, the parameter **n** appears to be more a function of the geometric characteristics of the network. As depicted by Equation 15, the value of **n** is increased as the fraction of one-way streets (X_2) is increased, thus implying a poorer quality of traffic service. At first glance this may be counterintuitive. It must be noted, however, that in an urban street network, in which land access is the primary function, one-way streets generally do result in a poorer traffic circulation pattern. This could translate into a higher level of interactions among vehicles and therefore a lower service quality. On the other hand, if the primary objective is to serve the through traffic at relatively high speeds and volumes, one-way streets will be most suitable. However, traffic in a CBD network generally has either its origin or destination within the CBD itself.

The impact of a 30 percent reduction in fraction of one-way streets (X_2) is examined in Figure 5. Figure 5 shows the current two-fluid trend for Fort Worth as well as the expected trend should the fraction of one-way streets be reduced by 30 percent. As can be seen, for the peak-period conditions in Forth Worth, as represented by the line $f_s = 0.36$, the peak trip time and stop time values would decrease by about 0.59 and 0.21 min/mi (10 percent each), respectively. This reduction would result from a decrease in the Fort Worth CBD **n**-value from the original 0.88 to 0.64.

A similar analysis is performed by the effect of a 30 percent increase in the average number of lanes per street (X_3) in Forth Worth, say, by prohibiting curb-side parking, by re-striping, or by widening streets. Keeping all other variables the same, the value of **n** would decrease to 0.72, resulting in an expected improvement in the quality of traffic service. The peak-period trip time and stop time would be reduced by about 0.40 and 0.15 min/mi (7 percent each), respectively.

Likewise, a 30 percent increase in the fraction of actuated signals (X_9) is examined. As expected, the trip and stop times would both decline should fixed-time signals be converted to actuated signals. However, as is also to be expected, the magnitude of the impact of this change would not be significant during the peak period when an actuated signal is likely to max out every cycle, thus operating virtually like a fixed-time signal. In this case, the reductions would be only about 0.02 and 0.01 min/mi (0.3 percent each) in the peak-period trip and stop times for the Fort Worth CBD.

In studying the influence of signal density (X_5) on the traffic service quality, it should be noted that although an increase in signal density would reduce the value of **n** (Equation 15), the opposite would be true for T_m (Equation 14).

Figure 6 examines the impact of increasing the signal density in the Fort Worth CBD by 30 percent. Such an increase would

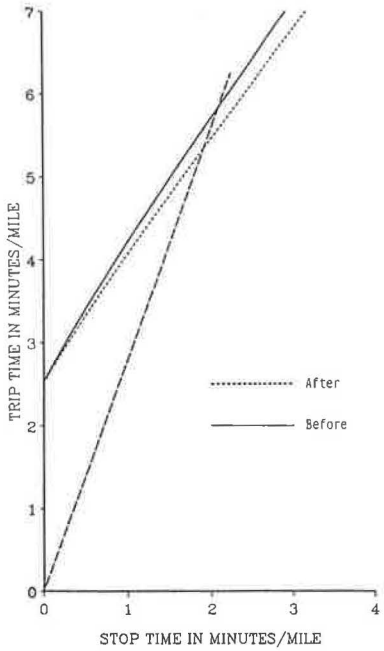

FIGURE 5 Two-fluid trends before and after 30 percent decrease in fraction of one-way streets (X_2) in network; Fort Worth CBD, f_s = 0.36.

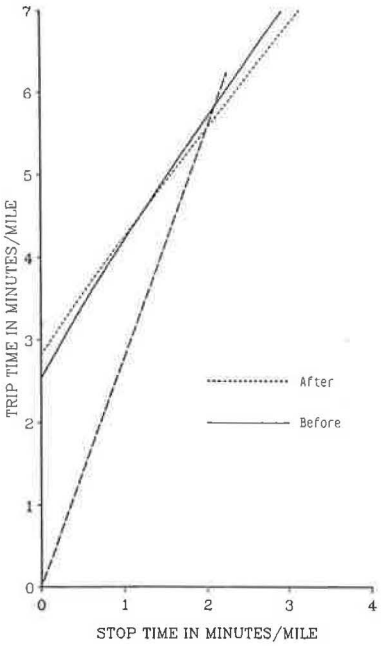

FIGURE 6 Two-fluid trends before and after 30 percent increase in signal density in network; Fort Worth CBD, f_s = 0.36.

lower the **n**-value to 0.54 and increase the T_m-value to 2.80. Consequently, the new two-fluid trend would have a higher intercept but a flatter slope, thus crossing the existing trend. The net result would be a 4.5 percent reduction each in peak-period trip time and stop time. However, the trip time and stop time values would increase during the off-peak period should more intersections in the area be signalized. It can be concluded that increasing the number of signals would harm the off-peak traffic operations but benefit peak traffic. This provides a strong argument for converting signals to flashing operation during off-peak periods.

Although these conclusions are somewhat intuitive, the procedure enables the traffic engineer to quantify the impact of such policy in terms of reductions in trip time and stop time.

CONCLUSIONS AND DISCUSSION OF RESULTS

The influence of the geometric and control features studied on the traffic service quality may be intuitive, but the degree of impact of each is not. The models derived provide analytical tools for quantifying the impact of such changes on the quality of traffic service. Without resorting to network simulation models that are time-consuming and expensive to code, engineers can use these macroscopic tools to examine the consequences of various policy decisions such as converting a two-way street to one-way, prohibiting on-street parking, adding or removing signals, converting signals to flashing operations, and adding or removing lanes.

Although the proposed methodology is promising, the models presented are based on a limited data base. More networks with varied geometric and control conditions must be added. More network features should also be considered. To perform a robust statistical analysis, the number of city networks should be more than twice the number of variables to provide a reasonable degree of freedom.

Despite the need for more field data, obtaining such data is expensive and extremely labor-intensive. The use of simulation may be an alternative in expanding the data base. NETSIM has been used successfully to calibrate the two-fluid model (*2,3,9*). The simulation environment will also allow a much wider range of variation to be achieved in the network feature variables. Furthermore, a higher number of variables can be examined without much additional effort, since the characteristics of the simulated network are readily known during the coding process. Simulation studies are being conducted using the TRAF-NETSIM package. The Fort Worth CBD network, with about 180 intersections and about 400 street links, has been coded for this purpose, and initial runs have been successful. The simulation studies will allow a more detailed examination and expansion of the relationships developed on the basis of field studies reported.

ACKNOWLEDGMENTS

This work has been sponsored by a Texas Higher Education Coordinating Board research project. The authors wish to acknowledge the help of traffic engineers in Arlington, Dallas, Fort Worth, and San Antonio: they are Ali Mozdbar, Beth Ramirez, Russell Wiles, and Rick Denney, respectively. The authors also wish to thank Robert Herman for many fruitful discussions in the course of years. Thanks are also extended to those students at the University of Texas at Arlington who participated in data collection.

REFERENCES

1. S. Ardekani and R. Herman. Urban Network-Wide Variables and Their Relations. *Transportation Science*, Vol. 21, No. 1, 1987.
2. J. C. Williams, H. S. Mahmassani, and R. Herman. Urban Traffic Network Flow Models. In *Transportation Research Record 1112*, TRB, National Research Council, Washington, D.C., 1987.
3. H. S. Mahmassani, J. C. Williams, and R. Herman. Investigation of Network-Level Traffic Flow Relationships: Some Simulation Results. In *Transportation Research Record 971*, TRB, National Research Council, Washington, D.C., 1984.
4. R. Herman and S. Ardekani. Characterizing Traffic Conditions in Urban Areas. *Transportation Science*, Vol. 18, No. 2, 1984.
5. S. Ardekani and R. Herman. A Comparison of the Quality of Traffic Service in Downtown Networks of Various Cities Around the World. *Traffic Engineering and Control*, Vol. 26, No. 2, 1985.
6. R. Herman and I. Prigogine. A Two-Fluid Approach to Town Traffic. *Science*, Vol. 204, 1979.
7. J. Neter, W. Wasserman, and M. H. Kutner. *Applied Linear Statistical Models*, 2nd ed., Richard D. Irwin, Homewood, Ill., 1985.
8. M. T. Ayadh. *Influence of the City Geometric Features on the Two-Fluid Parameters*. M.S. thesis. Virginia Polytechnic Institute and State University, Blacksburg, 1986.
9. J. C. Williams, H. S. Mahmassani, and R. Herman. Analysis of Traffic Network Flow Relations and Two-Fluid Model Parameter Sensitivity. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Advanced Traffic Management System: Real-Time Network Traffic Simulation Methodology with a Massively Parallel Computing Architecture

THANAVAT JUNCHAYA, GANG-LEN CHANG, AND ALBERTO SANTIAGO

The advent of parallel computing architectures presents an opportunity for transportation professionals to simulate a large-scale traffic network with sufficiently fast response time for real-time operation. However, it necessitates a fundamental change in the modeling algorithm to take full advantage of parallel computing. Such a methodology to simulate traffic network with the Connection Machine, a massively parallel computer, is described. The basic parallel computing architectures are introduced, along with a list of commercially available parallel computers. This is followed by an in-depth presentation of the proposed simulation methodology with a massively parallel computer. The proposed traffic simulation model has an inherent path-processing capability to represent drivers' route choice behavior at the individual-vehicle level. It has been implemented on the Connection Machine with 16,384 processors. Preliminary simulation experiments indicate that massively parallel computers are a practicable alternative for achieving real-time application. The experiment shows that the Connection Machine with 16k processors can simulate 32,000 vehicles for 30 min at 2-sec intervals within 2 min of running time.

Many metropolitan areas around the world face serious congestion problems that threaten to deteriorate the quality of life and increase air pollution. It was estimated that traffic congestion in 1987 accounted for more than 2 billion vehicle-hr of delay and 2.2 billion gal of excessive fuel consumption in the United States (*1*). In the next decade, the unavoidable dramatic increase in travel demand coupled with the diminishing construction of new transportation facilities will certainly worsen the traffic condition unless innovative congestion-relief methods can be developed and implemented in time.

One area that seems most promising in alleviating congestion is the development of intelligent vehicle-highway systems (IVHSs), specifically in the form of advanced traveler information systems (ATISs) and advanced traffic management systems (ATMSs). Significant improvements in mobility, highway safety, and productivity can thus be achieved through integrated applications of advanced technologies to surveillance, communications, route guidance, and control process (*2*). Research is being undertaken in such areas as adaptive traffic control, incident detection, real-time traffic assignment, and corridor optimization. Successful implementation of these ATMS developments would ensure optimal networkwise performance.

The anticipated benefits of these control methods depend on the complex interactions among principal traffic system components. These systems include driver behavior, level of congestion, dynamic nature of traffic patterns, and the network's geometric configuration. It is crucial to the design of these strategies that a comprehensive understanding of the complex interrelations between these key system components be established. Because it is often difficult for theoretical formulations to take all such complexities into account, traffic simulation offers the unique capability to conduct performance evaluations. In addition, an effective on-line simulation model would enable the ATMS control center to project promptly future traffic patterns considering any previously implemented strategies in a real-time operating environment. A graphical illustration of a traffic simulation model's function in ATIS-ATMS implementation is presented in Figure 1.

Such a real-time network traffic simulation model is required to have at least the following features: (a) a realistic representation of traffic characteristics and geometric configurations; (b) the capability to simulate both freeway and surface street networks at different levels of detail; and (c) a path-processing capability to represent drivers' route choice behavior at the individual-vehicle level. In addition, in order to be operational in a real-time basis, the software design must be efficient and well structured and maximize the utility of the hosting hardware. A comprehensive literature review clearly indicated that none of the existing simulation and assignment models fully meets these functional requirements. A detailed discussion in this regard can be found elsewhere (*3*).

In terms of providing sufficiently fast response to simulate a large-scale network, the use of advanced parallel computing architectures appears to be one of the most promising methods. However, adoption of this posture may require a fundamental change in the modeling algorithm. This cannot be achieved through the existing traffic simulation methodologies developed mainly for conventional computing machines.

The objective of this paper is to introduce a real-time traffic network simulation methodology that fully utilizes the capability of massively parallel machines. Some basic parallel computing architectures are introduced along with a list of

T. Junchaya, G. L. Chang, Department of Civil Engineering, University of Maryland, College Park, Md. 20742. A. Santiago, Federal Highway Administration, McLean, Va. 22101.

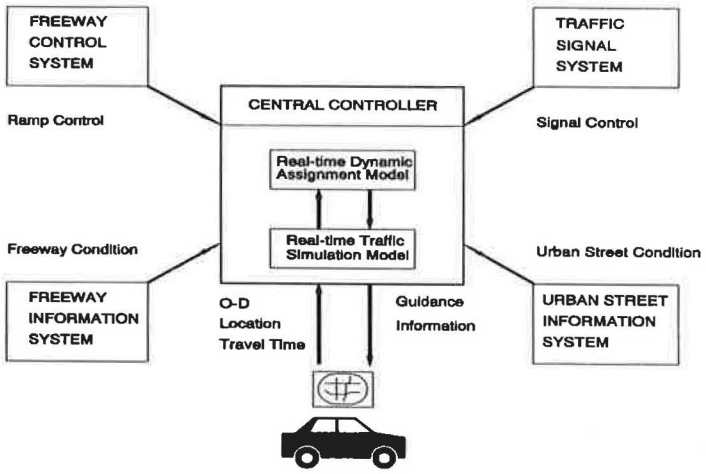

FIGURE 1 ATIS-ATMS system component.

commercially available parallel computers. There is also an in-depth presentation of the proposed simulation methodology with a massively parallel computer, specifically the Connection Machine, and some empirical results are given. Ongoing research activities and potential integration with related studies are also presented.

REVIEW OF PARALLEL PROCESSING METHODOLOGY AND ARCHITECTURE

Prodigious advances in computer architectures and capabilities have taken place in the past two decades. New technologies and innovative architectures will continue to appear in addition to the already bewildering array of configurations. However, it has become apparent to most researchers that the most promising long-term approach to achieve affordable, accessible supercomputing is parallel processing.

Parallel processing has been defined as follows:

> Parallel processing is an efficient form of information processing which emphasizes the exploitation of concurrent events in the computing process. Concurrency implies parallelism, simultaneity, and pipelining. Parallel events may occur in multiple resources during the same time simultaneous events may occur at the same time instant; and pipelined events may occur in overlapped time spans. These concurrent events are attainable in a computer system at various processing levels. Parallel processing demands concurrent execution of many programs in the computer. It is in contrast to sequential processing. It is a cost-effective means to improve system performance through concurrent activities in the computer. (*4*)

Classification of Parallel Processing Systems

Although no fully satisfactory taxonomy of multiprocessors has been established, parallel processing systems can still be classified according to their design alternatives (*5*). These include program control, interconnection methods, form of information exchange, and processing element granularity.

Program Control

The most widely used classification scheme was proposed by Flynn (*6*). He classified computer architectures into four categories:

1. Single instruction stream–single data stream (SISD): one instruction at a time is performed on one piece of data.
2. Single instruction stream–multiple data stream (SIMD): one type of instruction can be executed simultaneously on multiple data.
3. Multiple instruction stream–single data stream (MISD): different instructions can be performed simultaneously on the same data.
4. Multiple instruction stream–multiple data stream (MIMD): different instructions can be performed currently on multiple data.

Interconnection Methods

One important factor in determining the performance of the multiprocessor is the technique selected to connect the processing elements. Several alternatives have been proposed for the topology of the interconnection network in a multiprocessor (*7*), depending on whether the interconnections are dynamic or static. Dynamic networks, in which the interconnections are under program control, include shuffle exchange networks and the crossbar switch. Static topologies include ring, star, nearest-neighbor mesh, systolic array, and hypercube configuration.

Form of Information Exchange

There are two major forms of information exchange: shared memory and distributed memory (message passing). In a shared memory system, the processing elements have access to common memory resources and exchange data by successive read-write operations. In a distributed memory or message-passing

systems, each processing element has its own local memory, and elements exchange data by transmitting messages through the interconnection network.

Processing Element Granularity

The number of processors in parallel computers ranges from two elements to many thousands. Some computers have a few very powerful processors (coarse-grain), such as Cray Y-MP; others consist of a very large number of simple processors (fine-grain), such as the Connection Machine.

Review of Existing Parallel Computers

Most supercomputers explore parallel processing in the SIMD or MIMD mode. SIMD machines, the simpler of the two, use either the array processor or the pipeline approach. The fundamental differences between the SIMD and MIMD machines can be summarized as follows:

- SIMD
  - –All processors are given the same instruction.
  - –Each processor operates on different data.
  - –Some processors may "idle" during a sequence of instructions.
- MIMD
  - –Each processor runs its own instruction sequence.
  - –Each processor works on a different part of the problem.
  - –Each processor communicates data to the others.
  - –Some processors may have to wait for the results of processes being performed by other processors or for access to data being used by other processors.

SIMD machines use a single instruction to act on many sets of data simultaneously. This architecture, also called an array processor, features one control unit, multiple processors, multiple memories, and an interconnection network. The control unit broadcasts instruction to all the processors, but only active processors execute the same instruction at the same time using the data taken from their local memory.

MIMD machines consist of multiple processors with either multiple memories (distributed) or shared memory, in which each processor can follow an independent instruction stream. Whereas many tightly coupled multiprocessors, such as the Encore Multimax multiprocessor, use shared memory as a major means of communication between processors, the Intel's iPSC/860 and nCUBE's nCUBE2 are loosely coupled distributed memory multicomputers and employ the message passing communication mechanism. The shared-memory MIMD allows the use of conventional programming methods with which the user or compiler does not need to worry about the location of data. For example, one can take conventional FORTRAN code and obtain concurrency from parallel execution of DO loops automatically, with user directives, or both. In contrast, distributed-memory MIMD machines require the communication among the processors to be made explicitly by the user in programming.

Several parallel computers and their key features are summarized in Table 1. Of course, this is only a small portion of existing systems, not a comprehensive list. A more detailed review of parallel processing systems can be found in work by Miller et al. (*8*). A thorough review of SIMD machines has been given by Hord (*9*).

Comparison of Programming Methods

To run efficiently in a parallel environment, a sequential application must be partitioned or decomposed into subsets. This involves dividing the data or program code (or both) among the available or allocated processors. It may also in-

TABLE 1 Examples of Parallel Computers

Manufacturer	Model Number	Program control	Maximum number of processing elements	Topology	Interprocessor communication	Processor technology	Maximum memory capacity	Peak performance	Operating system	Language supported
Alliant (*10*)	FX/2800	MIMD	28	Crossbar and bus	Shared memory	Intel i860	1 GBytes	1000 MFLOPS	Unix	Fortran-77 Pascal C
BBN(*11*)	TC2000	MIMD	512	Butterfly switch	Message passing	Motorola 88100	1 Gbytes	1260 MFLOPS	Unix	Fortran C
Encore(*12*)	93	MIMD	32	Bus	Shared memory	Motorola 88100	640 MBytes	128 MFLOPS	Unix	Fortran Pascal C
FPS(*13*)	System 500	MIMD	84	Bus	Shared memory	SPARC	1 Gbytes	6.7 GFLOPS	Unix	Fortran C
Intel(*14*)	IPSC/860	MIMD	128	Hypercube	Message passing	Intel i860	8 Gbytes	7.6 GFLOPS	Unix	Fortran C
NCUBE(*15*)	NCUBE 2 Model 80	MIMD	8,192	Hypercube	Message passing	Custom	512 Gbytes	27 GFLOPS	Unix	Fortran-77 C
Thinking Machine(*16*)	Connection Machine CM-2	SIMD	65,536	Hypercube	Message passing	Custom	8 Gbytes	10 GFLOPS	Unix	Fortran C

MFLOPS = million floating-point operations per second
GFLOPS = giga FLOPS

volve changing DO-loop limits, array dimensions, and subroutine parameters so that each processor can operate on a subset of data.

The differences between the SIMD and MIMD machines in this regard can be characterized by the two basic approaches in partitioning: control parallelism and data parallelism.

Control Parallelism

Control parallelism breaks up a standard program into more or less independent subsets of instructions and assigns one such subset to each processor. It is used by vector supercomputers as well as by the many MIMD computers.

Though the multitasking method allows all processors to work on different parts of the same problem, it has several disadvantages. For instance, it is difficult to scale a very large number of processors well into a massively parallel regime. The breakdown of the instruction set into independent subunits is usually possible only to a certain level of granularity beyond which no further division is possible. In most engineering applications, the number of such independent subunits is typically measured in tens. Furthermore, synchronization and load balancing between the different subunits often become difficult tasks in program design.

Data Parallelism

The hardware and software paradigms that can scale well into the massively parallel regime base the parallelization on a program's data rather than on its instruction stream. Most programs manipulate tens of millions of pieces of data; very few programs have tens of millions of lines of code. This approach is used mainly by the SIMD machines. In a data parallel program, a single instruction can affect all elements of a parallel data structure simultaneously. The same operation in a serial program, however, needs to be expressed as a loop and executed sequentially for each element of the array.

Features of Connection Machine

The Connection Machine, a massively parallel SIMD supercomputer, has been used in scientific disciplines such as structural mechanics, molecular dynamics, and image processing (*17*). It consists of up to 65,536 bit-serial processors, each with 1 Mbit of local memory, and 2,048 Weitek floating-point processors when fully configured (*16*). Every chip contains 16 processors, and each pair of chips shares a Weitek processor. The chips are connected in a 12-dimensional hypercube; the processors on each chip are connected in a 4-dimensional hypercube.

The CM-2 system consists of a parallel processing unit that contains thousands of data processors, a front-end computer, and an I/O system. The front-end computer broadcasts instructions to all processors in parallel. The instructions are broadcast through a sequencer that decodes the front-end instructions into a series of low-level microinstructions and broadcasts them to individual processing elements.

The CM-2 processors are interconnected by a high-speed communication device called a router. The router allows general communication in which processors can send data or receive data from any other processors in parallel. It also supports a faster but more structured form of communication called grid communication, which allows processors to communicate with their neighbors in a multidimensional grid.

General communication involves the concept of parallel left indexing (*18*). A parallel left index rearranges the elements of the parallel variable on the basis of values stored in the elements of the index. Graphical illustrations of these two operations are presented in Figure 2.

In this paper, the proposed modeling concept is tailored for the SIMD machines—in particular, the Connection Machine CM-2, which uses the data parallel paradigm. A detailed discussion of the simulation methodology on a MIMD machine is available elsewhere (*19*).

MODELING METHODOLOGY FOR TRAFFIC NETWORK SIMULATION

The massively parallel traffic simulation model is adapted from the macroparticle traffic simulation (MPSM) approach (*20*) with the addition of vehicle path-processing capability. It follows a fixed time-step logic and uses macroscopic traffic relations to approximate the prevailing speed in a given link. Vehicles are then moved individually through the network according to predetermined paths. Because each vehicle is simulated individually, the proposed model can certainly incorporate microscopic features such as car-following and lane-changing mechanisms in the simulation process. However, for

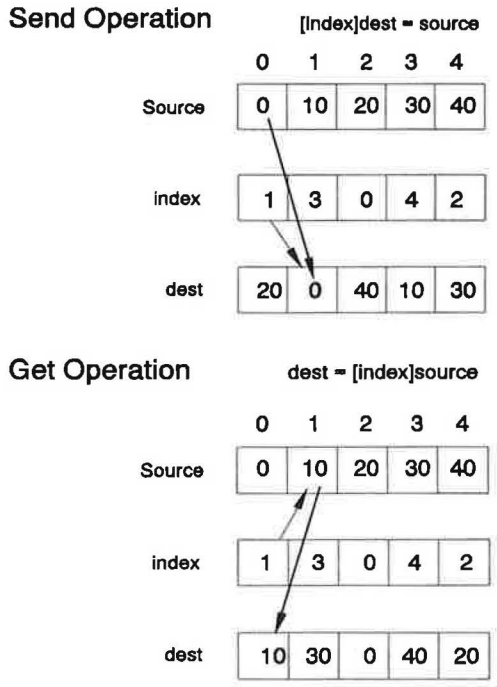

FIGURE 2 Examples of general communication.

simplicity of illustrating the data parallel modeling concept, only the simple MPSM logic is discussed.

Modeling Concept

There are three basic data entities for a real-time traffic simulation model: (a) urban streets and highways network, (b) traffic signal controls, and (c) vehicles. As an example, the network in Figure 3 (*top*) can be normally structured into a set of nodes and links as shown in Figure 3 (*bottom*). However, in an ATIS-ATMS application, the data structure for vehicle must be able to support the vehicle path-processing capability and to distinguish vehicles with and without access to in-board ATIS systems. The first requirement is achieved by explicitly embedding a predetermined path into each vehicle for a given origin-destination (O-D). These paths can be stored as a series of links or turning movements at each intersection. The second requirement is easily accomplished since each vehicle is simulated individually. Equipped vehicles will periodically update their paths on the basis of the optimal results of a real-time route assignment model, whereas the unguided vehicles will essentially follow predetermined paths.

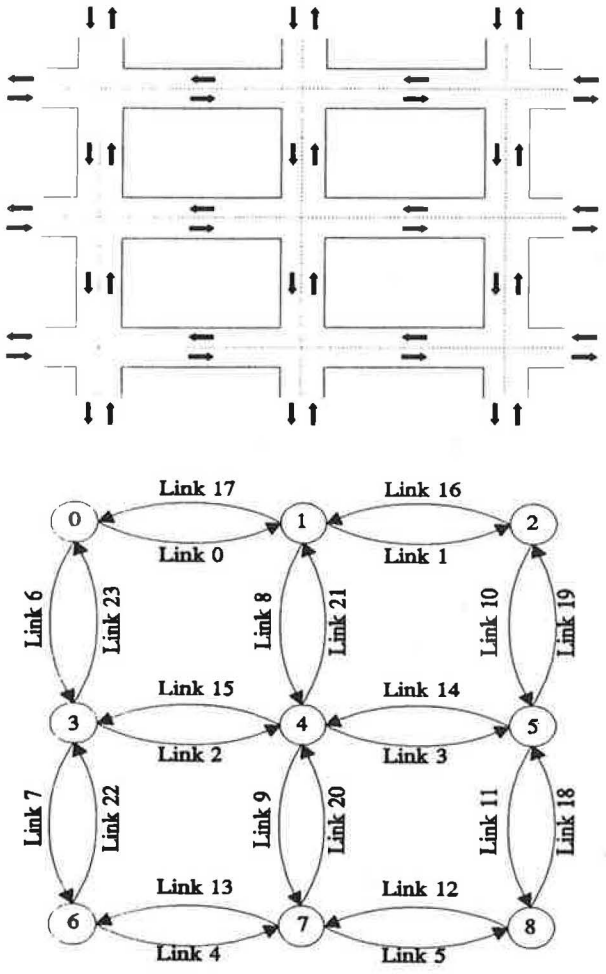

FIGURE 3 Example network: (*top*) street network; (*bottom*) representation of network in nodes and links.

The integration with a real-time dynamic assignment model, however, is beyond the scope of this paper, which concentrates on illustrating the massively parallel traffic simulation concept. Thus, all vehicles will be considered as unguided and will follow predetermined paths throughout the simulation period.

One of the most important aspects of data parallel programming is the choice of parallel data structure, since good data organization can significantly simplify computations and interprocessor communications. The aforementioned data entities can be structured as parallel variables so that an operation can be applied to all data elements simultaneously.

Description of Principal Model Components

There are several ways to organize these data entities as parallel variables. The parallel variables presented here are simple, yet they can be used in more complex microscopic models that use car-following and lane-changing mechanisms. Each set of data entity—vehicle, link, and node—is kept as a separate set of parallel variables.

Vehicle Parallel Variable

The vehicle parallel variable (VPVAR) is shaped as a one-dimension parallel variable with *NV* positions, where *NV* is the maximum number of vehicles to be simulated in the network at any time slice. Currently, the Connection Machine requires the number of positions for parallel variable to be a

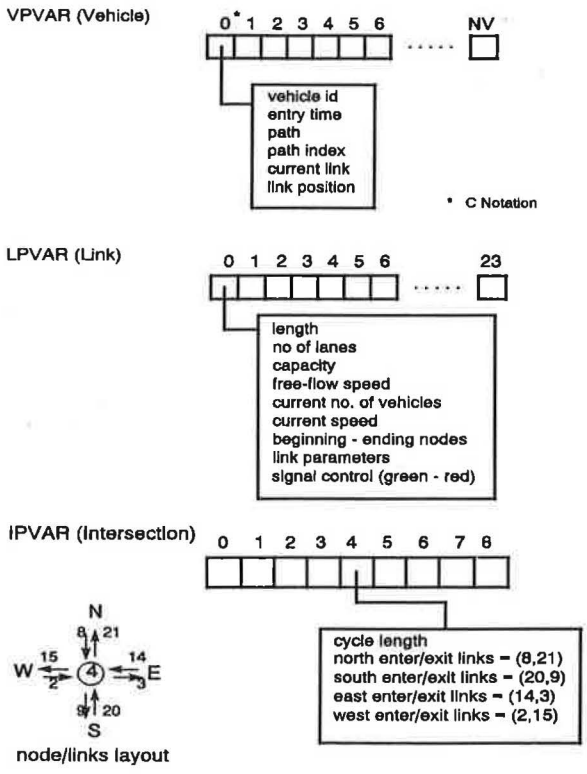

FIGURE 4 Graphical illustration of parallel variables.

power of two and must be some multiple of the number of physical processors: *NV* for the CM with 8,192 (8k) processors can be 8,192, 16,384, and so on. Other vehicles not yet entered in the network are kept in the front-end computer and sent to VPVAR at preset intervals. Each element in the VPVAR keeps track of one vehicle's key characteristics in the network [Figure 4 (*top*)], including its path, location, speed, and scheduled departure time. The scheduled departure time is used to determine a vehicle's entry time to the network and the entry order of vehicles in each link.

At each time step, a vehicle moves along a link by some distance that depends on the link's prevailing speed and the time increment. The time-dependent speed can be either computed with the embedded speed-density function or governed by a car-following mechanism. Once the vehicle reaches the end of a link, it will be moved onto the next downstream link in a path toward its destination. Upon arriving at its destination, the vehicle is removed from the network.

Notice that moving vehicles through the network requires the VPVAR to communicate with the link parallel variable via general communication (*18*). The first involves the send-with-reduction operation, which combines communication and computation. Each vehicle in the network sends a signal to its current link in parallel. These signals are then combined for each individual link, which is equal to the number of vehicles currently traveling in its link. Such information allows each link processor to compute the new prevailing speed with an embedded speed-density function. The second communication step is the get operation, in which each vehicle receives the speed information from its current link and updates its current position accordingly. Once a vehicle reaches the end of its current link, it will attempt to change to a downstream link in its path provided that it has not reached its destination. To determine whether such an action is possible, each vehicle in the VPVAR will use the get operation to communicate with the link parallel variable (LPVAR) in order to check whether the signal controlling this link is green or red and whether the downstream link volume is under capacity or not. The vehicles that can satisfy both conditions can then update their current link information.

The path structure of each vehicle has been developed to take advantage of typical vehicle movements through the network. In general, vehicles will travel along the same street in one direction for several blocks, turn left or right, and travel several more blocks in the same fashion. The number of turns that each vehicle makes tends to be relatively small in comparison with the number of links traveled. Suppose that we limit the number of turns that each vehicle can make to x; then we need to keep only x pairs of numbers. The first number in a pair corresponds to direction: north, south, east, or west. The second number corresponds to the number of links to be traveled in this direction. The trade-off required for this type of path structure is for the node parallel variable [Figure 4 (*bottom*)] to have indexes of entry-exit links for four directions.

Link Parallel Variable

The LPVAR's main function is to compute current speed according to the number of vehicles currently in the link using the macroscopic speed-density relationship. Each element in the LPVAR contains information such as link type, capacity, number of lanes, free-flow speed, and speed-density function parameters. Because it has been declared as a parallel variable, the operations for computing current speeds can be applied to all LPVAR elements simultaneously.

Using the network shown in Figure 3, the corresponding LPVAR and the intersection parallel variable (IPVAR) can be constructed as shown in Figure 4 (*middle*, *bottom*). The LPVAR is shaped as a one-dimension parallel variable.

Intersection Parallel Variable

The IPVAR is shaped as a one-dimension parallel variable, each element corresponding to an intersection and containing information on the cycle length and entry and exit links for four directions to each node [Figure 4 (*bottom*)]. At each time increment, the IPVAR updates the signal settings for all incoming links by communicating with LPVAR simultaneously using entry links as index for parallel left index operation. In this paper, only pretimed signals are modeled for the network. However, it can be extended to actuated signals by including an additional send operation from the VPVAR whenever a vehicle crosses the detector location.

Parallel Logic Flow Chart

Figure 5 illustrates the real-time traffic simulation logic. Notice that it can be used not only on a Connection Machine, but also on other SIMD machines with minor modifications. Basically, compared with the need of using three nested loops in sequential computers, the proposed method contains only one time loop and three stages of execution. In the first stage, parallel variables for vehicles, links, and intersections are initialized from external files. These external files include (a) traffic demands generated from an O-D matrix of individual vehicles with predetermined paths, and (b) network information of nodes, links, and signal control. The second stage involves the main simulation routine (Figures 6–8), which consists of four steps: updating signals, counting vehicles and updating link speeds, moving vehicles, and updating vehicles. A summary of simulation statistics is generated in the last stage.

The first step in the second stage is to update signal settings at all incoming links at all intersections in parallel. As shown in Figure 4 (*middle*), each link element contains signal information of start green and red time. At each step, start-red is checked against the system clock. If the start-red is less than the system clock, both start-green and start-red are incremented by cycle length from the IPVAR using the destination node as the index in the parallel left index operation.

The second step of the main simulation loop is to update the new prevailing speed for each link element in the LPVAR. Figure 6 shows how to compute each link element's volume and new prevailing speed at each time increment simultaneously. In Figure 6 (*top*), each vehicle element already in the network sends a signal to its current link element. Each link element then uses such information along with the speed-density function to compute the new prevailing speed as shown in Figure 6 (*bottom*).

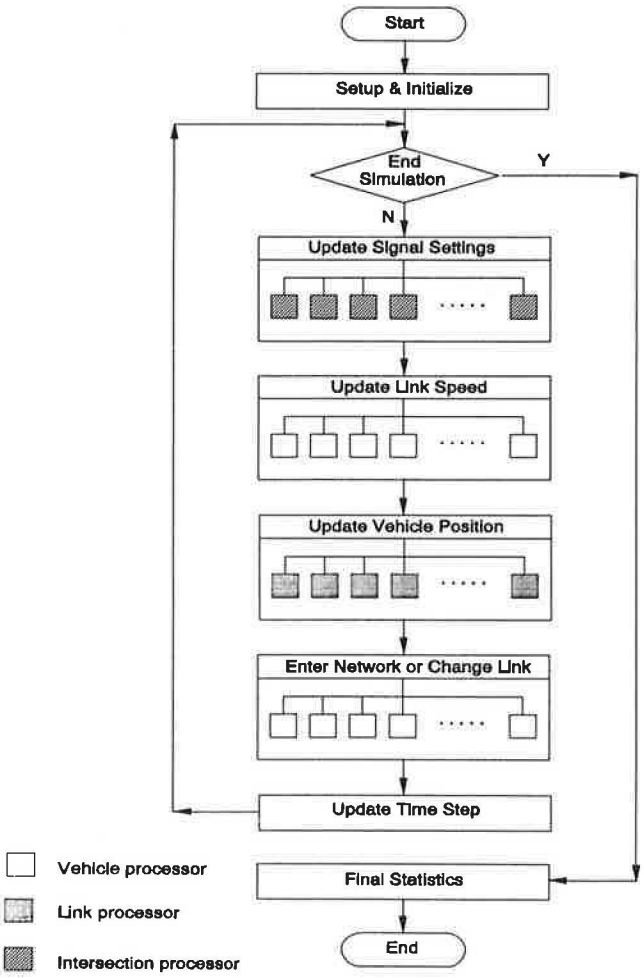

FIGURE 5 Logic flow chart for massively parallel simulation model.

Given the updated speed information, vehicles can then move to their new positions in the link with their updated speeds (Figure 7). In such a process, each vehicle element will use its current link field as an index in general communication (i.e., GET operation) and to receive its new speed information from the LPVAR. Vehicles that reach the end of the link are eligible for changing links in the next step; those that arrive at their destinations are removed from the network.

The next step is to enter vehicles into the network and to move them from link to link (Figure 8). As shown in Figure 8(*a*), it begins with a selection of vehicle elements that will be involved in the computation. These elements include vehicles that are about to enter the network and vehicles that have reached the end of links and intend to move along their own paths. All these identified vehicle elements will be activated and moved according to the signal control and the available link capacity. A graphical illustration of such a parallel moving process is presented in Figure 8 (*b*, *c*, and *d*).

Illustrative Example

In this example, although only the movement of one vehicle (i.e., Vehicle 3) will be presented, the same computation will be executed simultaneously for all vehicles in the simulation. Assume that Vehicle 3 is already in the network and has the following characteristics at time t:

Description	*Data*
O-D	Node 0–Node 8
Path	(1,2), (2,1), (1,2) (= Link 6, 2, 3, 11)
Current link	2
Path index	First link of second turn
Link position	1,100 ft
Link length	1,200 ft
Current speed	25 mph
Next intersection	4

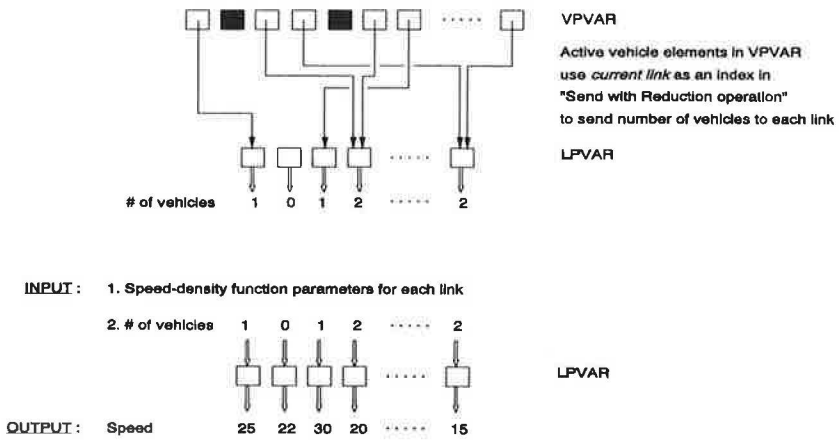

FIGURE 6 Graphical illustration of link-speed updating process: (*top*) each link "counts" vehicles currently in its link in parallel; (*bottom*) each link element computes the new speed in parallel.

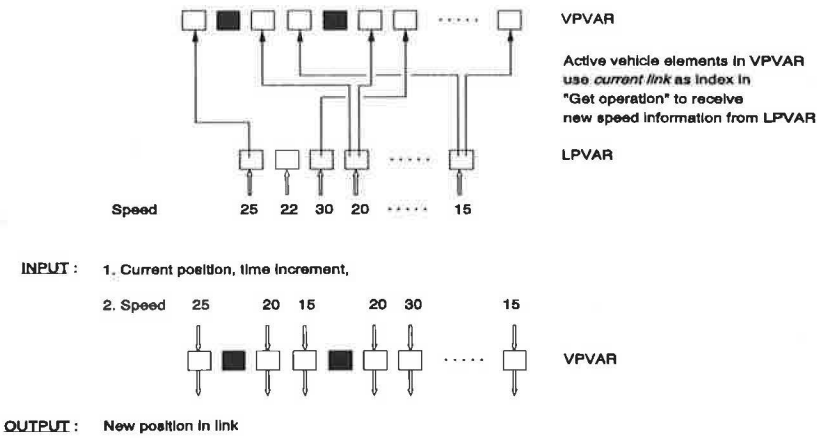

FIGURE 7 Graphical illustration of updating process for each vehicle element's position and speed: (*top*) active vehicle elements inquire new speed information in parallel; (*bottom*) active vehicle elements update link position.

The parallel simulation process from time t to $t + \Delta t$ is illustrated as follows:

For simulation time step t

Step	*Action*
Update signal settings	Signal setting information for each movement in all elements of IPVAR is updated.
Enter network or change links	The computation in this step affects those vehicles that are entering the network or changing links. Vehicle 3 is already in the network, so it will not be involved in the computation and will remain idle until the next step.
Update link speed	First, all vehicles that are in the network send signals to the current link using link-ID as an index in parallel left indexing. In this case, Vehicle 3 sends a signal to Link 2 along with other vehicles that are in Link 2. Link 2 then uses this information to compute the new prevailing speed.
Update vehicle position	Vehicle 3 receives the new average speed information from Link 2 and updates its new position. Suppose Vehicle 3 travels an additional 100 ft to reach the end of Link 2 in this time step, which makes it eligible to change to a new link in the next time step, $t + \Delta t$.
Update time step	Simulation time is increased by Δt.

For simulation time step $t + \Delta t$

Step	*Action*
Update signal settings	Signal setting information for each movement in all elements of IPVAR is updated.
Enter network or change link	Vehicle 3 is now eligible to change to its downstream link. First it uses general communication to get signal information from LPVAR using current Link 2 as an index in a parallel left indexing operation. It also uses downstream Link 3 as an index in a get operation to receive link capacity information from LPVAR. If traffic signal is green and there is no spillback, Vehicle 3 is allowed to move to Link 3.

The information currently stored in Vehicle 3 data elements will be

Description	*Data*
O-D	Node 0–Node 8
Path	(1,2), (2,1), (1,2) (= Link 6, 2, 3, 11)
Current link	3
Path index	Second link of second turn
Link position	0
Link length	1,500 ft
Current speed	20 mph
Next intersection	5

These models have been implemented on the Connection Machine in C* (*18*), an American National Standards Institute C-standard with parallel extension. Several simulation experiments have been carried out to test various factors affecting the running time on the Connection Machine. The results for these simulation experiments will be fully reported later (*21*). However, from our preliminary simulation experiments, we have been able to simulate 32,000 vehicles for 30 min at 2-sec increments within 2 min using the Connection Machine with 16,384 processors.

ONGOING RESEARCH ACTIVITIES

This paper presents our preliminary research in evaluating the applicability of massively parallel SIMD machines to simulate networkwide traffic in real time. The proposed model is part of the ongoing research to develop a real-time traffic simulation model for IVHS application. These research activities and related studies can be categorized into three areas: enhancement of SIMD models, development of traffic simulation methodology for MIMD machines, and integration with a real-time dynamic assignment model.

Main enhancements to the proposed SIMD model are to incorporate microscopic mechanisms such as car-following and lane-changing logic. Such logic can be added to the model using the same structure for parallel variables with some modifications. Further extensions of the model include the capability to model traffic incidents, lane closures, actuated signals, real-time surveillance systems, and ramp metering. The

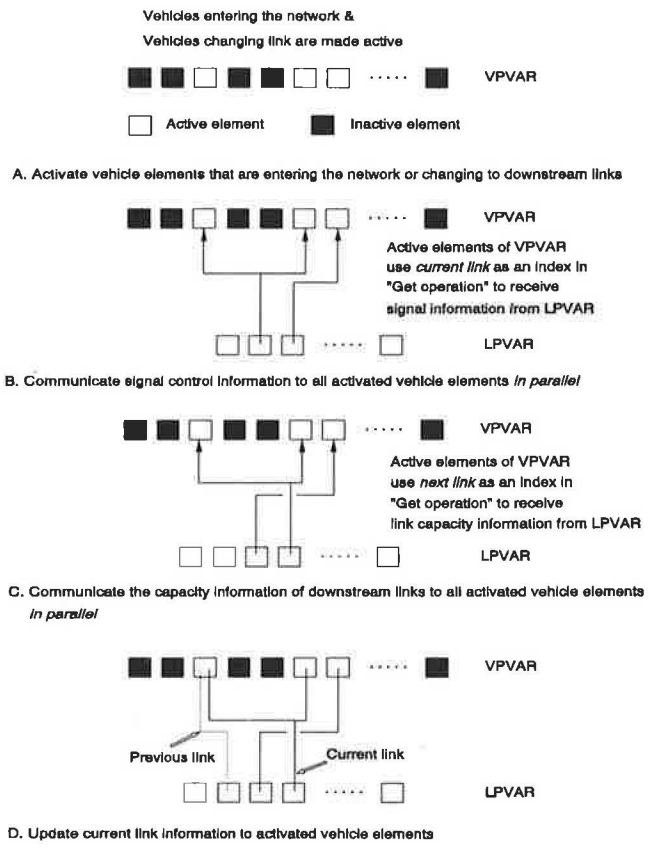

FIGURE 8 Graphical illustration of parallel vehicle moving process.

initial simulation experiments have shown that interprocessor communication constitutes the main fraction of running time in massively parallel computers. Various optimization techniques and data structure alternatives will be further explored and compared. Trade-offs between different data structures, programming methodology, and interprocessor communication need to be examined.

The second area of research involves the development of a traffic simulation model for MIMD machines. Many MIMD machines use control parallelism, which divides a standard program into more or less independent subsets of instructions and assigns one such subset to each processor. The traffic simulation model for control parallelism may involve dividing a program into vehicle, link, and traffic control subsets of instruction and assigning each subset to each processor, or dividing the network into several subnetworks and assigning each subnetwork to each processor. The main computation issues for the MIMD programming model are synchronization and load balancing among processors. An exploration of using the MIMD machines for real-time traffic simulation is being conducted in parallel with the development of SIMD model at the University of Maryland (*17*).

The third area of research involves integrating the real-time simulation and real-time dynamic assignment, rather than simply interfacing them. The integration is necessary for both models to operate efficiently.

ACKNOWLEDGMENTS

The contributions of Hani S. Mahmassani and Stavros Zenios are acknowledged. The authors would like to thank Jerry Sobieski of the University of Maryland Institute of Advanced Computer Studies for his valuable inputs and comments. This research is partially supported by a FHWA project.

REFERENCES

1. J. Lindley. Urban Freeway Congestion Problems and Solutions: An Update. *ITE Journal*, Vol. 59, No. 12, 1989, pp. 21–23.
2. G. Euler. Intelligent Vehicle/Highway Systems: Definitions and Applications. *ITE Journal*, Vol. 60, No. 11, 1990, pp. 17–22.
3. H. Mahmassani, S. Peeta, G. L. Chang, and T. Junchaya. *A Review of Dynamic Assignment and Traffic Simulation Models for ADIS/ATMS Applications*. Technical Report DTFH61-90-R-0074-1. Center for Transportation Research, University of Texas, Austin, 1991.
4. K. Hwang and F. A. Briggs. *Computer Architecture and Parallel Processing*. McGraw-Hill, New York, N.Y., 1984.
5. W. J. Karplus. Vector Processors and Multiprocessors. In *Parallel Processing for Supercomputers and Artificial Intelligence* (Hwang and DeGroot, eds.). McGraw-Hill, New York, N.Y., 1989, Chapter 1.
6. M. J. Flynn. Some Computer Organizations and their Effectiveness. *IEEE Transactions on Computers*, Vol. C-21, 1972, pp. 948–960.
7. T. Feng. A Survey of Interconnection Networks. In *Supercomputers: Design and Applications* (K. Hwang, ed.). IEEE Computer Society Press, 1984.
8. R. K. Miller and T. C. Walker. *Parallel Processing*. Fairmont Press, Lilburn, Ga., 1990.
9. M. R. Hord. *Parallel Supercomputing in SIMD Architectures*. CRC Press, Boca Raton, Fla., 1990.
10. *FX/2800 Product Summary*. Alliant Computer Systems Corp., Littleton, Mass.
11. *TC2000 Technical Summary Rev. 2.0*. BBN Advanced Computers, Inc., Cambridge, Mass., 1989.
12. Product Overview. ENCORE Computer Corp., Fort Lauderdale, Fla.
13. *System 500 SPARC Supercomputer Product Overview*. FPS Computing, Beaverton, Oreg.
14. *iPSC/860 Technical Summary*. Intel Corp., Beaverton, Oreg., 1990.
15. *Technical Overview: System*. nCUBE, Beaverton, Oreg.
16. *Connection Machine Model CM-2 Technical Summary*. Thinking Machines Corp., Cambridge, Mass., 1991.
17. *Proc., Conference on Scientific Applications of the Connection Machine* (H. Simon, ed.). World Scientific, Teaneck, N.J., 1989.
18. *The Connection Machine System, Programming in C**. Thinking Machines Corp., Cambridge, Mass., 1990.
19. G. L. Chang. *Development of A Dynamic Real-Time Traffic Simulation Model with a MIMD Computing Structure*. Waking Paper. Transportation Studies Center, University of Maryland, College Park, 1992.
20. G. L. Chang, H. S. Mahmassani, and R. Herman. A Macroparticle Traffic Simulation Model To Investigate Peak-Period Commuter Decision Dynamics. In *Transportation Research Record 1005*, TRB, National Research Council, Washington, D.C., 1985, pp. 107–121.
21. T. Junchaya and G. L. Chang. *Exploring Real-Time Traffic Simulation with Massively Parallel Computing Architectures*. Working Paper. Transportation Studies Center, University of Maryland, College Park, 1992.

Publication of this paper sponsored by Task Force on Advanced Vehicle and Highway Technologies.

Standards for Intelligent Vehicle-Highway System Technologies

JONATHAN L. GIFFORD

Alternative approaches to technological development for intelligent vehicle-highway systems (IVHSs) were investigated by reviewing the standards literature and interviewing key individuals. The standards literature suggested that for certain technologies, market forces can sometimes lead to suboptimal de facto standards, which would support government intervention to protect the public interest. There may be only narrow windows in time during which government or other collective action to establish such standards can be effective at reasonable costs. The greatest power to influence standards setting, however, may come exactly when the information available to inform action is most limited. Market pressure to disseminate a technology sometimes argues for an imperfect standard in a timely fashion over the alternative of no standard at all, but the sheer complexity of technical and marketing issues may confound and extend the duration of the standards-setting process. For automatic vehicle identification (AVI) technologies, concerns about suboptimal de facto standards may be misplaced, because those selecting the technologies are not mass-market end users but large-scale monopoly service providers. However, market pressures from such users to disseminate AVI technology are acute and may overwhelm standards-setting procedures. For other IVHS technologies, concerns about suboptimal de facto standards may also be misplaced, because the more fundamental issue of what end users are willing to pay for remains largely unresolved. Technological and market uncertainty and complexity may severely impede and extend the standards-setting process.

There is a broad consensus among transportation experts that the successful deployment of intelligent vehicle-highway systems (IVHSs) in the United States depends critically on the early development of technological standards and that success is much less dependent on major technological breakthroughs (*1*,*2*). Systems necessary for various aspects of IVHS are available and currently being pilot-tested in this country. A significant barrier to the dissemination of IVHS in the United States, however, is seen to be the absence of national standards for these existing and developing technologies. The supposition is that such standards must precede deployment and that some consensus-oriented cooperative effort is needed to develop such standards. Debate centers on the appropriate procedures for establishing these consensus standards.

The view that standards are a necessary precursor to the dissemination of IVHS is not self-evident, nor is the view that consensus is the appropriate mechanism for developing standards. Other complex technological systems have developed without early consensus standards. Indeed, there is some cause for concern that early development of IVHS standards may prematurely lock the technology into formats that are inefficient in the long term. Alternative technological development approaches, such as rivalry between competing manufacturers' systems and formats, may yield the greatest consumer benefits.

To investigate alternative approaches to technological development as it applies to IVHS, the author and a research assistant surveyed the literature on IVHS standards and on standards and technological development and interviewed key figures who are active in the development of IVHS standards. This paper reports the results of that research.

Department of Public Affairs, George Mason University, Fairfax, Va. 22030.

DESCRIPTION OF IVHS TECHNOLOGIES

IVHS embraces a broad range of technologies that incorporate advanced communications and control into the operation of highway vehicles and infrastructure. Briefly, the applications fall into several major areas, as indicated, although exact terminology is somewhat fluid.

1. Advanced traffic management systems (ATMSs) focus on traffic control devices such as conventional traffic signals and newer technologies such as changeable message signs as well as vehicle detection and monitoring.
2. Advanced traveler information systems (ATISs) provide drivers or transit users with travel information such as route selection, navigation, congestion, and delay. Transit applications are sometimes referred to as advanced public transportation systems (APTSs).
3. Commercial vehicle operations (CVOs) focus on improving the management of commercial fleets by enhanced vehicle identification and tracking.
4. Advanced vehicle control systems (AVCSs) focus on systems that automate driving, either by enhancing information available to the driver through, for example, radar detection of obstacles in a car's "blind spot," or by replacing driver control with automated control, at least for some portion of a trip.
5. Automatic vehicle identification (AVI) is a system whereby a vehicle carries a small identification device that allows roadside mechanisms to identify each vehicle uniquely. In the highway domain, AVI has been used to identify properly equipped vehicles as they cross certain points on the highway, without requiring action by an observer or the driver. AVI technologies can be used for many transportation applications, including electronic toll collection and vehicle monitoring. (This paper treats AVI as a part of IVHS, although some definitions do not.)

More-detailed descriptions are widely available (*3*–*5*).

STANDARDS AND TECHNOLOGICAL DEVELOPMENT

Technology standards have been the subject of research and development for more than a century. During the 19th century, they played an integral role in the development of the American system of manufacture, since standardized parts were essential to mass production (*6*). Since then "private organizations have developed tens of thousands of standards that serve to coordinate the productive efforts of American businesses" (*7*). During the first half of the 20th century, researchers examined the relative impact of standards on the efficiency of production and the social implications of a standardized society (*8–14*). Contemporary research has begun to focus on the relationship between standards and the technological development.

Standards perform a variety of functions: (a) a compatibility function, whereby a standard ensures the compatibility of complementary products from different manufacturers; (b) an informational function, whereby a standard informs the market about the characteristics of a standard product; (c) a quality function, whereby adherence to a standard indicates some level of quality (including a regulatory or safety standard); and (d) a variety-reduction function, whereby a standard allows the reduction in variety of a set of products (e.g., screw sizes) with little or no loss of consumer utility and producer gains in reduced production costs and lower inventory costs (*15*).

Technological standards are typically developed in one of three ways: through a government regulatory process resulting in mandatory standards; through a consensus process undertaken by standards-setting groups resulting in voluntary consensus standards; or through competitive rivalry between different technologies eventually resulting in one or more de facto standards. The U.S. public policy stance on standards, especially in the last decade, has generally been "to avoid mandatory standards, but . . . encourage . . . the formation of widely representative committtees to write voluntary technical standards . . ." (*16,17*).

Standardization is sometimes beneficial because it can lead to "cost savings through economies of scale" and the "lowering of entry barriers." In such cases, early standardization is probably more desirable than late, if the same standard is set (*18*). Early standardization also removes the incentive for potential users of the technology to "wait for the standard to settle down, and thus encourages early adoption of the technology" (*19*).

But there are also reasons to wait to establish standards for existing yet continually advancing technologies. Information on advances will continue to flow, information that may modify the view of the optimal standard to be established. In 1961, for example, IBM promoted its 6-bit computer code as a U.S. standard, but rapid technological change led it to shift its support to an 8-bit code only 4 years later (*20*).

Standards condition the rate and direction of technological development. When the technology is advancing, market forces can cause de facto standards to emerge in the absence of public intervention. Moreoever, historical chance events can exert powerful influences over those de facto standards and give rise to less-than-optimal standards. A striking example is the almost universal "QWERTY" keyboard, which came into predominant use not through any particular technical superiority but instead through a series of historical accidents (*21*).

Such research has developed the notion of "path dependency," which suggests that market forces, left to their own devices, may not yield technically or economically superior de facto standards (*22,23*). Path dependency, as a market failure, provides a rationale for government intervention into market processes in order to protect the public interest.

Two critical policy dilemmas emerge from such conditions. First, public policy interventions to affect standards may be effective or affordable only during a narrow window in time. And second, government's greatest power to influence the path of technological development may come at just the time when the necessary information on which to base such decisions is lacking. But adopting a wait-and-see policy runs the risk of locking into an inappropriate standard (*16*).

These policy dilemmas are central to the topic of this paper, for it is not at all clear whether resolving those dilemmas through consensus-based standards development will produce outcomes superior to those yielded by competitive rivalry. QWERTY is a case in which rivalry yielded lock-in on an inferior technology. But rivalry can also yield tremendous innovation (*24*).

Consider, for example, the competition over the last decade for dominance in the microcomputer market between DOS-based systems (developed by IBM and Microsoft), Apple's Macintosh system, and UNIX. Recent developments suggest that IBM and Apple will now join forces to create a system that synthesizes the benefits of both systems. Further study of this development process is clearly in order, but there appears to be at least an arguable case that competitive rivalry drove both parties to improve their own systems to a greater extent than they would have had the two joined forces in the early 1980s to produce a consensus standard.

What is clear is that both laissez-faire and policy intervention involve risks. Rivalry risks the emergence of de facto standards that are technically inferior. A consensus approach risks diminishing the incentives for innovation.

Another major avenue of research inquiry has been in health and safety standards, specifically in the appropriate role for government in setting standards and the extent to which the public interest is served by reliance on private voluntary standards. Beginning in the mid-1960s, private voluntary standards in several industries came under intense scrutiny. In the automobile industry, they were prompted by Nader's *Unsafe at Any Speed* (*25*), which led to the National Traffic and Automotive Safety Act in 1966. Similar concerns led to the Gas Pipeline Safety Act of 1968 and the formation of the Consumer Product Safety Commission in 1967. Research in this area has focused on how to conjoin the democratically motivated consideration of the public interest—especially the interests of consumers, workers, and small businesses—with the experience and expertise provided by the private standards-writing organizations, which governmental agencies cannot readily duplicate (*7,26–28*).

These concerns have raised the level of interest in standards as a general area of inquiry (*29*). Researchers have also focused on investigations of standards in various substantive areas, including communications and computers (*20,30*), housing (*31,32*), highway design (*33,34*), land surveying and

ownership (*35*), agricultural technology (*36*), and electrical supply (*37*).

These case studies contain some particularly relevant conclusions and generalizations for inquiry into IVHS. For packet switching standards for computer communications, economic and competitive pressures were forcing the rapid implementation of computer networks, with or without standards. As a result, if standards were going to contribute to the technology, they had to be developed on a compatible time scale. In such cases, the study concluded, an imperfect standard developed in a timely fashion is better than no standard at all, and the best time to develop a standard "appears to be during a very narrow window" after there has been some operating experience with a particular technology and "when there has been a commitment by other organizations to enter the field, but before these same organizations" commit themselves to divergent approaches (*20*).

In the standardization of computerized local-area networks (LANs), the "sheer complexity of the issues" surrounding the development of LAN strategies and standards meant that few engineers, if any, understood all the technical and marketing issues involved. Most of the participants in the standards-setting group conceded they were "there to learn rather than support any particular position." But even though there was a desire to reach a standard, the process became lengthy as group members struggled to understand the issues and the various arguments being presented. The LAN situation also indicated that a standard adopted before the technology has gained significant market experience leads to very lengthy standards that attempt to accommodate many options, because it is not clear ex ante what functions and formats will satisfy market preferences (*30*, p. 20).

The difficulty, then, lies in treading the narrow path between developing standards that adequately serve the public interest—soon enough to effect dissemination of the technology, but not so soon as to lock into inferior technology—all the while working in a domain that is fraught with complex technological and marketing issues that themselves are highly uncertain, indeed most uncertain, at the time the decision should be made.

IVHS STANDARDS

An abundance of literature addresses potential IVHS technological applications such as AVI, route guidance systems, and vehicle sensing and control strategies, but little material focuses specifically on IVHS standardization and technological compatibility issues. Much of the literature acknowledges the importance of system and technology standards and protocols for successful implementation, but, with few exceptions (*38*), most does not focus on standardization specifically (*39–41*).

The central standardization issues fall into three categories: timing, content, and process and participation. The timing issue turns on when it is appropriate to establish standards. "Standardization needs to be viewed in the context of an overall process of system design," and even the most mature IVHS application, ATMS, "has not yet reached the stage in its development that the system design trade-offs are understood." It is not advisable to wait until all "system implications" are understood and all "technical uncertainties are resolved" to initiate the standards development process, but it is necessary to determine the physical media and network topologies for the IVHS functions before establishing comprehensive standards (*42*, p. 15).

The content of standards obviously varies for each technology application, such as AVI. One general content issue is the question of performance versus design standards. Generally, design standards are seen to be inferior to performance standards since they tend to be more restrictive to innovation, but they are more difficult to develop (*43*). A more difficult aspect of the technical content of standards is the speed of events as of this writing. Multiple committees are meeting, establishing scopes and charges, creating task forces, and such. A general report on where matters stand would therefore be outdated almost immediately, and a more detailed discussion of the technical content issues for each standards-setting effort is beyond the paper's scope. Hence, this paper does not focus on technical content.

Although still somewhat fluid, process and participation issues appear to be stable enough to merit description. There is wide recognition of the need for a process to establish and coordinate national, and potentially international, IVHS standards and protocols (*44*). In separate studies recently completed for Congress by the General Accounting Office and the U.S. Department of Transportion (DOT), both agencies recommended the development of a national cooperative effort for the identification of technical standards (*1*,*45*). DOT's report also indicated that this effort should provide the forum not only for identifying the areas in need of technical standards, but also for deciding on the necessary standards and protocols as well. But the exact relationship between the public and private sectors is still in question (*40*), and even within the public sector, there is substantial disagreement over the respective roles of local, state, and federal governments (*46–49*).

There also appears to be relatively wide agreement that such a process be based on a voluntary consensus approach. A communications standards workshop held in June 1990 identified as its highest-priority action item the establishment of an IVHS Standards Oversight Committee with accreditation from the American National Standards Institute (ANSI). ANSI accreditation would ensure adherence to such various procedural protocols as open meetings and dispute handling (*50*). "[T]his Committee would observe, track, and coordinate all IVHS standards activities in the U.S., regardless of the originating organization. . . ." Workshop participants also recommended that IVHS standards be developed by existing standards-making organizations and coordinated by a newly established oversight committee (*2*, p. 13).

The Intelligent Vehicle Highway Society of America (IVHS AMERICA), incorporated in July 1990 as a nonprofit, public/private association, was a response to the desire for an organization to direct, coordinate, and provide structure for all IVHS efforts in North America, including standards-setting. It is anticipated that DOT will use it as a formal advisory committee on IVHS matters (subject to the provisions of the Federal Advisory Committee Act, 5 U.S.C. App.). As one of its organizational responsibilities, IVHS AMERICA will help identify needed standards, specifications, and protocols.

IVHS AMERICA has created a Standards and Protocols Committee, which will act in an oversight and coordinating capacity for all U.S. IVHS standards activities. It will function

as a "clearinghouse between requirements and standards-developing organizations" to identify areas in which technological standards are needed and to enlist the help of the appropriate voluntary standards-making organizations. It will not operate as a standards-setting group but will work to become ANSI-sanctioned (R. Weiland, author's files, 1991).

The actual development of the standards in the United States appears to be falling to four organizations: SAE, IEEE, AASHTO, and ASTM (G. Euler, W. D. Toohey, R. Weiland; personal communication; 1991).

SAE will most likely develop IVHS vehicle and human factors standards. Currently, SAE has a Database Standards Task Group working as a part of its Navigation Aids Subcommittee. This group has been working toward map data base standards for in-vehicle navigation systems for about 1 year. In addition, SAE recently formed an IVHS division under its existing Standards Board, which will "provide for the development and maintenance of SAE Standards, Recommended Practices, and Information Reports so as to aid the manufacturer in design consistency of vehicles and equipment that fall within the scope of the IVHS Division and to provide guidance and input to the IVHS AMERICA Standards Committee to coordinate harmonized national and international IVHS standards, protocols, and systems" (W. D. Toohey, personal communication, June 1991).

IEEE will focus on developing communications and electromagnetic technology standards. IEEE recently created a Standards Coordinating Committee, which will cooperate closely with the Standards and Protocols Committee of IVHS AMERICA to write standards in the communications and electromagnetic technology areas in response to requests from IVHS AMERICA (J. May, personal communication, July 1991).

AASHTO will most likely become involved in standards for technologies that affect highway facilities and the overall highway infrastructure. The AASHTO committees that developed the current roadside, geometric, and pavement design standards will have a substantial interest in the standards-setting process for IVHS technologies that will affect existing highway infrastructure standards. AASHTO currently has a temporary Special Committee on Transportation Systems Operations, which, among other responsibilities, tracks IVHS activities and the potential needs for infrastructure standards. This special committee, which reports to the Standing Committee for Highways, was established in December 1988 and has a 5-year temporary charter. As IVHS systems and technologies mature, AASHTO will make more permanent organizational decisions in terms of how to handle standards development in its areas of expertise (D. J. Hensing, personal communication, July 1991).

AVI STANDARDS

One component of IVHS—AVI, which is used in toll collection—uses a technology that extends well beyond IVHS and overlaps with standards-setting activities in several other areas. The U.S. Department of Defense is developing an accounting application of this technology for identifying aircraft during refueling and for freight container identification. In the refueling application, for example, a fuel truck could identify an aircraft during refueling for a potentially paperless transaction (J. Carnes, personal communication, July 1991). Working on a much smaller scale, Hughes has developed a 3/8-in. long by 1/16-in. diameter transponder for injection into fingerling salmon. The transponder allows the unique identification of each fish that returns upriver to spawn (D. S. Fleming, personal communication, Aug. 1991).

Standards for this technology are developing rapidly on several fronts. IVHS AMERICA's Committee on Standards and Protocols has established a subcommittee for AVI, and SAE's IVHS division has established a committee on AVI. Also, the trucking industry is experiencing rapid and extensive innovation and experimentation with communications technologies that may overlap with the development of AVI (*51*).

At the subnational level, several states are developing or have developed specifications for procurements that include the technology. California is developing compatibility specifications for AVI systems for electronic toll collection in the state. Several state agencies anticipate using the technology, including the Department of Transportation, which operates several toll roads; the Golden Gate Bridge Authority; and the Transportation Corridor Agencies (which are developing three toll roads in Orange County). The specifications define the "compatibility requirements for AVI equipment to insure that one transponder will operate at all future AVI facilities" in California. Once developed, the state intends to promulgate the specifications as administrative regulations (L. Kubel, personal communication, 1991).

The Virginia Department of Transportation (VDOT) has also established AVI specifications for toll collection. The specifications are not for statewide systems; they are part of a procurement process for an automatic toll system on the Dulles Toll Road in Northern Virginia. VDOT sees the Dulles Toll Road project in part as a proving ground for automated toll technology in the state. If the system operates successfully, it will more than likely be implemented elsewhere in Virginia (*52*). Several other state-level efforts are under way, as well as a coalition of New York and New Jersey that has agreed to use compatible AVI technology (L. Kubel, L. F. Yermack, personal communication, 1991–1992). One recent study identified operational or expected AVI activities in 22 locations (*53*).

With respect to non-IVHS applications, the Computer and Business Equipment Manufacturers Association recently created an ANSI-accredited standards committee for Non-Contact Information Systems Interface (called X3T6). Its purpose is to develop a non-contact interface between computer devices for the transfer of information. The committee will review "current technology in radio frequency data/communication, infrared and similar non-contact data transfer technologies with the objective of standardizing the interface between like devices." Although the technical committee will develop the standard for U.S. activities, the committee eventually intends to submit it to ANSI for approval as an international standard. The committee is open to all potential identification device applications (J. Carnes, 1991; author's files, March 1991).

The International Standards Organization is also developing standards for identification devices and recently adopted International Standard 10374, Automatic Equipment Identification (M. Bohlman, author's files, May 1991). Although it was not developed specifically for highway applications in potential AVI systems, it may well influence the groups now

aligning themselves to undertake AVI standards efforts and those firms that develop and implement automatic identification devices.

As of this writing, attempts to coordinate these various parties are moving forward. IVHS AMERICA's Committee on Standards and Protocols convened a special coordinating meeting on AVI standards in October 1991 intended to coordinate the development of specifications for North America and to "encourage restraint regarding the implementation of standards with less than continental scope . . ." (author's files, Oct. 1991). ASTM appears to be leading the effort for AVI standards, although some concern has emerged over whether it would adequately incorporate the views of trucking interests. At the same time, Virginia, California, and the New York–New Jersey coalition all have issued or will soon issue such specifications.

The case of AVI standards is interesting not only on its own merits, but also insofar as it can enlighten consideration of standards for other IVHS technologies. Two scenarios for the development of AVI standards capture the range of possibilities. Under the first, the laissez-faire scenario, states and operating agencies would promulgate specifications for their jurisdictions along the lines of ongoing efforts in California and Virginia or in multistate coalitions, as in New York and New Jersey. In all likelihood, these specifications would be incompatible, so that participating vehicles would require separate transponders for each jurisdiction in which they routinely operated. Standards-setting efforts, under the auspices of IVHS AMERICA or another organization, would likely produce a standard somewhat later, perhaps in 2 years. Such a standard would be informed by the experience of the lead states, and states implementing systems after the standard was available would likely adhere to it. Further, lead states might procure equipment consistent with the standard in a later replacement of their original equipment, or "gateway technologies" might be developed that would allow lead states' equipment to read standard transponders.

Under this scenario, lead states would reap the benefits of the technology while the standard was being developed, benefits that would be entirely foregone if they waited for the standard before implementing their systems. Popular estimates suggest payoff periods of less than a year, so that lead states' equipment would have paid for itself before the standard was even ready. Further, lead states would reap benefits even if the standards-setting process became bogged down and did not yield a standard for several years. On the other hand, lead states might resist converting their equipment to be consistent with the subsequently developed standard, thereby raising costs for multijurisdictional users.

Under a second scenario, operating agencies would defer procurements and specifications and participate in a consensus standards-setting process. Once the standard was developed, lead states would move forward to deploy systems consistent with the standard. Under this scenario, a single transponder would suffice for all jurisdictions, and users would be presented with the lowest costs. On the other hand, the standards-setting process would probably require approximately 2 years according to popular estimates, perhaps longer. In the meantime, no benefits of the technology would accrue, and such benefits would be permanently foregone—they could not be recaptured later.

From the standpoint of AVI, the laissez-faire scenario appears to be materializing, as several states issue their own specifications. The success of standards-setting efforts remains to be seen, but the ASTM initiative is promising. From the more general standpoint of IVHS, however, some conclusions are clear.

CONCLUDING REMARKS: ALTERNATIVE IVHS DEVELOPMENT PATHS

From the information collected in this research, a clear framework for the development of IVHS technologies appears to be emerging. IVHS AMERICA's Committee on Standards and Protocols will seek to coordinate the efforts of private standards-developing organizations such as IEEE, SAE, AASHTO, and ASTM. This framework is quite distinct from its major alternatives, rivalry among competing firms and government development of standards. The exception is AVI, for which many applications of the technology are moving forward rapidly and coordination efforts only recently have begun to emerge.

The standardization literature review identified several policy issues: (a) there may only be narrow windows in time during which collective action can be effective at reasonable costs; (b) the greatest power to influence may come at exactly the time when the information available to inform its action is most limited; (c) market pressure to disseminate a technology might argue for an imperfect standard in a timely fashion over the alternative of no standard at all; and (d) the sheer complexity of issues may confound and extend the duration of the standards-setting process.

The first two of these policy issues are not strictly applicable to the AVI case, since agencies operating highway facilities are not firms in a competitive market but monopoly or near-monopoly providers of road services. End users may choose to participate or not through the purchase of transponders, but they probably cannot choose between competing formats in the same way that consumers choose, say, between VHS and Beta videocassette recorders. The agencies can act to modify transponders at their will, and end users cannot elect another technology. Thus, the window in time during which a collectively determined standard can be effective may be much longer.

The third issue—the preference for imperfect standards over no standard at all—seems particularly apt for AVI. Market pressure for implementing systems is intense at present, and agencies are going it alone in the absence of a standard. Although agencies may later elect to convert to a standard system, or implement "gateways" or converters, some coordination to make such gateways technically feasible might be beneficial.

Finally, the complexity of the AVI derives both from its technical content and, perhaps more significantly, from marketing and implementation issues such as privacy and confidentiality of data and the use of transponders for law enforcement.

As for IVHS standards, the concerns over the timing of standards and the information on which standards decision can be based are more applicable. IVHS technologies that will rely on individual consumer choices between competing formats may be at risk for the emergence of suboptimal de

facto standards. Technologies whose formats are dictated by the decisions of monopoly operating agencies, on the other hand, may be less at risk.

All IVHS technologies would appear to suffer from uncertainty and poor information on which to base standards decisions. These are new technologies, and consumer preferences and willingness to pay for various services are simply not knowable at this time.

Unlike AVI, market pressures for other IVHS technologies have not yet become acute, suggesting that the choice may not be between an imperfect standard and no standard at all and that efforts to facilitate standards development will be fruitful. And finally, the issues of technological and market complexity are perhaps more acute for other IVHS technologies than for AVI, which may extend the duration of the standards-setting process.

ACKNOWLEDGMENTS

David J. Osiecki provided invaluable research assistance on this project. Four anonymous referees provided valuable comments and suggestions. This research was supported in part by grants from the Fenwick Library and the Graduate School of George Mason University. The author is responsible any remaining errors, mistakes, or misstatements.

REFERENCES

1. *Smart Highways: An Assessment of Their Potential to Improve Travel*. Report GAO/PEMD-91-18. U.S. General Accounting Office, Washington, D.C., May 1991.
2. *Transportation Research Circular 383: Intelligent Vehicle Highway Systems Communications Standards: Research Needs and Implementation Requirements*. (E. R. Case, M. I. Chung, R. L. French, and P. J. Tarnoff, eds.) TRB, National Research Council, Washington, D.C., 1991.
3. Mobility 2000. *Reports on Major Aspects of IVHS*. Texas Transportation Institute, College Station, Tex., 1990.
4. Castle Rock Consultants. *NCHRP Report 340: Assessment of Advanced Technologies for Relieving Urban Traffic Congestion*. TRB, National Research Council, Washington, D.C., 1991.
5. *Special Report 232: Advanced Vehicle and Highway Technologies*. TRB, National Research Council, Washington, D.C., 1991.
6. D. A. Hounshell. *From the American System to Mass Production, 1800–1932: The Development of Manufacturing Technology in the United States*. Johns Hopkins University Press, Baltimore, Md., 1984.
7. R. W. Hamilton. The Role of Nongovernmental Standards in the Development of Mandatory Federal Standards Affecting Safety or Health. *Texas Law Review*, Vol. 56, 1978, pp. 1329–1484.
8. L. L. Bernard. Social Control by Means of Regimentation and Standardization. In *Social Control in Its Sociological Aspects*. Macmillan, New York, N.Y., 1939, pp. 408–450.
9. J. Gaillard. *Industrial Standardization: Its Principles and Applications*. H. W. Wilson, New York, N.Y., 1934.
10. L. E. Gilbreth. *The Psychology of Management: The Function of the Mind in Determining, Teaching and Installing Methods of Least Waste*. Hive, Easton, Pa., 1973.
11. N. F. Harriman. *Standards and Standardization*. McGraw-Hill, New York, N.Y., 1928.
12. D. Reck. The Role of Company Standards in Industrial Administration. *Advanced Management*, April 1954, pp. 19–23.
13. D. Reck. National Standards in Industrial Administration (Parts 1 and 2). *Advanced Management*, 1954.
14. J. V. Coles. *Standardization of Consumers' Goods: An Aid to Consumer Buying*. Ronald Press, New York, N.Y., 1932.
15. D. Bottaro. *Analysis of Factors Affecting the Demand for and Supply of Voluntary Consensus Standards*. Working Paper MIT-EL 82-003WP. Massachusetts Institute of Technology Energy Laboratory, Cambridge, Aug. 1981.
16. P. David. New Standards for the Economics of Standardization. In *Economic Policy and Technological Performance* (P. Dasgupta and P. Stoneman, eds.). Cambridge University Press, New York, N.Y., 1987.
17. *ASTM Workshop on Regulatory Alternatives and Supplements* (H. R. Piehler, ed.). ASTM, Philadelphia, Pa., 1985.
18. Putnam, Hayes & Bartlett, Inc. *The Impact of Private Voluntary Standards on Industrial Innovation*. Report NBS GCR 82-420. National Bureau of Standards, U.S. Department of Commerce, Nov. 1982.
19. J. Farrell and G. Saloner. *Economics Issues in Standardization*. Working Paper 393. Department of Economics, Massachusetts Institute of Technology, Cambridge, Oct. 1985.
20. M. A. Sirbu and L. E. Zwimpfer. Standards Setting for Computer Communication: The Case of X.25. *IEEE Communications Magazine*, March 1985.
21. P. A. David. Clio and the Economics of QWERTY. *American Economic Review*, Vol. 75, 1985, pp. 332–337.
22. W. B. Arthur. Competing Technologies, Increasing Returns and Lock-In By Historical Events. *The Economic Journal*, Vol. 99, March 1989, pp. 116–131.
23. W. B. Arthur, E. Yurim, and K. Yurim. On Generalized Urn Schemes of the Polya Kind. *Cybernetics*, Vol. 19, 1983, pp. 61–71.
24. R. Cowan. Tortoises and Hares: Choice Among Technologies of Unknown Merit. *The Economic Journal*, Vol. 101, July 1991, pp. 801–814.
25. R. Nader. *Unsafe at Any Speed: The Designed-In Dangers of the American Automobile*. Grossman, New York, N.Y., 1965.
26. *The Voluntary Standards System of the United States of America: An Appraisal by the American Society for Testing and Materials*. ASTM, Philadelphia, Pa., n.d.
27. R. G. Dixon, Jr. *Standards Development in the Private Sector: Thoughts on Interest Representation and Procedural Fairness*. National Fire Protection Association, Boston, Mass., 1978.
28. P. J. Harter. *Regulatory Use of Standards: The Implications for Standards Writers*. Report NBS GCR 79-171. Office of Standards Information, Analysis and Development, Office of Engineering Standards, National Engineering Laboratory, National Bureau of Standards, Washington, D.C., Nov. 1979.
29. D. Hemenway. *Industrywide Voluntary Product Standards*. Ballinger, Cambridge, Mass., 1975.
30. M. Sirbu and K. Hughes. Standardization of Local Area Networks. *Proc., 14th Annual Telecommunications Policy Research Conference*, Airlie, Va., April 1986.
31. C. E. Clark, Jr. *The American Family Home, 1800–1960*. University of North Carolina Press, Chapel Hill, 1986.
32. C. G. Field and S. R. Rivkin. *The Building Code Burden*. Lexington Books, Lexington, Mass., 1975.
33. J. L. Gifford. The Innovation of the Interstate Highway System. *Transportation Research*, Vol. 18A, 1984, pp. 319–332.
34. B. E. Seely. The Paradox of Cooperation and Expertise: Consensus Highway Standards, 1921–39. In *Building the American Highway System: Engineers As Policy Makers*. Temple University Press, Philadelphia, Pa., 1987, pp. 118–135.
35. H. B. Johnson. *Order Upon the Land: The U.S. Rectangular Land Survey and the Upper Mississippi Country*. Oxford University Press, New York, N.Y., 1976.
36. P. A. David. The Landscape and the Machine: Technical Interrelatedness, Land Tenure and the Mechanization of the Corn Harvest in Victorian Britain. In *Essays on a Native Economy: Britain After 1840* (D. N. McCloskey, ed.). Meuthen, London, England, 1975, Chapter 5.
37. P. A. David with J. A. Bunn. The Battle of the Systems' and the Evolutionary Dynamics of Network Technology Rivalries. High Technology Impact Program Working Paper 15. Department of Economics, Stanford University, Calif., Jan. 1987.
38. V. Mangematin. The Place of the Standardization in RTI/IVHS Technology Competition. In *Road Transport Informatics/Intelligent Vehicle-Highway Systems*. Automotive Automation Limited, Croydon, England, 1991, pp. 301–309.

39. R. A. Johnston, M. A. DeLuchi, D. Sperling, and P. Craig. Automatic Urban Freeways: Policy Research Agenda. *Journal of Transportation Engineering*, Vol. 116, No. 4, 1990, pp. 442–460.
40. A. B. Boghani and J. R. Catmur. IVHS: A Public/Private Sector Partnership. In *Road Transport Informatics/Intelligent Vehicle-Highway Systems*. Automotive Automation Limited, Croydon, England, 1991, pp. 297–299.
41. J. H. Rillings and R. J. Betsold. Advanced Driver Information Systems. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, Feb. 1991, pp. 31–40.
42. S. E. Shladover. Issues in Communication Standardization for Advanced Vehicle Control Systems (AVCS). In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991.
43. A. M. Parkes and T. Ross. The Need for Performance-Based Standards in Future Vehicle-Man-Machine Interfaces. In *Advanced Telematics in Road Transport*, Elsevier Science Publishers, 1991, pp. 1312–1319.
44. M. R. Norman. Intelligent Vehicle Highway Systems in the United States—The Next Steps. *ITE Journal*, Vol. 60, No. 11, Nov. 1990, pp. 34–38.
45. *Report to Congress on Intelligent Vehicle Highway Systems*. Report DOT-P-37-90-1. Office of the Secretary of Transportation, U.S. Department of Transportation, March 1990.
46. G. Farber. Human Factors and IVHS Standards. Presented at 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis, Minn., Aug. 1991.
47. H. M. Heywood. U.S. DOT Perspective on IVHS Standards. Presented at 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis, Minn., Aug. 1991.
48. R. Shields. Communications and IVHS Standards. Presented at 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis, Minn., Aug. 1991.
49. R. Stehr. State DOT Perspective on IVHS Standards. Presented at 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis, Minn., Aug. 1991.
50. *Procedures for the Development and Coordination of American National Standards*. American National Standards Institute, New York, N.Y., 1983.
51. D. A. Scapinakis and W. L. Garrison. *Communications and Positioning Systems in the Motor Carrier Industry*. Report UCB-ITS-PRR-91-10. University of California, Berkeley, 1991.
52. N. Robertson. In *Inside IVHS, Intelligent Vehicle Highway Systems Update*, Waters Information Services, Inc., June 24, 1991.
53. *Analysis of Automatic Vehicle Identification Technology and Its Potential Application on Florida's Turnpike*. Center for Urban Transportation Research, College of Engineering, University of South Florida, Tampa, Dec. 1990.

Publication of this paper sponsored by Task Force on Advanced Vehicle and Highway Technologies.

Policy Implications of Driver Information Systems

KAN CHEN

The potential effects of driver information systems (DISs) have strong policy implications for traffic management authorities. Privacy and standardization issues are of major concern when implementing electronic toll collection. Road pricing, a simple technical extension of toll collection, has profound policy implications in terms of political acceptability. Real-time traffic information is being provided by an increasing variety of technological systems with financial viability, and their market acceptance and dominance can strongly influence the future direction of dynamic route guidance. Public authorities must soon make policy decisions about their roles in facilitating multijurisdictional agreements on traffic diversion and in facilitating multimode transport choices to be made by travelers. As for driving assistance, policy decisions may help early implementation of automatic emergency signaling; issues of safety regulations and legal liability should be settled before a host of technologies for driving assistance can become marketable. DIS functions that are to provide business and personal information do not seem to have any direct policy implications for traffic management. However, to the extent that such DIS functions affect the efficiency and regulation of transport, and the marketability and future standards requirements of DIS in general, traffic management authorities cannot ignore such DIS development. The impacts of policy making and technology management are so intertwined that frequent in-depth exchange of views between policy and technology developers should be routinized through joint projects and periodic reviews.

The rapid development and amalgamation of information technology with automobile and road technologies have given rise to new programs in intelligent vehicle-highway systems (IVHSs) in Europe, Japan, North America, and other parts of the world. These programs aim to improve road transport efficiency, safety, comfort, and environment. Driver information systems (DISs) is one of the important components of IVHS that is developing rapidly and is primarily centered on the vehicle to provide motorists with information of interest to them.

The newness and the evolving nature of DIS are such that the potential impacts of these systems and the associated policy implications are difficult to predict. Technology developers and promoters understandably have concentrated on the potential benefits—individual and social—of these systems. Most of the impact assessments typically raised have tried to answer such questions as how much traffic delay can be reduced by dynamic route guidance and how much such reduction would mean economically in tangible terms (*1*). Policy issues were brought up occasionally in these assessments, but only sporadically and for the purpose of considering how to circumvent them by technology (*2*).

There have been some efforts to assess the impacts and policy implications of IVHS in Europe (*3–7*), as well as in the United States (*8,9*). However, most publications have either treated IVHS in general or discussed only a selected issue (e.g., legal liability). The objective of this paper is to provide a comprehensive framework for policy makers and advisors to monitor, assess, and influence the development of DIS from the perspective of traffic management authorities. The work in this paper was based on the author's interviews in late 1990 with a number of organizations in Europe engaged in selected PROMETHEUS and DRIVE projects. However, relevant information was also drawn from the author's knowledge of IVHS activities in Asia (especially through 5-week interviews in Japan in early 1991) and in North America (especially through IVHS activities at the University of Michigan).

In this paper, four categories of DIS will be considered:

1. Toll collection and road pricing,
2. Traffic information and route guidance,
3. Driving assistance, and
4. Business and personal information.

Because DIS is the part of IVHS that involves a great deal of technological development within the private sector that will market DIS products with or without complementary investment by the traffic management authorities, cost estimates of marketable DISs are essential input to the task of assessing realistic policy implications of DIS. Through interactions with a number of private firms, the author has assumed that the mass market for DIS will support low-end in-vehicle units (IVUs) on the order of a few hundred dollars—or the cost of a luxury car radio or an automobile air conditioner. High-end IVUs can probably sell for a few thousand dollars—or 15 to 20 percent of the price of a luxury car. For commercial vehicle operations (CVOs) and for infrastructure investment by the public sector, tangible cost-benefit analyses will be the basis for the necessary justification for DIS investment. However, this may not be sufficient, because there are other policy issues such as privacy and safety, which will be discussed in this paper. The point that should be made at the outset is that many of the policy implications of DIS are intertwined with the specific technologies to be used for DIS and their costs to the users. Therefore, each section of this paper will take a sociotechnological approach by mixing technoeconomic descriptions of DIS with policy-relevant discussions.

College of Engineering, 4112 EECS, University of Michigan, Ann Arbor, Mich. 48109.

TOLL COLLECTION AND ROAD PRICING

Electronic toll collection is an IVHS technology ready for deployment. The current technology centers on automatic vehicle information (AVI) using electronic transponders. It may be considered as the first wave of IVHS technology entering the mass market. At a relatively low cost (about 10 percent of the normal toll) and with high reliability (about 99 percent accurate), the efficiency of road travelers and the efficiency of infrastructure toll collectors can be increased substantially. In some installations, such as the Crescent City Connection bridge in New Orleans, the public authorities share their savings by offering a 30 percent discount on tolls for AVI-equipped vehicles (Paisant, unpublished data, 1991). Nonequipped motorists also benefit from the shortening of the queues as the equipped vehicles zip through the toll booths without stopping. It appears to be a "win-win" arrangement for all major stakeholders.

There are two basic technologies for the AVI transponder. The ones available on the market operate at a UHF. One basic technology uses the surface acoustic wave (SAW) approach. The other basic technology uses an electronically erasable and programmable read-only memory (EEPROM) chip, which can have its code number or message modified electronically. The EEPROM chip may be passive or active.

The existence of several AVI technologies and automatic toll collection system designs gives rise to the issue of standardization, which is getting hotter as this category of DIS application spreads and as the various vendors vie for market share and market dominance. As with other information technologies such as computers, there are two general approaches to standardization. One approach is in the marketplace, in which the market leader sets the de facto standards and compels other suppliers to go along with an obvious disadvantage, since the inner workings of the standards are often opaque and proprietary. The other approach is in the conference room, in which committees consisting of major industrial suppliers, often with the involvement of major users, public authorities, and academics, try to agree on a minimum set of standards that will allow multiple vendors to coexist with mutually compatible hardware and software products. The market-leader approach may be considered an unfair practice to the dominated suppliers, but it is realistic and works as fast as market penetration. The committee approach usually comes up with more-open systems that favor the users, but it is a slow process that often appears wasteful and can be overtaken by the market-leader approach.

Perhaps one important reason that standardization for automatic toll collection has been, and perhaps should be, slowed is the yet-to-be-resolved issue of privacy. The functioning of the toll collection systems that have been described assume that the equipped users have no objection to their identity's being revealed to the system, thus allowing their whereabouts at what time to be known to those with access to the system. One way to reduce this concern is to use a technology similar to that for prepaid phone cards, whose users' identities cannot be known to the telephone system.

Phone cards have been used in Europe with success, but adapting the concept to a "smart travel card" for automatic toll debiting and other such applications is technically not quite straightforward. First, for automatic toll collection, unlike for pay phones, there can be no contact between the smart travel card and the interrogating system. Second, the required two-way communication will be a dialogue, and not a two-way monologue (as in AVI), because the amount to be debited on the smart travel card will depend on the location and the time of the interrogator signal—and the complicated dialogue needs to be completed within a short time if the vehicle is not required to slow down excessively.

As indicated previously, electronic road pricing in its primitive form is technically a simple extension of electronic toll collection. There are two general types of road pricing: (a) the charge for the use of a road, which has long been in practice for toll roads, regardless of whether the toll is collected manually or electronically; and (b) the charge for keeping a vehicle in a particular district, no matter whether the vehicle is in motion or not (analogous to having the vehicle in a large parking structure). It is the second type of road pricing that has recently been proposed to be put into practice and that has been particularly controversial.

The concept of road pricing as a means for demand management is attractive to economists who argue that excessive congestion is a phenomenon of inefficient allocation of scarce resources. An efficient way to reduce congestion is thus to introduce a market mechanism to road transport. Without road pricing, more road building would simply attract more traffic to the new roads, and the previous level of congestion would return as the system seeks a new equilibrium. In the long run, the only way to reduce congestion is to charge the less urgent users—some opponents would say the less affluent users—sufficiently to keep them off the congested routes. This concept is not new at all, but the low-cost electronic means to make road pricing practical is new and has given the concept a new life (*10*).

Road pricing has many opponents. Besides those who believe that road pricing favors the rich, the strongest public sentiment against road pricing is its appearance as another tax. Opponents have also raised the privacy issue as a negative factor. On a rational basis, the proponents of road pricing appear to have answers to all the objections that have been mentioned (*11*). For example, reduced rates may be charged to the poor, privacy may be protected by the use of anonymously prepaid smart cards, and so on.

In Singapore, a manually operated road pricing system (an area licensing scheme) to keep most of the motor traffic from its central business district has been in operation since the mid-1980s. The scheme was highly successful in reducing traffic congestion in the central business district. In fact, it was so successful that the roads became highly underutilized in the district, and the price was reduced from Singaporean $5 to $3 for any vehicle to enter the restricted zone during peak hours (*12*). Recently, Singapore has planned to convert its road pricing system from manual to electronic.

In Europe, there is a joint manual and automatic toll cordon for Oslo, Norway; similar plans are under consideration for Stockholm, Sweden. In the United Kingdom, serious consideration for road pricing has been coupled with very innovative ideas for its implementation. For example, a "timezone" concept has been proposed for London, which would be ringed with roughly concentric circles representing progressively more expensive tolls as one approached the center (*11*). This approach would prevent traffic diversion at zone boundaries as

has happened around the central business district of Singapore, causing congestion around its boundaries. An even more radical concept, known as congestion metering, is being considered by the city of Cambridge (*13*). Unlike the usual road pricing scheme—as in Hong Kong, where a congested zone is predetermined and a fixed fee for entry is charged whether the zone is congested or not—congestion metering will levy a charge only when a vehicle experiences actual congestion (defined by a threshold of vehicle speed and number of stops per unit distance). It is believed that such a scheme will induce a more economically rational behavior from the driver and result in more effective relief of congestion. However, the political acceptability of these schemes is still to be tested.

TRAFFIC INFORMATION AND ROUTE GUIDANCE

For years, real-time traffic information collected by traffic management authorities has been provided to the general public, including drivers on the road, through commercial and public radio and television stations at no incremental cost to the users. However, such information may not be timely and relevant enough to help drivers make strategic route choice decisions. Even dedicated traffic stations (highway advisory radio, or HAR) may give too much irrelevant information, and the driver may not tune to these stations at the critical moments for decision making. On the other hand, the incremental cost to the driver for these services is negligible, since most car radios can tune to dedicated stations just as easily as to other stations. To ensure that the driver does get new significant traffic information when it becomes available (such as after a major incident has been identified), automatic HAR (AHAR) radio has been designed to turn the radio on, or interrupt other audio devices and programs, and automatically tune to the HAR station with the latest information (*14*). However, the added cost of the AHAR has not made it very marketable, and most HAR stations have not been sending out the automatic interrupting signal to make AHAR work.

Recently radio data systems (RDSs) have become available in Europe to provide low-rate (about 1,000 bit/sec, or bps) coded information to properly designed receivers, using the sidebands of the frequency spectrum assigned to the broadcaster. Originally designed to provide program identification and alternative frequencies for better reception of the desired radio program or program type, RDS is now being promoted to provide real-time traffic information through a traffic message channel (TMC), which may provide effectively 74 bps of traffic information codes to specially equipped radios (*15*).

There is an active European project working on the standards of RDS-TMC (*16*). As indicated, the output may be in the form of textual or synthetic voice displays—textual display may be more distracting but it would not interfere with or require interruption of other audio programs. For owners of high-priced (about $300) car radios, the added cost of acquiring the RDS feature is marginal. The cost to the stations to add RDS features would be on the order of a few thousand dollars, and the incentive for them to make the investment could be significant because of competition. To go to TMC, however, there would be additional infrastructure costs for traffic data collection and collation and for message management systems.

The European RDS standards have almost been set, but the Japanese have not been satisfied with them and have been developing and testing new approaches, using more sophisticated digital modulation schemes and error correcting codes. One of Japan's FM multiplex broadcasting approaches has been tested for mobile reception; it promises to provide much higher bit rates than the current European approach—10 to 12 kilobit/sec (kbps) versus 1 kbps (Yamada, unpublished data, 1991). These high bit rates can be used to transmit a great deal more traffic information and even projected link times without the need for additional frequency spectrum. Given the still-fluid situation with RDS in North America, it behooves the traffic management authorities to work with telecommunication policy makers there to review all the competing schemes of RDS before final standards are set.

The latest technological system on the market for providing real-time traffic information is the TrafficMaster, which has been in operation for the London area since mid-1990 (*17*). Average vehicle speed in the fast lane of the motorways around London (M25 and a limited range of other motorways connecting to it) is automatically measured every 2 mi and reported every 3 min by infrared sensors. When the speed drops below a threshold of 25 mph, the average speed and location of the congested segment are transmitted over a paging service frequency to owners of the TrafficMaster's simple map display unit, which may be carried by the owners or put on top of their vehicle dashboards. Paging service is also available. The price of TrafficMaster is about $500 for the portable unit and about $32/month for the service charge. The paging service option is another $30/month. By late 1990, there were about 300 customers (35 to 40 percent of whom choose the paging option), and the number exceeded 2,000 by the end of 1991. Because the breakeven point is only about 2,500, TrafficMaster's financial viability appears to be very promising (Martell, unpublished data, 1990).

From the perspective of traffic management authorities, at least two policy issues are related to traffic information. The first has to do with the fusion and authentication of traffic information, and the other has to do with the potential risk of a specific traffic information provider's near-term success preempting certain forms of dynamic route guidance that must use real-time traffic information.

As in any private service-providing activities, there is a risk of "cream-skimming" so that the easiest and most profitable services are provided while the more difficult but perhaps much needed services are neglected. For example, TrafficMaster has focused on M25 not only because it is the most traveled motorway, but also because the average speed can be measured most conveniently and the operating company needs to deal with only one public authority, namely, the U.K. Department of Transport. It is encouraging to learn that TrafficMaster has recently considered expanding to the arterials (the A-roads) in the London area (*18*).

If drivers just want to find their way to their destination under normal traffic conditions, they can use static route guidance, which does not require real-time traffic information and thus can be autonomous. Philips' CARIN system and Bosch's TravelPilot system can provide static route guidance through a combination of dead-reckoning and map-matching with the digital map information stored in a magnetic tape cassette or compact disk. To cover a wide area of possible travel (e.g.,

Western Europe or the entire United States), the most practical way to store the vast amount of map information is through the use of a compact disk system, which has been rather expensive. For example, TravelPilot with compact disks started to sell for about $3,500 in the United States, although the price quickly dropped to $2,500 and is expected to drop with an expanding market. This may be compared with currently available navigation systems with GPS plus dead-reckoning and map-matching for more than $4,000 in Japan, where compact disk drives can be acquired for as low as a few hundred dollars, substantially below the cost of those drives in Europe. In any event, a fully functional navigation system will probably be too expensive in the near future to be accepted in a "baseline" IVHS system for the mass market in the United States (*19*). This is an important point for the public authorities to consider if privatizing IVHS is an important goal.

Because one of the major frustrations in driving is to get caught in unexpected traffic congestion, dynamic route guidance that takes into account real-time traffic information in route guidance is highly desirable. Dynamic route guidance is by no means a new concept. Over the past 20 years, there have been three discernible generations of such systems (*20*). The first-generation systems include ERGS in the United States (*21*), CACS in Japan (*22*), and ALI in Europe (*23*)—all programs in the 1970s—using low-rate inductive loops for communication. For example, CACS transmitted 144 bits of information at each junction at 4.8 kbps (*24*). The second generation is typified by ALI-SCOUT (*25*), whose development began in the early 1980s, using beacons for short-distance communication at major junctions. At each junction, the beacon transmits 8 K of information at 125 kbps. [The next generation of ALI-SCOUT, known as EURO-SCOUT, is to be ready by 1992; it will transmit 30 K at 500 kbps (Sodeikat, unpublished data, 1990)]. The third-generation systems include CARMINAT (*26*) (one-way wide-area communication into the vehicle via RDS), SOCRATES (*27*), and ADVANCE (*28*) (two-way communication via a digital cellular radio link), which are systems under development in the late 1980s and early 1990s. In these systems, information will be transmitted *continuously* at a relatively high rate—about 9.6 kbps in SOCRATES (McQueen, unpublished data, 1991). As we progressed from the first to the third generation, the required number of equipped junctions on the infrastructure decreased while the typical cost of an IVU increased (from $80 to $250, to an estimated $800, hopefully, for the third-generation of IVU at mass production).

It is interesting to note that the first two generations of dynamic route guidance systems are infrastructure-based and the third generation is primarily vehicle-based. The former searches for the best routes with equipment on the infrastructure (such as with ALI-SCOUT); the latter does that with equipment on board of the vehicle (such as with CARMINAT). It is important to discuss the policy implications of these two types of system. First is the financial implication, which may affect the system viability. Infrastructure-based systems naturally put major capital investment requirements on the infrastructure side. This may be a slow process, not only because of the magnitude of the investment but also because of the need for multijurisdictional cooperation in both system installation and operation. Drivers who are early investors in the IVUs for such systems may be dissatisfied with their limited usability for a long time. The whole system may then become unacceptable to the users, and the infrastructure costs may be too high for a government or operating company to bear. On the other hand, the vehicle-based systems require expensive IVUs, and the early users are likely to be the more affluent drivers and the commercial vehicles. Second is the question of who is in control. The infrastructure-based systems would be more amenable to the presumption that traffic should be controlled by a central authority (public or a private surrogate) that will guide the traffic to reach a "system optimum" according to a systemwide objective (such as minimum total vehicle hours of delays) subject to centrally determined constraints (such as no traffic diversion to schools and residential areas). Such systems would also make it easier for public authorities to encourage multimode transport options (e.g., park-and-ride). On the other hand, the vehicle-based systems would be more amenable to the presumption that drivers will determine their own objective functions and will search for "user optimum" (such as a weighted average between travel time and tolls paid) subject to privately determined constraints (such as avoiding subjectively perceived high-crime areas).

Regardless of the approach taken, the projection of link time (up to about 90 min into the future) will be needed for dynamic route optimization, and this projection must be done centrally. This implies new tasks and improved technologies to be used by the traffic management authorities. For example, credible and reliable projection of link time will need the speedy and accurate estimation of the duration of an incident after it is detected and verified. Such estimations are being done only haphazardly at present. Some form of simple model based on artificial intelligence is needed to predict incident duration as a function of the link location, time of the day, number of lanes blocked, weather condition, whether bodily injuries are involved, and so on. At present, the projection of link times in the ALI-SCOUT system is done by a sort of moving average of current and historical data with modification through link models, which appear to ignore the interactions between links in a network (Janko, unpublished data, 1990). A more sophisticated real-time simulation model may be needed to provide more accurate link-time projections (*29*).

Another important task in this area for the traffic management authorities is to provide updated information of road changes (construction, reconfiguration, new roads and roundabouts, new one-way and turning lane restrictions, etc.). Autonomous systems that provide navigation and static route guidance would need this updated information. As they get such information from the infrastructure, they might as well get the dynamic route guidance information from the infrastructure, if it is available. Thus there is a strong motivation to explore dual-mode systems that combine the autonomous and infrastructure-based systems (*30*).

With the coexistence of multiple basic approaches to dynamic route guidance, standardization is clearly a very important policy issue. It is interesting to note that the Japanese have abandoned the infrastructure-based approach as a result of the evaluation survey among drivers involved in the CACS program during the 1970s (*31*). The North Americans, notably through TravTek (*32*) and ADVANCE projects, have been

working during the last few years exclusively on vehicle-based systems for dynamic route guidance. The substantive progress being made in SOCRATES may lead to a vehicle-based system in Europe that would be more congruent with the trends in Japan and North America. On the other hand, the low-cost IVU is clearly an attractive feature of infrastructure-based route guidance systems. If a sufficient number of cities in Europe and North America are convinced in the near future to make the substantial investment on the infrastructure for them, such systems as ALI-SCOUT may preempt the market for dynamic route guidance.

Because the social benefits of DIS are supposed to be relieving of traffic congestion and increase of driving safety, not just helping a small group of users to reduce their travel times, it is important for traffic management authorities to evaluate the traffic and safety impacts of DIS. Fortunately, such projects have been launched in several countries, including the United Kingdom (*33*). Preliminary assessment of dynamic route guidance has indicated that the typical reduction of vehicle hours of delays would be in the range of 7 to 15 percent (*34*). Recently more detailed evaluation using computer models has indicated some interesting but problematical results. For example, the user benefits have been shown to decrease—they may even become negative—under certain circumstances as the percentage of cars equipped with the ALI-SCOUT system increases beyond about 15 percent (McDonald, unpublished data, 1990). This was based on the assumption that drivers of equipped cars would follow the routes along the main roads determined by the traffic center while the drivers of unequipped cars who know the local geography would "rat race" through the small urban streets. On the other hand, research at the University of Michigan has shown that the benefit to users in a vehicle-based route guidance system (*29*) can continue to rise for the equipped drivers as the penetration rate increases from 0 to 100 percent. There is an important policy issue here regarding the degree and the means of control that the traffic management authority should exercise to discourage rat race and to make dynamic route guidance attractive to those who have invested in the system.

DRIVING ASSISTANCE

The DIS application category of driving assistance consists of new electronic devices and subsystems that will augment the functions of driving tasks. From the perspective of communications, they may be subcategorized as autonomous, vehicle-highway, and vehicle-vehicle systems. Vehicle-high way systems involve modification of the highway infrastructure, so they are of obvious and immediate concern to traffic management authorities. However, significant public policy issues exist in the other subcategories as well.

Autonomous driving assistance includes, first of all, new electronic display panels that have been designed to show only the key information of interest to the driver at the moment, thus reducing information (output) overload to the driver and saving the very limited space on the dashboard. Similar arrangements have been designed to reduce the number of driver (input) control knobs through a hierarchy of touch-screen commands. The safety and human-factor merits of these devices, under various external lighting conditions and driving situations, are yet to be proven by extensive tests.

Head-up displays of key information on the windshield, a technology borrowed from military aircraft, are generally considered to be desirable because they can provide information without requiring much eye movement from the driver. The technology is still expensive [about $300 (Spreitzer, unpublished data, 1991)] and is only beginning to be offered as an option on some of the most luxurious cars. Vision enhancement through infrared or ultraviolet system is another example of autonomous driving assistance DIS. Still another autonomous system that is attractive is to superimpose a red bar on the side mirror when a sensor detects a vehicle on the driver's blind spot to caution him about changing lanes; this is shown in the PROMETHEUS video tape. Such a system provides only warning signals with the responsibility of vehicle control remaining entirely with the driver. However, there may still be questions of legal liability if collisions occur because the warning red bars malfunction.

In the subcategory of vehicle-highway communication for driving assistance is the example of in-vehicle safety advisory and warning system, in which some of the radio-frequency warning signals—such as hidden stop sign, dangerous curve, and construction ahead—may come to the vehicle from ground points. To receive such active signals at a close distance requires relative low cost receivers on the vehicle—no more than several hundred dollars (*35*). They can be implemented soon on the presumption of willing cooperation between traffic management authorities to provide the active signals and private manufacturers to make the receivers. Another vehicle-highway system that the traffic management authorities may wish to promote is the automatic emergency calls (or distress signals) to be transmitted by vehicles in trouble. This would require three elements: (a) an automatic as well as a manual triggering device for transmitting the signal, because the driver may be unconscious in an emergency situation; (b) an automatic indication of the vehicle location with sufficient accuracy for the rescue crew to find the vehicle; and (c) cooperating traffic management authorities to take up the rescue mission. In a recent survey among American truck drivers and operators, this was considered the highest-priority function of IVHS (*36*). To speed up such applications, some government leadership is in order to get the system set up and maintained in full operation, at least in those segments of the motorways and arterials where accidents have occurred frequently and where emergency calls from telephones or from motorists are either impractical or unsafe.

In the subcategory of vehicle-vehicle communication for driving assistance, the examples are fewer but interesting. The PROMETHEUS program seems to have emphasized communication between vehicles moving in opposite directions. The Handshake project features fog warning by vehicles just coming out a fog zone to oncoming vehicles that are about to enter the fog zone. Similar communication schemes have been proposed so that vehicles, after observing a traffic jam on the other side of the motorway, can send signals to vehicles in the opposite direction upstream of the jam so that they may make a timely diversion. Another recent idea that achieves similar functions, but uses vehicle-highway communication rather than direct vehicle-vehicle communication, is to rely on an electronic transponder installed on the middle strip of

a divided motorway as a mailbox for vehicles moving in opposite directions to deposit and pick up messages for driving assistance.

It should be pointed out that all such vehicle-vehicle communication applications, no matter whether they are for vehicles moving in the same or opposite directions, would not be practical until the percentage of vehicles with appropriate and compatible communication equipment becomes significant. The implementation would be difficult if everything were left to the free market because of the lack of incentives for early investment. Perhaps the way to bootstrap the market is for the traffic management authorities to install the electronic mailboxes discussed previously and provide plentiful information for driving assistance (e.g., icy bridge ahead, etc.) so that early as well as late investors can receive immediate benefits.

TABLE 1 Matrix Summary

Policy Impacts	Toll & Rd Price	Route Guidance	Driving Assistance	Business & Pers Info
Privacy	XX	X		
Standards	XX	XX	X	X
Financing	X	XX	X	X
Political Acceptance	XX			
Multijurisdiction	X	XX		
Multimode	X	XX		
Safety		X	XX	X
Legal Liability			XX	

BUSINESS AND PERSONAL INFORMATION

The provision of business and personal communication to drivers on the move used to be considered a luxury, at least for noncommercial vehicles. However, with the decreasing cost of information technology and the increasing demand for such services in our information society, provision for business and personal information is deemed highly desirable if not essential for long-distance commuters, tourists, high-level executives, and drivers of commercial vehicles—such as trucks, taxis, ambulances, and police cars, some of which are not strictly commercial.

From a policy perspective, the regulation and frequency allocation for business and personal communication appears to be entirely in the domain of telecommunication authorities. However, to the extent that the technology and communication media for this application category of DIS share and overlap with those for the other three application categories that are of central concern to traffic management authorities, the latter cannot ignore the policy implications of this fourth category of DIS as well. Driving safety issues in the use of some of business and personal information systems also would require evaluation and regulation activities of the traffic management authorities.

Personal information can now be conveyed to the motorist (driver and passenger) by car phone, facsimile, or paging—some luxury cars, especially in Japan, even have commercial television for the passengers, or for the driver while the car is standing still. Clearly some of these communication channels may be used for traffic information and route guidance as well. As mentioned previously, TrafficMaster provides real-time traffic information by way of paging service frequency (*17*). The cellular car phone system may be used to provide link-time information for on-board route optimization, especially after the cellular system becomes digital as in the future GSM in Europe (*27*). With the expected advent of personal communication service, which may one day replace the home phone and the car phone, its continuing development should be monitored by the traffic management authorities as well as the communication regulatory authorities.

Tourist and other "Yellow Page" information is expected to be provided to drivers on the move as a feature of DIS. TravTek, one of the major American IVHS demonstration projects involving 100 test vehicles in Orlando, Florida, will feature the provision of tourist information around Disney World. The automobile clubs (both the AA in Europe and AAA in the United States) have libraries rich in tourist information that can be made readily available. There are strong commercial interests in making Yellow Pages available in digital form to facilitate information search and to make two-way communication for Yellow Page transactions such as hotel reservation practical. This growing business may have policy implications in both financing and technical decision making on the basis of safety and human factors.

SUMMARY AND CONCLUSIONS

The matrix shown in Table 1 summarizes the potential prominent policy impacts of the four categories of DIS that have been discussed in this paper. The double X's indicate the policy implications that demand the most critical attention. It should be clear from the matrix that policy implications are quite different for the various categories of DIS and that the relative significance of policy implications may vary from one country to another. However, this author has one general observation that is common to all the countries he has visited in Europe and Asia as well as North America: the DIS researchers in the private industry have not thought very much about policy implications except in a haphazard way, yet they are making technological choices and developmental efforts that may have profound policy implications. Similarly, the policy makers are not familiar with exactly what DIS technologies are being planned for deployment because they have been under development mainly in the private sector. Both sides can save time at the later stage of implementation and can be synergistic if they work more closely and frequently at the developmental stage. Therefore, the author suggests that frequent in-depth exchange of views between policy and technology developers be routinized through joint projects and periodic reviews.

REFERENCES

1. Working Group on Operational Benefits. Final Report. Mobility 2000, March 1990.

2. Working Group on Advanced Driver Information Systems. Final Report. Mobility 2000, March 1990.
3. W. J. Gillan. PROMETHEUS and DRIVE: Their Implications for Traffic Managers. *Proc., 1st Vehicle Navigation and Information Systems Conference*, 1989, pp. 237–243.
4. B. Stoneman. *The Potential Impacts of Future In-Vehicle Driver Information Systems*. Working Paper WP/TO/75. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1990.
5. T. Karlsson. *Financing*. Working Paper k:/secfo/docs/SC0259. DRIVE Program, March 1990.
6. T. Karlsson. *Road-Pricing*. Working Paper k:/secfo/docs/SC0261. DRIVE Program, March 1990.
7. T. Karlsson. *Privacy*. Working Paper k:/secfo/docs/SC0274. DRIVE Program, March 1990.
8. J. Vostrez. *Vehicle/Highway Automation: Policy Issues and Strategies in California*. SAE Paper 891721. SAE, 1989.
9. K. Chen and R. D. Ervin. *American Policy Issues in Intelligent Vehicle-Highway Systems*. SAE Paper 901504. SAE, 1990.
10. K. Small, C. Winston, and C. A. Evans. *Roadwork*. Brookings Institution, Washington, D.C., 1989.
11. N. Green. London "Timezone." Presented at the National Economic Development Organization Conference on Information Technology and Traffic Management, London, England, Nov. 1990.
12. B. G. Field. *From Area Licensing to Electronic Road Pricing: A Case Study in Vehicle Restraint*. SAE Paper 911677. SAE, 1991.
13. B. Oldridge. Electronic Pricing: An Answer to Congestion. Presented at the National Economic Development Organization Conference on Information Technology and Traffic Management, London, Nov. 1990.
14. H. Turnage. Development of Automatic Highway Advisory Radio in the United States. *Proc., International Conference on Road Traffic Signalling*, IEE, London, England, 1982, pp. 156–158.
15. M. Thomas. Telecommunications: The Backbone of DRIVE. *Traffic Engineering and Control*, April 1990, pp. 209–213.
16. P. Davies and C. Eng. Radio Data System–Traffic Message Channel. *Proc., 1st Vehicle Navigation and Information Systems Conference*, 1989, pp. A44–A48.
17. D. Martell. Navigation Systems. Presented at the National Economic Development Organization Conference on Information Technology and Traffic Management, London, England, Nov. 1990.
18. Trafficmaster Seeks Expansion of Network onto London A Roads. *Inside IVHS*, Vol. 1, No. 12, June 10, 1991, pp. 6–7.
19. M. P. Ristenbatt. A Communications Architecture Concept for ATIS. *Proc., 2nd Vehicle Navigation and Information Systems Conference*, IEEE/SAE, 1991.
20. *Dynamic Route Guidance*. Working Package 2.2. TARDIS (Traffic and Roads—DRIVE Integrated Systems, The Total Traffic Management Environment), DRIVE Project V1018, Nov. 1990.
21. D. Rosen, F. Mammano, and R. Favout. An Electronic Route-Guidance System for Highway Vehicles. *IEEE Transactions on Vehicular Technology*, Vol. 19, No. 1, 1970, pp. 143–152.
22. N. Yumoto, H. Ihara, T. Tabe, and M. Naniwada. Outline of the CACS Pilot Test Systems. Presented at the 58th Annual Meeting of the Transportation Research Board, Washington, D.C., 1979.
23. P. Braegas. Function, Equipment and Field Testing of a Route Guidance and Information System for Drivers (ALI). *IEEE Transactions on Vehicular Technology*, Vol. 29, No. 2, 1980, pp. 216–225.
24. H. Kawashima. Present Status of Japanese Research Programmes on Vehicle Information and Intelligent Vehicle Systems. Presented at the DRIVE Conference, Brussels, Belgium, Feb. 1991.
25. R. Von Tomkewitsch. Dynamic Route Guidance and Interactive Transport Management with ALI-SCOUT. *IEEE Transactions on Vehicle Technology*, Vol. 40, No. 1, 1991, pp. 45–50.
26. *CARMINAT Leads the Way*. Press Document. Renault, Oct. 1990.
27. I. Catling and B. McQueen. Road Transport Informatics in Europe—A Summary of Current Development. *Proc., 5th Jerusalem Conference on Information Technology*, Oct. 1990, pp. 702–715.
28. A. M. Kirson. RF Data Communications Considerations in Advanced Driver Information Systems. *IEEE Transactions on Vehicle Technology*, Vol. 40, No. 1, 1991, pp. 51–55.
29. D. E. Kaufman, K. E. Wunderlich, and R. L. Smith. *An Iterative Routing/Assignment Method for Anticipatory Real-Time Route Guidance*. IVHS Technical Report 91-02. University of Michigan, Ann Arbor, May 1991.
30. P. Haessermann. Route Guidance Systems: Combination of Autonomous Systems with Infrastructure-Based Systems. Presented at SITEV '90, Geneva, Switzerland, May 1990.
31. H. Kawashima. Two Major Programs and Demonstrations in Japan. *IEEE Transactions on Vehicle Technology*, Vol. 40, No. 1, 1991, pp. 141–146.
32. J. H. Rillings. *TravTek*. General Motors, June 26, 1991.
33. A. Stevens and M. Bartlam. Safety and Traffic Impact Evaluation of Driver Information Systems. Presented at PROMETHEUS Pro-General Workshop, London, England, Nov. 1990.
34. JMP Consultants, Ltd. *Study to Show the Benefits of Autoguide in London*. Final Report. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, Feb. 1987.
35. K. Housewright. Near-Term Automatic Driver Information Systems Application. Presented at the 1991 SAE Future Transportation Technology Conference, Aug. 1991.
36. B. E. Stone and R. D. Ervin. *Survey of the Trucking Industry's Preferences for IVHS*. University of Michigan, Ann Arbor, Oct. 1990.

Publication of this paper sponsored by Task Force on Advanced Vehicle and Highway Technologies.

Full-Scale Experimental Study of Vehicle Lateral Control System

WEI-BIN ZHANG, HUEI PENG, ALAN ARAI, PETER DEVLIN, YE LIN, THOMAS HESSBURG, STEVEN E. SHLADOVER, AND MASAYOSHI TOMIZUKA

An ongoing experimental research project focusing on vehicle lateral control is reported. This project is being jointly conducted by the University of California's PATH Program and IMRA America, Inc. The lateral control system to be tested uses the concept of cooperation between the vehicle-borne steering control system and a discrete magnetic reference system in the roadway. The components of this experimental setup are described. The analytical and experimental research conducted before this project on the various components are reviewed, some results of the completed tests are provided, and the planned tests are outlined.

The purpose of an automated vehicular lateral control system is to maneuver, or to assist a driver to maneuver, the vehicle. These steering functions may include lane keeping, lane changing, merging, and diverging. The automated lateral guidance and control system described in this paper includes four basic components: a roadway reference system that defines a path for the vehicle to follow, sensing devices that acquire information from the roadway reference system and collect other necessary information for steering control, a controller that generates steering commands based on the control algorithm, and a steering actuator that executes the steering command given by the controller. The system schematic is shown in Figure 1.

Automated lateral control systems have been studied for the past 30 years. Several systems using various technologies have been developed or simulated (*3*). However, only a few full-scale systems have been tested in a real road-vehicle environment. After an extensive literature review, PATH developed a concept of a cooperative lateral control system that allows the vehicle to receive information from a discrete magnetic reference and sensing system. The roadway reference system consists of a series of permanent magnetic markers buried in the roadway. On-board sensors measure the proximity of the vehicle to the markers and decode upcoming road geometry information. The lateral controller makes steering angle corrections according to the vehicle's tracking error and any upcoming road curvatures so that both the tracking accuracy and ride quality can be improved. This concept has been investigated in a series of PATH studies (*1–4,6–8*). In parallel to the PATH work on the roadway reference and sensing system and the lateral control algorithms, IMRA has been working on a design of a steering actuator for vehicle lateral control, in particular for lane keeping. This design is intended to allow automated steering control to augment the driver's steering performance on the highway.

Because of their mutual interest in automatic vehicle lateral control, PATH and IMRA agreed to conduct a joint experimental study on vehicle lateral control, beginning in January 1991. The goal of the joint project is to demonstrate the concept of automatic lateral control with a full-scale automobile. This paper reviews the work performed individually by PATH and IMRA before this project and the joint efforts in conducting the full-scale vehicle experiment.

W.-B. Zhang, P. Devlin, Y. Lin, S. E. Shladover, PATH Program, Institute of Transportation Studies, University of California, Building 452, Richmond Field Station, 1301 South 46th Street, Richmond, Calif. 94804. H. Peng, T. Hessburg, M. Tomizuka, Department of Mechanical Engineering, University of California, Berekely, Calif. 94720. A. Arai, IMRA America, Inc., Institute of Transportation Studies, University of California, Building 452, Richmond Field Station, 1301 South 46th Street, Richmond, Calif. 94804.

AUTOMATED VEHICLE LATERAL CONTROL SYSTEM

The system used for experiments in the PATH-IMRA project includes both roadway-installed and vehicle-borne components. The discrete magnetic markers are installed in the roadway. The vehicle-borne components include the sensors, computer and control algorithms, and a steering actuator. Figure 2 shows the test vehicle with various components used in this experimental setup. The following describes these components.

Roadway Reference/Sensing System and Other Sensors

Several sensors are used in the PATH-IMRA experiments to measure the vehicle lateral displacement and other dynamic vehicle state information. These sensors include magnetometers for the discrete magnetic reference and sensing system and sensors that measure vehicle lateral acceleration and yaw rate.

The roadway reference and sensing system measures the vehicle's lateral displacement and provides a preview of road curvature, which has been determined to be an important input for vehicle lateral control (*2,4,5*). An assessment of all identified techniques for measuring the lateral deviation of a vehicle was made (*3*). A variety of technologies including optical, acoustic, and radar were reviewed. A roadway reference and sensing system using discrete magnetic markers was determined to have good potential for practical application on existing roadways. The simplicity of this system

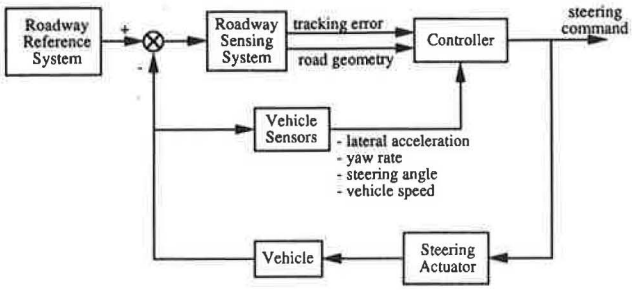

FIGURE 1 Automatic vehicle lateral control system.

leads to the expectation that the magnetic markers will be inexpensive and easy to install and maintain. The magnetic field from a magnetic marker appears to be less affected by environmental factors. More important, the reference system using magnetic markers can store roadway geometry information in a binary format, with each marker in a sequence representing 1 bit.

Figure 3 illustrates the configuration of the discrete magnetic marker reference and sensing system. The roadway reference consists of a series of permanent magnets installed at 1-m intervals along the center of the lane. The magnetic fields of the markers are measured to determine the lateral location of the vehicle relative to the center of the traffic lane. By using the polarities of the magnets, road geometry information such as radius and length of curvature can be encoded. The magnetic field sensing system consists of four magnetometers mounted under the front bumper of the vehicle 15 cm above the road surface. Two sensors, one measuring the vertical and the other measuring the horizontal component of the magnetic field, are located on the longitudinal axis of the vehicle. The other two sensors, which measure the vertical component of the magnetic field, are mounted on both sides of the vehicle longitudinal axis, about 30 cm from the two central sensors. The two central sensors are used to obtain accurate measurements of the vehicle lateral position; the outer two sensors were added to extend the measurement range. Figure 4 shows the sensors mounted on the test vehicle. The acquired signal is preprocessed by a low-pass filter and then digitized using an analog-to-digital (A/D) converter before being input to the controller.

Studies were conducted to investigate different approaches for extracting the vehicle lateral deviation information from the measured magnetic field. Both analysis and experiments indicate that the magnetic field of a permanent magnetic marker is sensitive to the distance between the magnetic marker and the sensor. To eliminate the effect of the vertical movement of the vehicle, both the vertical and horizontal components of the magnetic field are measured. The relationship between the lateral displacement of the sensor and the sensor measurements was determined by experiment. An algorithm based on this relationship has been developed that contains logic for recognizing the magnetic field of the marker and a look-up table for translating the magnetic field into a measurement of the vehicle lateral displacement. The theory and the experimental results were reported elsewhere (*4*).

The magnetic field sensing system, along with the roadway reference markers, can measure lateral displacement over a 100-cm range. The simulation analysis and bench experiments indicate that the resolution of this sensing system is reasonably good (about 1 cm) when the vehicle is near the center of the lane, but it degrades as the vehicle's lateral displacement increases. This degradation in accuracy at larger vehicle deviations should not affect the performance of the lateral control system because when the vehicle is far from the lane center, the controller does not need highly detailed information. In this experimental study, an independent measurement system consisting of two line-scan video systems, one mounted at the front and one at the rear of the vehicle, and a reflective tape strip placed on the roadway parallel to the roadway reference system is used for calibration of the discrete magnetic marker reference and sensing system and evaluation of lateral control system performance. Figure 5 shows the experimental vehicle with the independent measurement system.

A coding and decoding strategy has been developed to encode the road geometry information into the roadway ref-

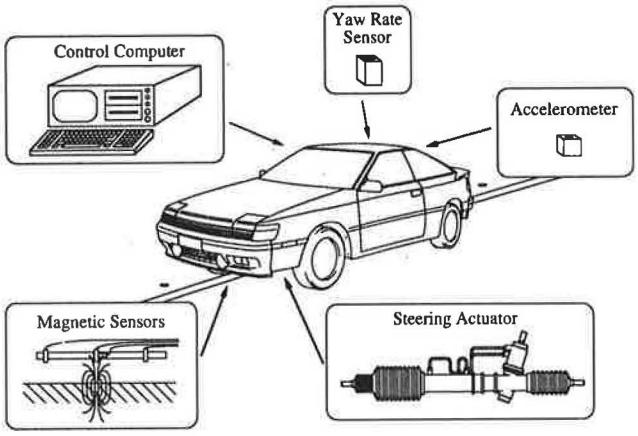

FIGURE 2 PATH-IMRA experimental vehicle.

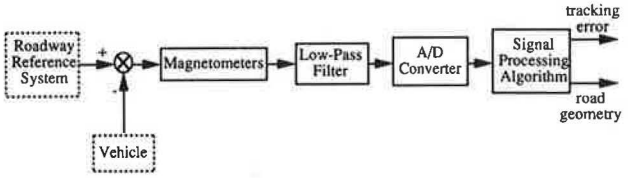

FIGURE 3 Discrete magnetic reference sensing system.

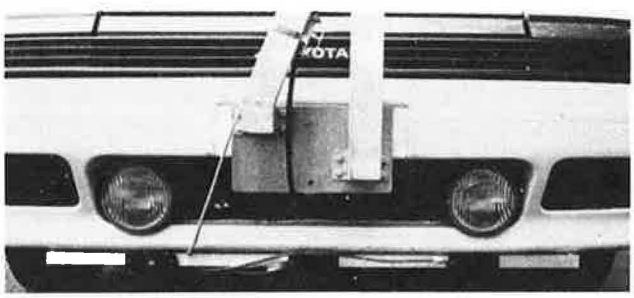

FIGURE 4 Magnetic sensors mounted on test vehicle.

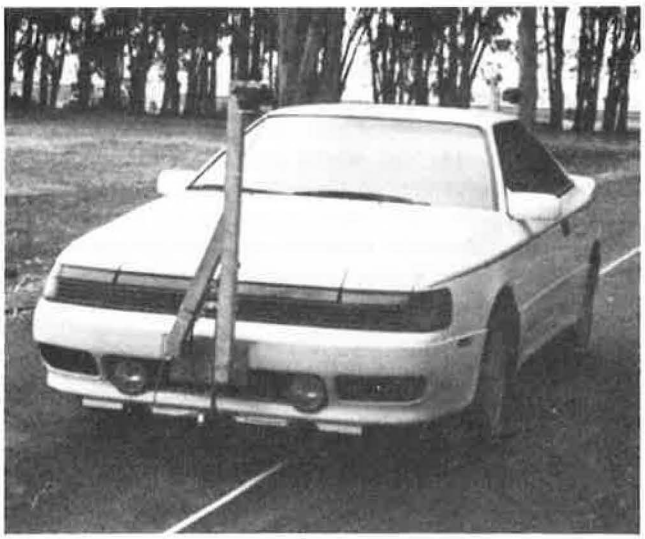

FIGURE 5 Experimental vehicle with independent measurement system.

erence system using the polarities of the magnetic markers. Using this method, after the vehicle passes over a series of markers, the road geometry information is decoded and passed to the controller for preview control. The number of markers necessary to code the information depends on the amount of information desired and the coding strategy.

In addition to the magnetic field sensing system, an accelerometer, a yaw rate sensor, and a steering angle sensor are also installed on the vehicle. The measurements of the vehicle's lateral acceleration, yaw rate, and steering angle are used to derive an estimation of the tire-road cornering force. The tire-road cornering force is represented by the tire cornering stiffness, which is defined as the ratio between the side force and tire slip angle. Simulation studies have shown that the steering control decisions should be made in accordance with several factors, including changes in the cornering stiffness. These results show that without considering the cornering stiffness in the control algorithm, the vehicle tracking performance deteriorates significantly when the cornering stiffness drops to 30 percent of its nominal value. Under the same situation, the vehicle's maximum displacement can be greatly reduced (about 50 percent) by incorporating the cornering stiffness estimation in the control algorithm.

Controller

The lateral control algorithm processes the information gathered by the sensors—road curvature, vehicle lateral deviation, yaw rate, lateral acceleration, forward speed, and front-wheel steering angle—and generates appropriate steering commands. The control algorithm balances the opposing goals of tracking accuracy and ride quality, where the ride quality is based on the lateral acceleration in a certain frequency range (*6*). The control algorithms must be designed in such a way that the system performance is robust with respect to variations in vehicle speed, vehicle load, road surface condition, and external disturbances such as wind gusts. To study the vehicle response and to derive control algorithms, two vehicle dynamics models have been developed.

The two mathematical models are used to represent the lateral dynamics of a front-wheel-steered, rubber-tired vehicle (*1*). In order to represent the vehicle dynamics, an elaborate nonlinear model was developed. This model includes the motion of the vehicle body in all six degrees of freedom plus suspension deflections and wheel motions. As a crucial element of the nonlinear model, a tire model was developed on the basis of data obtained from the tests of the tires used on the experimental vehicle. A simplified linear model that includes only lateral and yaw motions was also derived. The linear model was used to develop the feedback and feedforward controllers, and the nonlinear model was used to evaluate the performance of the control algorithms before the field tests. The steering actuator dynamics were incorporated into the plant model and are considered in the design of the control algorithms.

Two control algorithms have been developed and simulated, including a pure feedback control algorithm and a preview control algorithm. In the feedback control algorithm, the frequency-shaped linear quadratic (FSLQ) control theory was used (*6*). This control algorithm allows the frequency dependence of ride quality to be included in the performance index explicitly. By properly choosing the weighting in the performance index, the high-frequency robustness (with respect to unmodeled dynamics) of the control system can be enhanced.

The preview control algorithm combines a feedforward algorithm with feedback. The feedforward algorithm was designed to take advantage of the road geometry information available through the roadway reference system and is able to negotiate a curve without steady-state error. In this case, less feedback action is required, and better tracking performance can be achieved with improved ride quality. Analysis and simulation indicate that the preview control law improves the tracking performance in the low-frequency region and the lateral acceleration in the high-frequency region simultaneously (*7*).

The lateral dynamics of the vehicle strongly depend on the tire cornering stiffness (*6*) and the longitudinal velocity of the vehicle. A control algorithm that takes this into account was developed. A parameter identification scheme is used to estimate the cornering stiffness from the measurement of lateral acceleration, yaw rate, vehicle speed, and front-wheel-steering angle based on the dynamic equations of the system. The gain-scheduling technique is used to tune the feedback and feedforward (preview) controllers using the measured velocity and estimated cornering stiffness. Figure 6 depicts the block

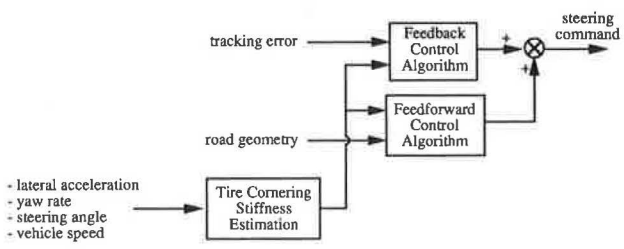

FIGURE 6 Preview control algorithm.

diagram of a lateral controller using the preview control algorithm.

PATH is also working on a fuzzy rule–based controller (*8*). This algorithm is currently being simulated on the nonlinear vehicle model. A fuzzy rule–based controller has the advantage of controlling a system based on rules formulated in a natural, linguistic setting. Also, a variety of linguistic inputs can be used efficiently in the rule base. This rule base addresses vehicle robustness to variations in vehicle parameters and to disturbances such as wind gusts. For curved sections of roadway, the rule base uses preview information provided from the discrete magnetic marker reference and sensing system.

The control algorithms and the signal preprocessing algorithm are implemented on a computer-based controller. This controller consists of a computer and a data acquisition system that includes a 32-channel A/D conversion board and a signal conditioning board. The sensor signals are first filtered through the signal conditioning board and then are acquired by the A/D conversion board at a fixed sampling rate. Upon receiving the digitized inputs, the computer converts the sensor signals into useful information and executes the control algorithm and generates steering commands accordingly.

Steering Actuator

The steering actuator steers the vehicle's front wheels on the basis of command signals issued by the controller. The steering actuator positioning is handled by the steering actuator controller that interfaces with the lateral control computer. IMRA's steering actuator is based on a standard rack-and-pinion steering system, modified to include a hydraulic servo, shown in Figure 7. In this design, the driver's steering input is in series with the computer controlled steering actuation. The controlled steering angle is limited to about 10 percent of the full range of normal steering angle. The series control arrangement and the limited range of the controlled steering angle were selected to allow the driver to override the control system at any time. This design is intended to allow the steering actuator to augment the driver's steering performance on the highway and as such should be limited in the latitude of its control. Further tests and evaluations from the viewpoints of safety, human factors, and system dynamics are necessary to determine the steering-angle range needed for automated steering under various operating conditions.

For the purpose of testing the lateral control algorithm, it was necessary to eliminate the influence of the driver on the steering system while the steering was under computer control. A spring-loaded detent mechanism was designed to attach to the steering column housing. The detent fixes the position of the steering wheel during testing. The spring load was set so that the driver can overcome the detent force and take control of the steering of the vehicle. The vehicle still includes the standard hydraulic, power-assist system so that the steering performance of the vehicle without the lateral controller active and with the detent disengaged is the same as that of a standard vehicle. Braking and acceleration remain under the control of the driver.

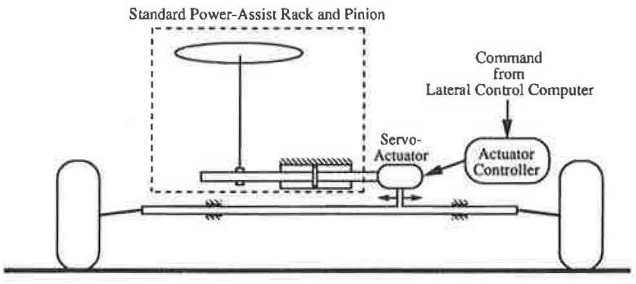

FIGURE 7 Steering actuator system.

EXPERIMENTAL STUDIES ON VEHICLE LATERAL CONTROL SYSTEM

The specifications of the components, systems, and methods used for the experimental studies were based on simulation and on previous experiments conducted by both PATH and IMRA researchers. Before this project, many bench tests and full-scale experiments on the individual components of the proposed system, such as the discrete magnetic marker reference and sensing system, the control algorithm, and the steering actuator, were conducted separately by PATH and IMRA.

Test of Discrete Magnetic Marker Reference and Sensing System

Bench tests and full-scale experiments on the discrete magnetic marker reference and sensing system have been conducted by PATH in order to verify the concept under a controlled environment and in a wide speed range. A test bed consisting of a rotary arm driven by a motor in a circular path in the horizontal plane was used to test the reference and sensing system. In these tests, a magnet was fixed vertically to the end of the arm. A bench that held the magnetometers was located such that the magnetic marker passed over the sensors on each rotation. Both the vertical and horizontal distance between the sensor and the magnetic marker were adjustable. Signals were collected at different speeds (up to 130 km/hr) and various distances (10 to 20 cm). The data were used to verify the feasibility of such system under a controlled environment over a wide speed range and to develop a signal processing algorithm and filters. The bench experiments were followed by a full-scale experiment. Magnets were installed in 200-m stretch of a road at the Richmond Field Station of the University of California, Berkeley (UCB). An experimental vehicle equipped with a set of magnetic sensors and a portable computer was used to collect a data at vehicle speeds of up to 40 km/hr. Figure 8 shows the vertical and horizontal signals collected from the roadway magnetic reference system and the estimated lateral displacement of the vehicle. The experimental results indicate that the discrete magnetic marker reference and sensing system can provide adequate lateral displacement information under the conditions tested (*4*). In the PATH-IMRA experiments, the accuracy of the roadway reference and sensing system will be verified using a measurement system that is independent of the vehicle control system.

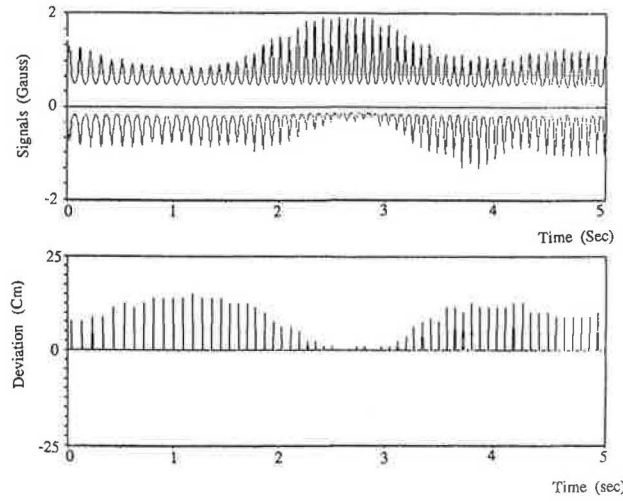

FIGURE 8 Magnetic signals (*top*) and calculated vehicle deviation (*bottom*).

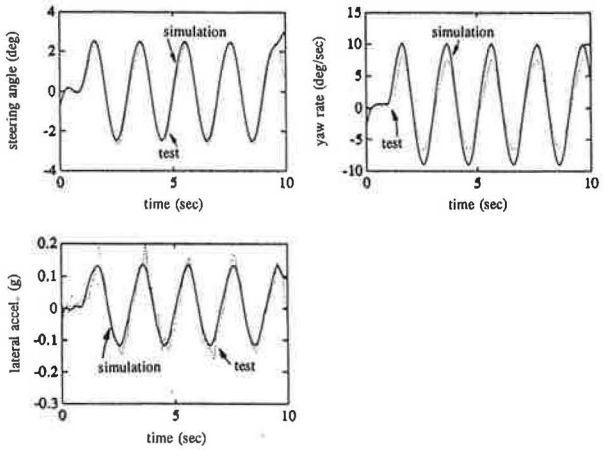

FIGURE 9 Open-loop experiment results.

Simulation of Control Laws and Tests Using Scaled Experimental Vehicle

Vehicle lateral control using feedback and preview control laws was simulated on a computer for a variety of road curvatures and under different road surface conditions. Simulations were also conducted to fine-tune the control laws using realistic vehicle parameters.

As a step before the full-scale vehicle tests, an experimental study on the proposed system was conducted using a scaled test vehicle. In this study, the accuracy of the roadway reference and sensing system and the performance of the control algorithm were evaluated over a wide range of extreme operating conditions such as variation of load, vehicle speed, and cornering stiffness. In the experiments, a PID controller with feedforward action was able to keep the tracking error within ± 10 cm at a vehicle speed of 2 m/sec. The test track that was used for the experiments included a curved section with a 4.5-m radius. The detailed experimental results are reported elsewhere (*2*).

Open-Loop Experiments

PATH and IMRA have conducted open-loop experiments to identify vehicle parameters and evaluate the performance of the steering actuator under normal load conditions. In the experiments, the steering actuator position was set by the control computer. Selected steering commands were sent to the steering actuator in the form of step and sinusoidal inputs. Input amplitudes and frequencies were tested at vehicle speeds ranging from 20 to 60 km/hr. The response of the vehicle—including the position of the steering actuator—the lateral acceleration, and the yaw rate were recorded at a sampling rate of 100 Hz. The tests were repeated to ensure the reliability of the data. Figure 9 gives a set of test results at a speed of 40 km/hr with a 0.5 Hz sinusoidal steering input. The test data have been used to validate the vehicle dynamics model and the control algorithms. In the analysis, the steering system and vehicle parameters were identified. The vehicle models were modified on the basis of the test data. In Figure 6, the dotted lines show the test results and the solid lines indicate the simulation results. These results show that the modified overall vehicle model predicts the vehicle response satisfactorily for the current test conditions.

Planned Closed-Loop Experiments

The closed-loop experiments will be performed at the UCB Richmond Field Station. A 480-m test track has been built. Figure 10(*a*) shows the geometric layout of the test track. The vehicle path will be defined by discrete magnetic markers. To use the test track effectively, two curves with radii of 60 and 75 m have been placed at the location of the 90-degree curve and several gentle reverse curves have been placed in one of the straight sections. Several hundred capsules, shown in Figure 10(*b*), have been installed on the test track. The capsules are designed to house the magnetic markers and to allow the

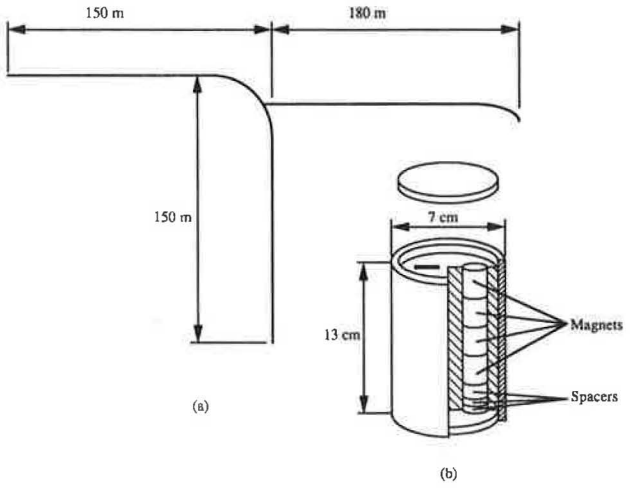

FIGURE 10 Test track layout (*a*); capsules for magnetic markers (*b*).

location of the markers to be adjustable so that the effect of placement accuracy of the magnetic marker on the performance of the lateral control system can be tested. This design also allows the road geometry information encoded in the magnetic reference system to be changeable. However, it should be noted that an operational system using the magnetic marker reference system would not need the capsules but would have the simple markers (10 cm long, 2.5 cm in diameter) installed directly in the pavement.

The closed-loop experiments are designed to examine the tracking performance of the lateral control system using the proposed sensing devices, control algorithm, and steering actuator. Several tasks will be performed in the closed-loop experiments. These tasks are

1. Testing of the feedback control algorithms: The tests will be conducted at various vehicle speeds from 20 to 60 km/hr.
2. Testing of the preview-FSLQ control law: The experiments in Task 1 will be repeated using the preview-FSLQ control law.
3. Determination of marker spacing: Experiments using both feedback and preview lateral control with different marker spacings will be conducted to determine the effect of different marker spacing.
4. Evaluation of the robustness of the integrated control system: Lateral control under some nonideal conditions will be performed. In performing this task, several nonideal conditions are created, including adjusting the locations of magnetic markers to create errors and reducing the pressure of some or all of the tires to vary the cornering stiffness. Tests will be conducted to evaluated the performance of the system with respect to various levels of resolution of the coded upcoming road geometry information.

These tasks are expected to be performed in early 1992. From these experiments, the proposed lateral control system using the previously described components and control algorithms will be evaluated. The performance specifications for the sensing systems and actuating devices will be verified. The performance of the system with different control algorithms will be compared. The dynamic models can be further validated so that they can be used to more accurately simulate the performance of the lateral control system under a wide range of conditions under which experiments cannot be easily performed.

CONCLUDING REMARKS

Extensive research in vehicle lateral control through computer simulation has shown that this technology has potential. To bridge the gap between computer simulation and full-scale hardware implementation in a roadway environment, PATH and IMRA initiated the collaborative experimental research project described in this paper. To date, the various components necessary for vehicle control and performance evaluation have been developed, tested separately, and integrated into the vehicle system. The project is now in its final stage of preparation for full-scale testing. The full-scale closed-loop experiments are planned to be conducted in early 1992. The test results will be reported in a future paper.

ACKNOWLEDGMENTS

The portion of research conducted by PATH is sponsored by the state of California, Business Transportation and Housing Agency, Department of Transportation. The authors wish to thank R. Bushey, M. Dearing, R. Parsons, and C. Price for their advice and assistance and L. Bell for his efforts in preparing the test facility.

REFERENCES

1. S. Shladover et al. Automatic Vehicle Control Developments in the PATH Program. *IEEE Transactions on Vehicular Technology*, Vol. 40, No. 1, Feb. 1991, pp. 114–130.
2. T. Hessburg, H. Peng, M. Tomizuka, W. Zhang, and E. Kamei. An Experimental Study on Lateral Control of a Vehicle. *Proc., 1991 American Control Conference*, Boston, Mass., 1991.
3. R. Parsons and W.-B. Zhang. PATH Lateral Guidance System Requirements Definition. *Proc., 1st International Conference on Application of Advanced Technology in Transportation Engineering*, ASCE, San Diego, Calif., Feb. 1990, pp. 275–280.
4. W.-B. Zhang, R. Parsons, and T. West. An Intelligent Roadway Reference System for Vehicle Lateral Guidance/Control. *Proc., 1990 American Control Conference*, San Diego, Calif., May 1990, pp. 281–286.
5. A. Y. Lee. A Preview Steering Autopilot Control Algorithm for Four-Wheel-Steering Passenger Vehicles. *Advanced Automotive Technologies—1989*, ASME, pp. 83–98.
6. H. Peng and M. Tomizuka. *Lateral Control of Front-Wheel-Steering Rubber-Tire Vehicles*. Publication UCB-ITS-PRR-90-5. PATH, Institute of Transportation Studies, University of California, Berkeley, July 1990.
7. H. Peng and M. Tomizuka. Preview Control for Vehicle Lateral Guidance in Highway Automation. *Proc., 1991 American Control Conference*, Boston, Mass., 1991, pp. 3090–3095.
8. T. Hessburg and M. Tomizuka. A Fuzzy Rule Based Controller for Automotive Vehicle Guidance. *Fuzzy and Neural Systems, and Vehicle Applications '91*, Japan, Nov. 1991.

Publication of this paper sponsored by Task Force on Advanced Vehicle and Highway Technologies.

Concept of Super Smart Vehicle Systems and Their Relation to Advanced Vehicle Control Systems

SADAYUKI TSUGAWA

The Super Smart Vehicle System (SSVS), proposed under support of the Japanese Ministry of International Trade and Industry, is a new information system that 20 to 30 years from now will solve problems caused by automobiles and automobile traffic. Its solutions will be based on previous research and development of driver information systems and vehicle control systems in Japan. The current status and problems of automobiles and automobile traffic in Japan are described. Large cities in Japan are characterized by excessively high density, which is the main cause of the problems. Then histories of vehicle control systems are explained. Japan has a long history in vehicle control systems, which include driver assistance systems and automatic driving systems. Advanced vehicle control systems (AVCSs) are a main research theme of the SSVS. The SSVS deals with four fields: information systems for a single vehicle, for inter-vehicles, and for vehicle-to-road relations; and studies on vehicle-to-driver relations. Some of system candidates relating to AVCS proposed for the SSVS are introduced.

Traffic accidents, congestion, and pollution caused by automobiles have been becoming more serious in Japan. Many systems for traffic management, driver information, and vehicle control have been studied and developed in Japan since the 1960s in efforts to solve the problems. Traffic control and surveillance systems were already installed in major cities in the 1970s. The number of control centers in Japan runs into 160. In addition, experiments of navigation systems have been conducted. A large-scale experiment of a dynamic route guidance system called the Comprehensive Automobile Traffic Control System (CACS) (*1*) was conducted in Tokyo in the 1970s; it aimed at increasing the efficiency of automobile traffic and decreasing traffic accidents and congestion. Experiments of the Advanced Mobile Traffic Information and Communication System (AMTICS) (*2*) and the Road/Automobile Communication System (RACS) (*3*), both navigation systems, were conducted in the mid-1980s. Now onboard navigation systems are commercially available.

Vehicle control systems such as collision warning systems and automatic driving systems have been studied in Japan since the 1950s, and some of them have been commercially available. However, automatic driving systems still remain in the area of research.

Mechanical Engineering Laboratory, Agency of Industrial Science and Technology, Ministry of International Trade and Industry, Namiki 1-2, Tsukuba-shi, Ibaraki-ken, 305 Japan.

For solving the problem, we have proposed an information system for automobiles and automobile traffic for 20 to 30 years from now. The Super Smart Vehicle System (SSVS) (*4*), based on the info-mobility concept (*5*), is an information system for drivers. Its purpose is making safety and efficiency compatible while taking aging and pollution into account.

The current status of automobile traffic and the history of vehicle control systems in Japan will be introduced; these are the background of the SSVS proposal. After the background and the info-mobility concept are introduced, the SSVS will be explained; some systems relating to advanced vehicle control systems (AVCSs) will be described in detail.

STATUS OF AUTOMOBILE TRAFFIC IN JAPAN

Japan is a mountainous country with little flatland, which leads to an excessive density of population and automobiles in cities. This high density, a major characteristic of Japan, has caused the problems.

Automobiles and Japanese Society

The number of automobiles has been rapidly increasing in Japan in recent years, as shown in Figure 1, and it became 58 million in October 1990. It is expected that there will be 64 million automobiles in 1995 and 72 million in 2000.

The role of automobiles has increased as well. Figure 2 shows the volume of passenger traffic by various means of transportation. The volume of passenger traffic by automobile has increased in indexes of the number of passengers and the passenger distance of travel.

The trend in cargo traffic by automobile is similar to that of passenger traffic by automobile. The ratio of cargo traffic by motor trucks is more than 90 percent in weight and more than 50 percent in weight distance. The role of motor trucks is much larger in Japan than in Europe or the United States.

Although the number of automobiles and drivers has been increasing, roads are not necessarily sufficient. The length of paved roads per vehicle is 14 mi in Japan, 17 mi in western Germany, and 20 mi in the United States. The length of expressways per 100 vehicles is 8 mi in Japan, 29 mi in western Germany, and 47 mi in the United States.

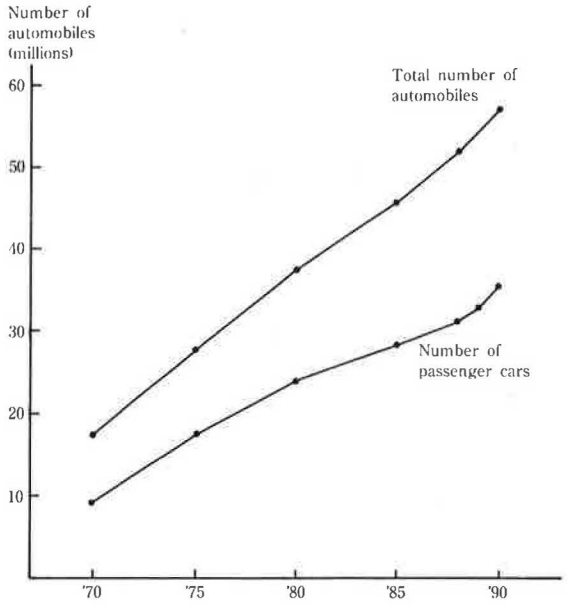

FIGURE 1 Numbers of automobiles (four-wheeled vehicles) and passenger cars.

Problems of Automobile Traffic

Accidents

As shown in Figure 3, the annual number of fatalities by traffic accidents began to increase in 1987, and it exceeded 10,000 in 1988. Fatal accidents have the following characteristics: first, there is a significant increase in fatalities for people between 16 and 24 and over 65 years of age. Second, fatalities of drivers and passengers have increased much more than those of pedestrians. Third, nighttime accidents have increased much more than daytime accidents, and the ratio of fatal accidents at night is higher than that in the day. Finally, accidents have increased during weekends. Sundays accounted for the most fatalities in 1989, and Saturdays did in 1990. Table 1 summarizes the characteristics of the accidents.

In addition, accidents and fatalities on expressways have also been increasing. Since the expressways became longer, obviously the total number of accidents and fatalities has increased. However, the ratio of accidents and fatalities on the expressways with the index of vehicle distance has also increased.

Congestion

Traffic congestion is also a serious problem, not only in large cities such as Tokyo and Osaka but also in rural cities. In Tokyo the average delay at one major intersection was 3.3 hr/day in 1983 and 5.1 hr/day in 1989. Also in Tokyo along urban expressways the average delay at one point was 4.6 hr/day in 1983 and 7.9 hr/day in 1989.

The delays caused by traffic congestion have resulted in a great deal of loss in social and economical activities. It was estimated that all the driving hours including the delays were 20.37 billion person-hr/year in 1980, but that it would have been 15.77 billion person-hr had there been no traffic congestion. Thus, the loss was 4.6 billion person-hr/year, which equals 2.19 million workers.

Factors in Future Automobile Traffic

Because the problems caused by automobiles and automobile traffic have become serious even recently, it is expected that they will become much more serious because the number of automobiles and drivers will grow—even though constructing roads will become harder because of the high price and shortage of lands. In addition, some factors will affect automobile traffic in the future: aging and the increase of female drivers and foreign people.

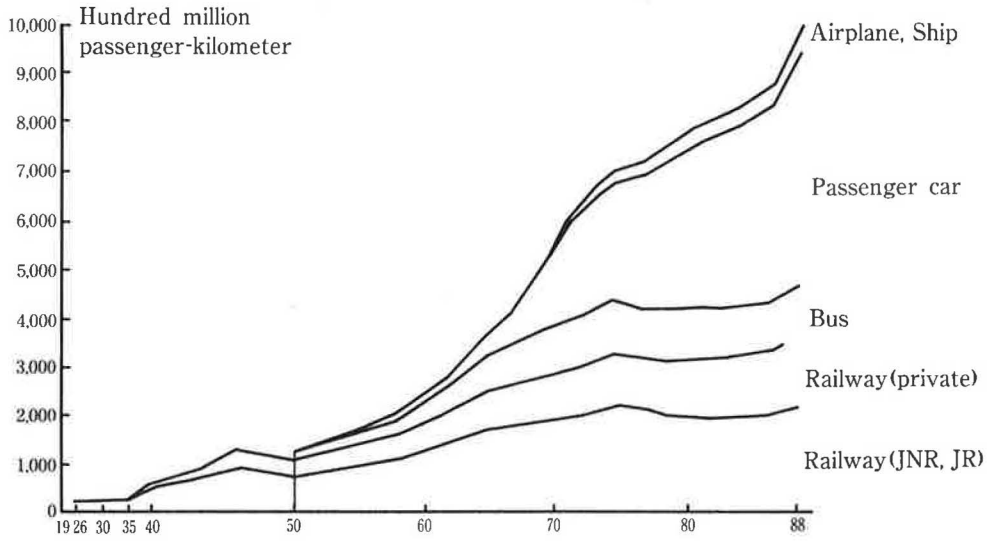

FIGURE 2 Passenger traffic.

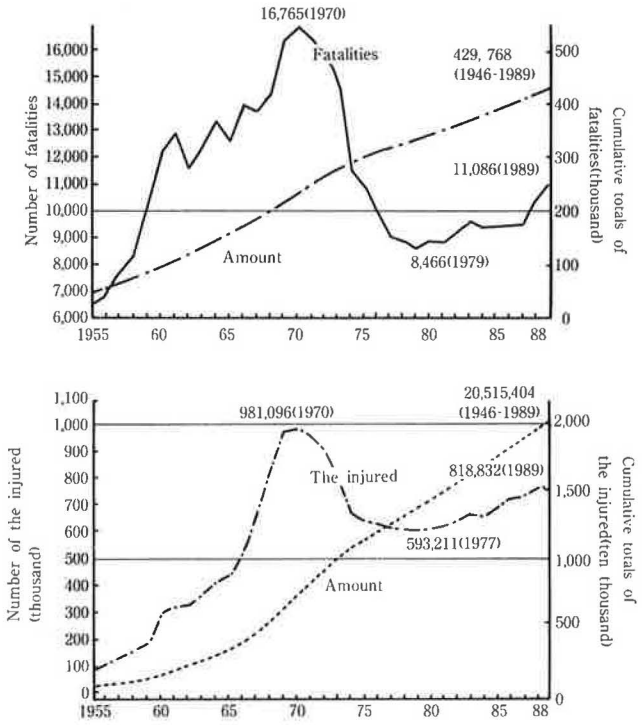

FIGURE 3 Numbers of (*top*) fatalities and (*bottom*) injuries from traffic accidents.

Aging is an important issue in automobiles and automobile traffic in terms of the increase of aged drivers and shortage of labor force, in particular, drivers of heavy-duty trucks. It is expected that at the beginning of 21st century, 20 percent of the people will be aged.

The increase of female drivers should also be taken into account. Their rate of increase is much larger than that of male drivers. The ratio of female drivers in the female population was 27.9 percent in 1980 and 43.5 percent in 1989. Besides, fatalities of female drivers have increased at a higher rate than that of male drivers. In 1989 the number of fatalities of female drivers was 362, which was 2.7 times the number in 1979; meanwhile, the number of female drivers increased 1.9 times.

The third factor is the increase of foreign people and foreign drivers. This requires that information for drivers be not only in Japanese but also in English, for example. Actually, message signs in Japanese as well as English are becoming common.

TABLE 1 Characteristics of Accidents

		1979	1990
Number of fatalities	all ages	8466(100)	11227(133)
	ages between 16 and 24	1845(100)	3158(171)
	ages over 65	1613(100)	2673(166)
Number of fatalities of	drivers/passengers	2998(100)	4501(150)
drivers and pedestrians	pedestrians	2888(100)	3042(105)
Number of accidents	daytime	349536(100)	437134(125)
in the daytime and nighttime		[4071(100)]	[4610(113)]
[ ]:fatal accidents	nighttime	122110(100)	205963(169)
		[3977(100)]	[6041(152)]
Number of fatal accidents per day	weekdays	21.1	27.9
in weekdays and at weekends	weekends	24.4	32.4

HISTORY OF VEHICLE CONTROL SYSTEMS IN JAPAN

System candidates proposed in the SSVS study have roots in vehicle control systems that have been developed since the 1960s. These are technological backgrounds for the proposal of the SSVS. The vehicle control systems can be classified as driver assistance systems and automatic driving systems. The driver assistance systems include a collision warning system and a lane detection system. The automatic driving systems include an automated vehicle system with inductive cable and an autonomous vehicle with machine vision.

Driver Assistance Systems

Studies on driver assistance systems have been conducted in many fields in Japan. Obstacle warning systems in the vicinity of an automobile using ultrasonic are already commercially available. Here, a laser radar system and a lane detection system based on machine vision will be explained. Although radar systems were studied also in Japan and many papers have been published, the systems have not yet become commercially available. In addition, driver perceptual enhancement systems have not yet been opened in Japan; these will not be referred to, either.

Laser Radar System

After the first development of a laser radar system for obstacle warning using a semiconductor laser in Japan (*6*), a new laser radar system was developed for heavy-duty trucks (*7*). The gap between the truck and the leading vehicle within 100 m is measured by the emission of infrared laser pulses of 70-nsec duration at 6 kHz. Their speeds and decelerations are recorded, and the driver is warned when necessary.

Lane Detection System

Lane detection systems have been used for visual navigation systems of mobile robots and intelligent vehicles. In the Personal Vehicle System (PVS; more detail later), white lines beside a road or lane markers are detected for lateral control.

Here, a lane marker detecting system for warning a driver of lane deviation is described. A study of a lane detecting system using machine vision with stereo television cameras has been conducted in the Mechanical Engineering Laboratory (MEL) (*8*). The system is characterized by real-time operation to detect lane markers in the field of view with a range of 4.5 to 21 m. It consists of two television cameras and hard-wired logic for video signal processing. Figure 4 shows

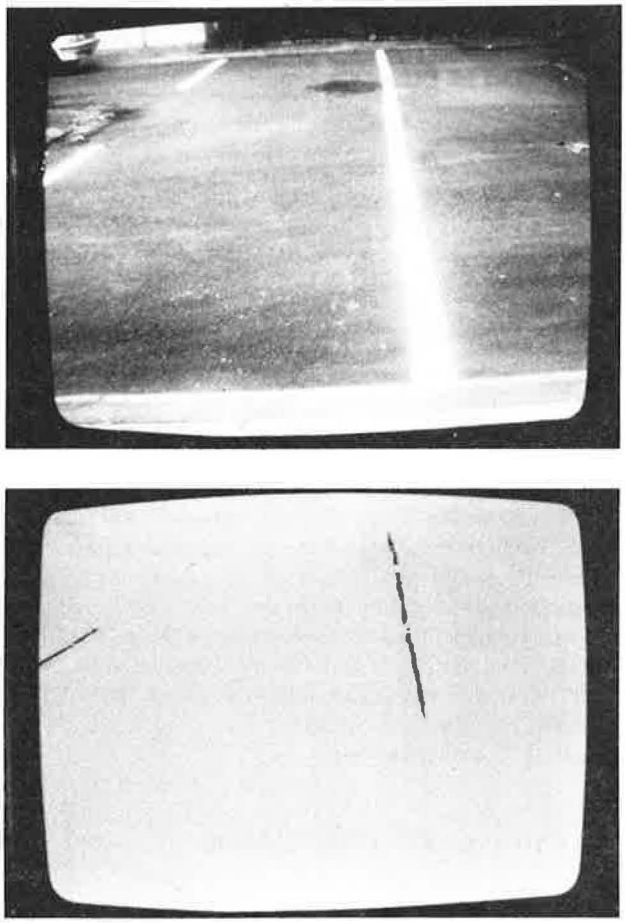

FIGURE 4 Detection of lane markers: (*top*) an original scene and (*bottom*) detected lane markers (courtesy of Takeshi Hirose, MEL).

an original scene and lane markers, where only lane markers are detected.

Automatic Driving Systems

Automated Vehicle with Machine Vision

Studies on automatic driving systems were started from the employment of inductive cable systems in Japan, as in the United States and Europe. In 1967 an automated vehicle developed by MEL was driven stably at 100 km/hr (*9*).

After the research on the automated vehicle with inductive cable, to eliminate restrictions on the system with inductive cable, a new system using machine vision was begun at MEL in the 1970s. The world's first automated vehicle with machine vision was developed at MEL, and it was named the Intelligent Vehicle (Figure 5). It ran autonomously at 30 km/hr while avoiding obstacles and guardrails (*10*). The steering control is based on table look-up. The machine vision consists of stereo television cameras and hard-wired logic for processing video signals to detect obstacles in the field of view between 5 to 20 m with a viewing angle of 40 degrees. The obstacle detection system features real-time processing. After scanning one image, which takes 33 msec, the presence and locations of obstacles are detected with another 2 msec of processing. The principle of obstacle detection is based on parallax: an obstacle yields images of the same heights on television cameras, but figures on a road yield those of different heights because of the positions of the television cameras on the vehicle. Figure 6 shows an original image of a scene and the location of obstacles in the quantized field of view.

In 1984 the Intelligent Vehicle was equipped with a dead-reckoning system using a differential odometer. Thus, the vehicle had functions of navigation and obstacle detection, which led to a completely autonomous vehicle (*11*). The vehicle was driven autonomously from its starting point to its goal while avoiding an obstacle after its goal was assigned and route planning was performed.

The Intelligent Vehicle was an experimental system. It was driven only under well-defined conditions; the performance entirely depends on weather, brightness, and direction of light.

Personal Vehicle System

In 1987 a study on a new Intelligent Vehicle named the PVS (*12*) was started. Figure 7 shows the PVS. It has functions of navigation and obstacle detection. Machine vision of the PVS has five television cameras; three of them are for detecting white lines indicating a lane, and the other two are for detecting obstacles. The PVS also has a dead-reckoning system that uses a differential odometer to locate its position. It enables the PVS to navigate from its starting point to its goal along an optimal route. In addition, the PVS has another obstacle detection system with a laser radar system. In 1989

FIGURE 5 Intelligent Vehicle.

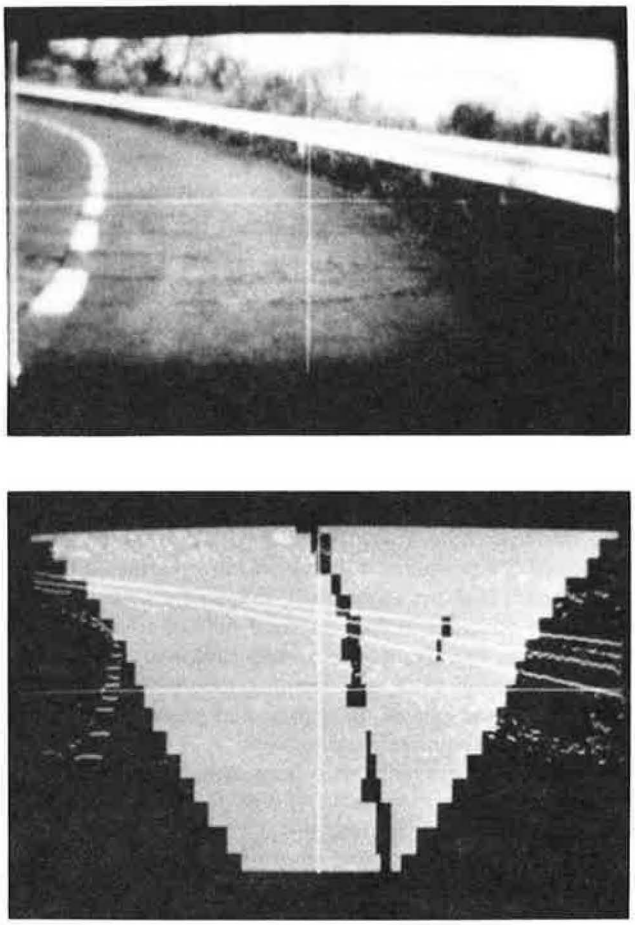

FIGURE 6 Obstacle detection for Intelligent Vehicle: (*a*) an original scene and (*b*) obstacles (guardrail) detected in the field of view (courtesy of Takeshi Hirose, MEL).

the PVS ran at 30 km/hr along straight lanes and 10 km/hr along curved lanes on a special proving ground and at 60 km/hr along a straight lane on another proving ground.

The PVS was remodeled in 1990 and 1991: the three cameras for lane detection were removed and a single camera on a turntable was installed inside the windshield. Experiments under rainy conditions or at night were conducted.

Vehicle-to-Vehicle Communication Systems

Vehicle-to-vehicle communication systems have a wide range of variations; they are applicable to many systems in driver information systems and vehicle control systems. In Japan, vehicle-to-vehicle communication has been studied at the Association of Electronic Technology for Automobile Traffic and Driving (JSK) since 1981 (*13*).

After the studies on the Intelligent Vehicle, an application of vehicle-to-vehicle communication to control of a group of autonomous vehicles was started at MEL in 1984 (*14,15*). It aims at a vehicle-following system with small gaps between vehicles. The system is called the soft-linked vehicle system after linking with vehicle-to-vehicle communication. Experiments using automated guided vehicles for factory automation

FIGURE 7 PVS.

were conducted to investigate the vehicle-to-vehicle communication with infrared and the control algorithms.

Simulation studies on a vehicle-following system with vehicle-to-vehicle communication have also been conducted (*16*). Simulation results of comparisons between vehicle following with the communication and vehicle following with human drivers show that the vehicle-following system with vehicle-to-vehicle communication helps increase road capacity and decrease rear-end collisions.

SSVS AND AVCS-RELATED SYSTEMS

Info-Mobility Concept

Recently we proposed a framework for info-mobility (*5*). The info-mobility system consists of information systems covering a mobility system. As shown in Figure 8, the mobility system consists of a driver subsystem, a vehicle subsystem, and a road environment subsystem. These subsystems are essential to automobile driving. However, the mobility system does not suffice for safe and efficient driving. It is pointed out that traffic accidents and congestion are caused by discord—or "gaps"—among the three subsystems. What fills the gaps is information systems; therefore, information systems make automobile driving safer and more efficient. The info-mobility consists of the mobility system and the information systems to cover it. As pointed out in intelligent vehicle-highway sys-

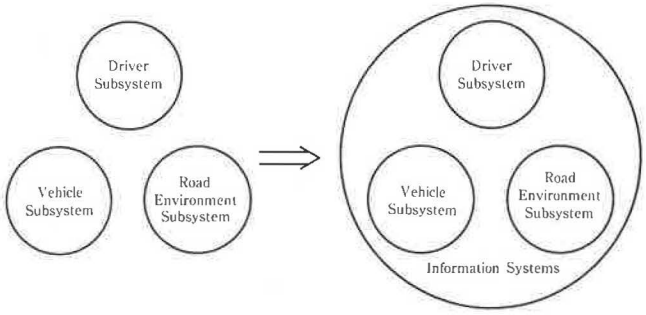

FIGURE 8 Mobility system (*left*) and info-mobility system (*right*).

tems, the information systems include advanced traffic management systems, advanced driver information systems, and AVCSs. Advanced traffic management systems are located mainly at the road environment subsystems, the advanced driver information systems are located mainly between the driver subsystem and the road environment subsystem, and the AVCSs are located mainly around the vehicle subsystem.

The SSVS is one way to realize the info-mobility concept.

SSVS and AVCS

In 1990, JSK started the 2-year preliminary study program on the SSVS under support of the Japanese Ministry of International Trade and Industry to promote the research and development of information systems for automobiles and automobile traffic to be used in 20 to 30 years (*17,18*).

The SSVS has been proposed on the basis of the previous work on driver information systems such as the CACS and vehicle control systems such as the Intelligent Vehicle and the PVS. It is AVCS that will provide substantial solutions to the future problems of traffic accidents and congestion, because the effects of traffic management systems and driver information systems are indirect to the problems and, therefore, bounded. For example, the safety of automobile traffic at nonsignalized intersections will not be guaranteed to the same degree as at signalized intersections. In addition, a dynamic route guidance system will not effectively shorten traveling time when the ratio of automobiles with on-board equipment is more than 50 percent (*19*). Thus, the main theme of the SSVS will be AVCS.

AVCS-Related Systems

The research themes of the SSVS are

1. Information systems for a single vehicle,
2. Information systems for inter-vehicles,
3. Information systems for vehicle-to-road relations, and
4. Studies on vehicle-to-driver relations.

Combining the fields and the purposes of the SSVS, a cooperative driving system, a control configured vehicle system with ultra-little vehicles, an active driver assistance system, an intelligent intersection system, and intelligent logistics and sensor systems represented by machine vision have been proposed, as shown in Figure 9.

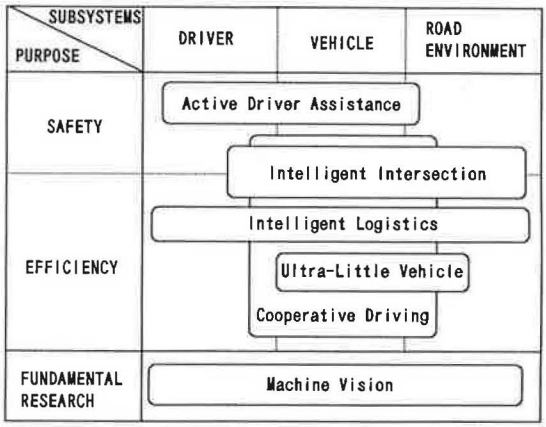

FIGURE 9 System candidates for SSVS.

Cooperative Driving System

The system coordinates driving of automobiles with radar systems and vehicle-to-vehicle communication systems to increase safety and efficiency. If each automobile communicates with other automobiles, and information is given to drivers while driving as shown in Figure 10, the effective road capacity would grow, lane changing and merging would be eased, and safety would be increased. Vehicle-to-vehicle communication is performed either directly among vehicles or indirectly through inductive cables under road surfaces or coaxial leakage cables beside roads.

Control Configured Vehicle System with Ultra-Little Vehicles

In the system, ultra-little vehicles with high performance are driven either independent of each other or linked to form platoons of two-by-one to four-by-four vehicles linked by mechanical coupling or vehicle-to-vehicle communication. The vehicles have functions of access to each other for linking using an omnidirectional radar and control for cooperative driving among vehicles linked together as well as drive-by-wire to ease driving under linking. The vehicle, either a one- or two-seater with a payload of 1 ton, for example, is about 0.8 m wide and 2 m long, which is half in the width and length of a normal automobile. The vehicle is shown in Figure 11. The capacity or payload is based on current use of automobiles, especially those in the downtowns of large cities. The system will help decrease congestion and increase the effective use of roads.

Active Driver Assistance System

The system not only assists a driver by indicating the presence of obstacles in the direction of movement as well as the presence of other vehicles and pedestrians near the driver's vehicle, but also drives automatically for a short time if the

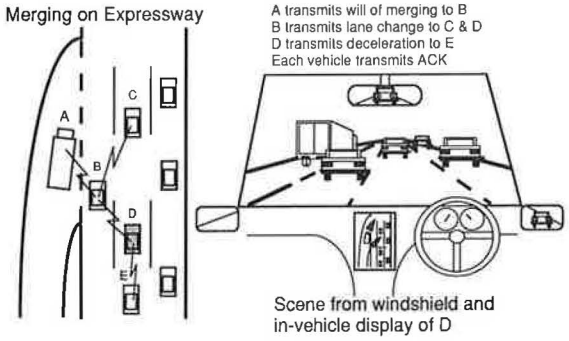

FIGURE 10 Cooperative driving system.

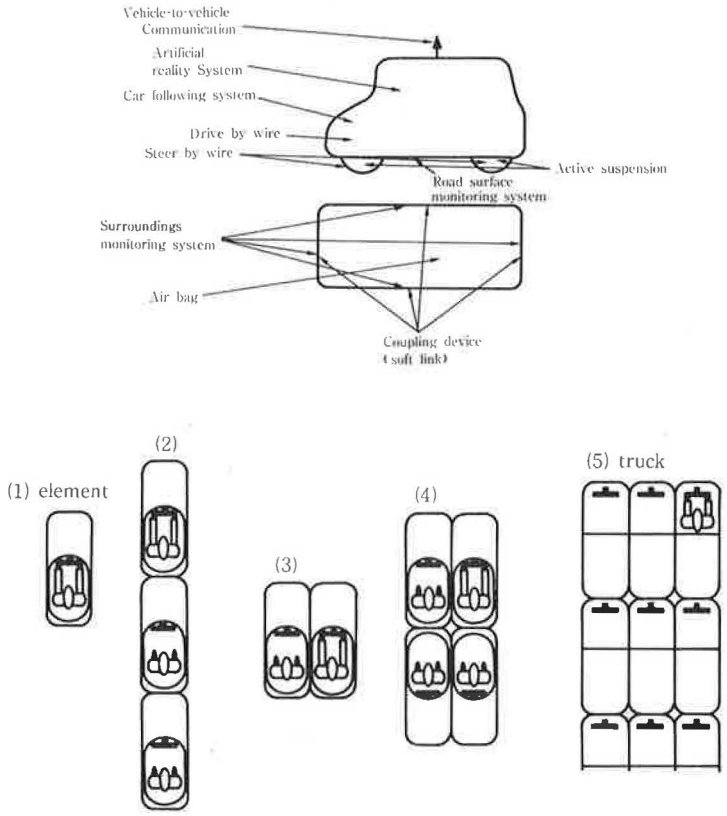

FIGURE 11 Control configured vehicle system with ultra-little vehicles: (*top*) element vehicle and its functions, and (*bottom*) examples of vehicle systems.

driver dozes off or suddenly gets sick. The system will be applicable to an assistance system at high-speed driving or for the aged and the handicapped. The vehicle is equipped with a sensor system including machine vision, a processing system for driver assistance and for automatic driving, a display system for a driver, and an actuator system to steer, accelerate, and brake the vehicle. In this system human factors will play an important role.

Automatic driving will be classified in two categories. One is a steady-state system in which the vehicle is driven without a driver as a default state. It can be called automatic chauffeuring. The other is an unsteady-state system, which is the active driver assistance system described here.

Intelligent Intersection System

It would be effective in decreasing accidents if intersections were made intelligent, because accidents at intersections account for 60 percent of all accidents in Japan. It would also increase road capacity. The intelligent intersection system warns drivers who are ignoring traffic signals, indicates the presence of pedestrians and bicycles to drivers, provides a bird's-eye view of an intersection to drivers, and tells the speeds and directions of vehicle movement—which is called a guide light system. The guide light system provides information on directions and speeds of vehicles by lengths of lights on roads.

CONCLUSION

Automobile traffic in Japan today is characterized by excessively high density: a narrow land, shortage of roads, and a great number of automobiles and drivers. Nevertheless, in Japan automobiles are one of life's necessities. The SSVS is aiming at a solution to the problems that will be caused by automobiles and automobile traffic 20 to 30 years from now. In this paper, based on the SSVS reports of FY 1990 and 1991, the status of automobile traffic in Japan was explained and the research and development history of the vehicle control systems was described. Drawing on previous work in Japan, we have proposed the SSVS. The SSVS is one way of realizing the new concept of info-mobility. Some AVCS-related systems proposed for the SSVS have been introduced. When developing the SSVS, driver acceptance and social acceptance of the systems should be considered.

REFERENCES

1. S. Matsumoto et al. Comprehensive Automobile Traffic Control System (in Japanese). *Journal of IECE*, Vol. 62, No. 8, 1979, pp. 870–887.
2. H. Okamoto. The Progress of AMTICS. *Proc., 22nd ISATA*, Vol. 1, 1990, pp. 215–222.
3. K. Takada et al. RACS: Results of the First Overall Field Trial. *Proc., 22nd ISATA*, Vol. 1, 1990, pp. 223–230.

4. S. Tsugawa et al. Super Smart Vehicle System—Its Concept and Preliminary Works. *Proc., 1991 Vehicle Navigation and Information Conference*, SAE, 1991, pp. 269–277.
5. S. Tsugawa et al. *Info-Mobility: A Concept for Advanced Automotive Functions Towards the 21st Century*. Paper 910112, SAE, 1991.
6. H. Endo. Japanese Patent 61-6349 (in Japanese). 1986.
7. M. Sakata et al. Laser Radar Rear-End Collision Warning System for Heavy-Duty Trucks. *Proc., ATA International Symposium on Future Automotive Microelectronics*, 1988.
8. T. Hirose. Position Prediction of Automobiles—White Line Detection (in Japanese). *Proc., 28th SICE Annual Conference*, Vol. 1, 1989, pp. 23–24.
9. Y. Ohshima et al. Control System for Automatic Automobile Driving. *Proc., IFAC Tokyo Symposium on Systems Engineering for Control System Design*, 1965, pp. 347–357.
10. S. Tsugawa et al. An Automobile with Artificial Intelligence. *Proc., 6th International Joint Conference on Artificial Intelligence*, 1979, pp. 893–895.
11. S. Tsugawa et al. An Intelligent Vehicle with Obstacle Detection and Navigation Functions. *Proc., IECON '84*, 1984, pp. 303–308.
12. M. Taniguchi et al. The Development of Autonomously Controlled Vehicle, PVS. *Proc., 1991 Vehicle Navigation and Information Conference*, 1991, pp. 1137–1141.
13. M. Aoki et al. An Inter-Vehicle Communication Technology and Its Applications. *Proc., 22nd ISATA*, Vol. 1, 1990, pp. 127–134.
14. S. Tsugawa et al. Vehicle Following System Using Vehicle-to-Vehicle Communication—Its Concept, Control Algorithm, and Communication System. *Proc., ASME/ISCIE USA-Japan Symposium on Flexible Automation*, 1988, pp. 621–628.
15. T. Yatabe et al. Control of Autonomous Vehicles with Vehicle-to-Vehicle Communication. Presented at IFAC/IFIP/IFORS Conference on Control, Computer and Communication in Transportation, 1989.
16. S. Tsugawa et al. Velocity Control for Vehicle Following Through Vehicle-Vehicle Communication. *Proc., 22nd ISATA*, Vol. 1, 1990, pp. 343–350.
17. Association of Electronic Technology for Automobile Traffic and Driving. *Report of Studies on the Super Smart Vehicle System in FY 1990* (in Japanese). March 1991.
18. Association of Electronic Technology for Automobile Traffic and Driving. *Report of Studies on the Super Smart Vehicle System in FY 1991* (in Japanese). March 1991.
19. Japan Industrial Technology Association. *Research and Development of the Comprehensive Automobile Traffic Control System*. Agency of Industrial Science and Technology, Ministry of International Trade and Industry, Japan, 1979.

Publication of this paper sponsored by Task Force on Advanced Vehicle and Highway Technologies.

Intelligent Vehicle-Highway System Safety: A Demonstration Specification and Hazard Analysis

A. HITCHCOCK

A complete specification and fault-tree analysis on one concept for an automated freeway system are described. The example chosen includes a single automated lane on a freeway that also admits manually driven vehicles. Proper execution of all maneuvers is independently verified by infrastructure instruments. There is maximal intelligence within the infrastructure. It is planned to do this again with different starting assumptions. This is only the first example in a set of two or three. Hazards have been specified. A safety criterion has been chosen by way of example. A design has been specified. A fault-tree analysis is also described. This analysis attempts to verify that the design does satisfy all criteria. The example demonstrates that full specification is possible and that design errors (there were four) can be detected by fault-tree analysis. After further development, the technique of full specification and fault-tree analysis can become a basis for safety standards that will apply to both design methods and verification of conformity to safety criteria. The initial assumptions were few and broad. Along with the need to avoid violation of the hazard criteria, they determine a very small set of possible system design structures. There is more than one safe way to design an automated road, but there is not an abundance of options.

Hitchcock has indicated (*1*) that the safety of an automated freeway can be examined, at the conceptual stage, thus

1. The first stage is to identify the safety-critical subsystem (S-CS). This should be of modular design so that each module can communicate with others only by defined protocols. In particular, each module in the S-CS has a fully specified interface with modules not in the S-CS. Thus, malfunction of a non-safety-critical module cannot cause danger. The work of Varaiya and Shladover on system architecture is relevant here (*2*).
2. Next, the S-CS must be specified completely. This means that what happens next in any condition of the system whatsoever may be determined.
3. Hazards must be specified. Here, a catastrophe is a collision that may cause death or injury, that is, a high–delta-V collision, and hazard is a condition in which a failure can result in a catastrophe.
4. A safety criterion is now selected. Fault-tree and other forms of analysis should be used to verify that it is satisfied (*3*).

PATH, Institute of Transportation Studies, University of California–Berkeley, Richmond Field Station, 1301 South 46th Street, Richmond, Calif. 94804.

A demonstration specification and a fault-tree analysis have been completed and reported (*4,5*). The safety criterion was that, for a hazard to occur, two independent faults must occur. Because a hazard is the precursor of a high–delta-V crash, three independent faults are necessary before anyone can be hurt as a result of automation. The specification and analysis are too complex to be reproduced here in detail. They demonstrate that the process just described is sufficient to ensure safe design at the conceptual stage.

Ultimately, the logical patterns described must be quantified. Safety criteria can then be stated in terms of catastrophe rates. The reliability of the critical components must be known in order to estimate casualty rates. Alternatively, the critical components can be identified and the required reliability specified.

If there is agreement about the safety criterion and the hazards, the method of complete specification and fault-tree analysis becomes a basis for procedural standards for safe design and evaluation.

SPECIFICATION OF HAZARDS

The hazards used in the fault-tree analysis, which form a basis for design, are specified elsewhere (*1*). They assume there is platooning, which is a basis of this design. The four hazards are then the possible precursors of a high–delta-V collision involving a platoon or vehicle under automatic control. These hazards are

1. A platoon (or single controlled vehicle) is separated from one ahead of it, or from a massive stationary object in its path, by less than platoon spacing (to be defined).
2. A vehicle not under system control is an unmeasured and unknown distance in front of a platoon or single controlled vehicle.
3. A vehicle is released to manual control before the driver has given a positive indication of readiness.
4. A vehicle is released to manual control at less than manual spacing from the vehicle ahead of it, or at such a relative speed that a spacing less than manual spacing will be realized within, say, 2 sec.

Here, platoon spacing is defined as the safe spacing between platoons according to the criterion of Shladover (*6*). A preceding vehicle is halted violently (say, with a deceleration of 1.0 g), and the follower must brake to rest without collision.

Manual spacing is that spacing at which drivers feel comfortable and use in normal driving. In the system proposed, these quantities, which depend on the condition of the road, are set by the system controllers.

During the fault-tree analysis, it became apparent that hazard specifications erred. The last one (Item 4 in the list) omitted to say that a vehicle should not be released to manual control while the brakes are being applied. Otherwise the driver may not be able to regain control.

No formulation of the hazards can exclude all possibilities of high–delta-V collisions. There are parts of the road (transition lanes, or TLs) where vehicles are taken from manual to automatic control or released from automatic to manual control. Both manual and automatic vehicles are present here. Thus, automatically controlled vehicles cannot be protected from all errors by manual drivers. Sideswiping or cutting in on an automated vehicle are particular examples. Hitchcock has shown that in any design, a fence must protect vehicles on the automated lane from such accidents (*1*). There is an exception when accident debris is projected through a gate just in front of a platoon. Vehicles on the transition lane are, however, open to such accidents, just as they would have been without automation. The collisions that may result, if a vehicle in a platoon fails, take place at low delta-Vs (*6*). Such collisions do not have to be guarded against in the same way as do collisions between vehicles in different platoons. This is the basis on which platooned designs are accepted (*6*).

If a platoon is fully formed and there is a vehicle failure within it, the ensuing collisions are slight. The entire platoon may then come to rest if a vehicle cannot continue. Provided the wreckage does stop without hitting something else, no occupants will suffer large, injury-provoking decelerations. The fences referred to earlier also have this effect. However, if vehicles are joining or leaving the platoon, the collisions can be more serious. Just how serious has not been made clear. We do not know the frequency of the vehicle faults that cause such accidents. In the current design, some automatic inspection of vehicles entering the system is envisaged. Whether dangerous faults will then be detected is not known.

Within-platoon collisions are not examined here. If there is only a single automated lane, system design can do little to reduce the numbers of such collisions. The time that elapses while vehicles join or leave a platoon can be minimized. If there are multiple lanes, the more serious incidents that occur when a vehicle is joining or leaving a platoon can be reduced. The wish to do so must be overriding. The system is arranged so that a platoon (or a single vehicle) joins another platoon only at the rear, by cutting into position from another lane. Platoons divide in the same way. It will often happen that a vehicle in the center of a platoon wants to leave the automated lanes (ALs). The platoon then must reform. The back of the platoon must make two lane changes in quick succession—which is likely to be uncomfortable. The cure seems worse than the disease.

SYSTEM CONCEPT

The system considered here has a single AL on which vehicles run in platoons on a freeway that is also used by manually controlled vehicles. Such systems are discussed by Hitchcock (*1*), who shows that, if the hazards are to be respected, the physical configuration is necessarily that which is presented in Figure 1.

The AL is separated from the unconcerned lanes (ULs) by a fence through which vehicles may pass only at brief on-gates and off-gates. The gates are grouped into one logical on-ramp (LONR) and one logical off-ramp (LOFR) per block (of 1 mi or so in length). The last gate in the block is an on-gate. To the right of the AL, there is a TL in some places. It stretches from some distance upstream of the LOFR and LONR to a short distance downstream of the LONR. It may be discontinuous, as it is in the case considered here.

The TL is instrumented with vehicle position detectors (VPDs). There are also VPDs on the AL in the neighborhood of the LONR and LOFR. Vehicles are taken under the control of the system on the TL and are under control as they are passed through the gates. Vehicles enter the AL only at the rear of a platoon. Along the length of both TL and AL runs a lateral guidance reference. This reference defines the proper course to vehicles' later controls. Close to the gates there are turning points marked on the TL. Turning points act as reference points in lane changing. Each vehicle bears a lateral and a longitudinal control system that keeps it on track and property spaced within a platoon. These systems contain sensors that can detect a vehicle ahead within a defined minimum range called sensor range. The speed at which platoon spacing equals sensor range is called sensor range speed.

All these conditions have been shown to be necessary for safety (*1*). In this design, the TL also contains identifiers and chicanes. At identifiers, vehicles that wish to enter the system identify themselves. At chicanes, their claim to have an operative control system is verified. Appropriate control signals are sent. It is checked that the vehicle accelerates, decelerates, or steers to the left or the right. The chicanes are, of course, much less severe than those on racing circuits—the occupants may not notice the test.

Varaiya and Shladover have described a possible control architecture of an automated freeway—it is shown, with the addition of a top layer (law), in Figure 2 (*2*). In this case, the link layer is concerned to organize the formation of platoons

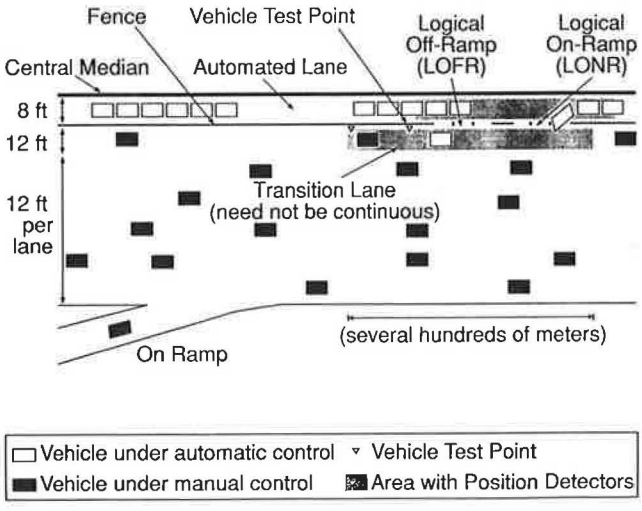

FIGURE 1 Layout of AL and TL.

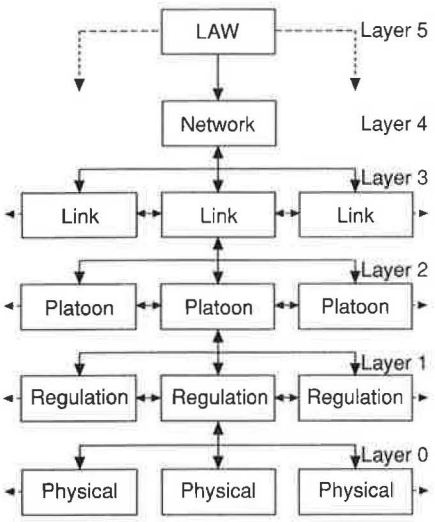

FIGURE 2 IVHS control architecture (2).

and selection of gates for entry and exit. It also organizes the movement of platoons on AL and TL so that there are appropriate gaps when lane change is desired. The link layer thus determines system capacity. The link layer is not part of the S-CS, and the maneuvers it recommends are not started until the platoon level has verified that they can be executed safely. The verification uses information from the VPDs. Control and interpretation of VPD signals are carried out at the regulatory level. The platoon and regulatory and physical layers compose the S-CS and are localized within each block.

When some fault occurs, the system in one block is put into a degraded mode. The fault may arise because the presence of a faulty vehicle in the system is detected. Alternatively, the roadside system detects an internal error. The law requires that vehicles with faulty (or no) controls do not enter the system. Entrance tests try to enforce this. However, faults may develop after entry. Deceit ("hacking") is also possible. In the simplest degraded modes, speeds on the AL are reduced to sensor-range speed. In one of these modes, vehicles may be required to exit the AL at the off-gates. In another, vehicles may continue on the AL the next block, where there may be no restrictions. If operation in these modes is not safe, further degradation is necessary. All vehicles will be brought to rest. In these conditions human intervention is necessary. The system controllers should be able to direct unusual operations. These include automated backing up on the AL so as to clear the way for removal of casualties or debris. Reversion to normal is also under human supervision.

INITIAL DESIGN CONSIDERATIONS

No previous descriptions of an automated freeway system have covered such events as joining and leaving the automated lanes, nor do any cater for fault conditions. The purpose here is to design a system on which a method of safety analysis is to be tested. The hazards are therefore defined first. In addition, the designer has three concepts that shape much of the design. These are

- The system should verify that each event demanded has in fact occurred. If it has not, suitable action should be defined. This ensures the completeness of the specification.
- There will be a strong temptation for some people to attempt to beat the system. Such people might send signals that falsely allege that a vehicle has automatic controls. They might suppress data indicating a fault. This can bring economic benefit to—and gratify the ego of—the driver. If this is done, however, it is possible that things will go wrong. This can be a source of catastrophe.
- Misinterpretation or nonreceipt of transmitted messages is a potential source of hazard. Noise can cause this, so it must be kept to a minimum. A strict discipline of sequenced transmissions is desirable. Each should identify the vehicle referred to, that is, which one is transmitting or being addressed.

The second of these considerations suggests that the output of vehicle-borne intelligence is untrustworthy. The intelligence should be emphasized in the infrastructure. This has therefore been decided on for the present design. Thus, much of the transmission between units will involve the infrastructure, and there will probably be economic advantage in eliminating any vehicle-vehicle communication. The third consideration supports this decision.

These decisions are basic to this design. They reflect unproven assumptions. It is certainly possible to construct system designs based on vehicle-borne intelligence—the author has done so. It may be possible to avoid the deleterious effects on hacking and noise by other means, or perhaps hacking and noise are less significant than assumed. Nevertheless, in this case, infrastructure-mounted intelligence is emphasized.

It is possible to arrange that a message to or from a vehicle is possible only if the vehicle is at one particular spot. This is necessary at the identifiers, for there is no other way in which the system can identify the vehicle seeking entry. It may also be desirable at chicanes. In general, however, the message identifies the vehicle being referred to. To ensure receipt it should be heard over the whole length of the block. Transmission and reception are therefore accomplished by way of a line of devices operating in parallel.

Functions of Architectural Levels

The regulatory layer must maintain each vehicle on track and in motion. By control of VPD signals, it provides performance data that enable an independent check on vehicle behavior. The platoon layer must initiate maneuvers, such as platoon formation or lane change, advised by the link layer. It does this only after verifying, in the light of data from the regulatory layer, that it is safe to do so. The platoon layer also analyzes the performance and position data provided by the regulatory level. It checks for a hazardous condition. If there is an incipient hazard, the platoon layer will revise maximum speeds. The platoon layer also passes data to the link level describing the position and speed of platoons and solo vehicles.

The link layer is outside the S-CS. It advises in maneuvers, including platoon and vehicle speed changes, that will

- Enable vehicles to leave at their selected exit,
- Form platoons as appropriate,
- Organize the pattern of traffic near gates in order to make gaps, and
- Enable changes from TL to AL and the reverse.

The link level thus has an optimizing function. Link maximizes the achieved capacity of the system so as to meet demand. However, it only advises on actions. The platoon levels mediate these for safety—that is, the local controllers at the platoon level check that each maneuver requested by the district controllers at the link level violates no hazards. Only then does the platoon level initiate the maneuver. This is achieved in part by sending a maximum speed to the vehicle-borne regulatory level. The regulatory level determines the speed of the vehicle it controls as follows:

1. The platoon level's maximum speed will not be exceeded, whatever the other rules say.
2. The speed of a vehicle in platoon is determined by the control algorithm, sensor readings, and other data.
3. The speed of a platoon leader or solo vehicle is the target speed, set by the link layer.

Alternative Designs

A very considerable superstructure has been built on initial considerations that are not without merit, but should not be overriding. The resulting design is not unsound, but the concentration on infrastructure intelligence and communication is extreme. Other designs are possible and would be arrived at if there were different initial considerations (for instance, that vehicle owners should pay directly for their benefits). It is therefore planned to make a parallel demonstration of the specification and fault-tree analysis technique, starting from the partial design of Hsu et al. (7). This is based on the idea that the greatest possible amount of intelligence should be vehicle-borne.

Development of Design

Returning to the initial design, the considerations rehearsed so far lead to the concept of the iterator as the roadside component of the regulatory layer. An iterator is a control and communication computer controlling a number of similar elements. An iterator communicates with each in element in rotation. Iterators on ALs and TLs address each solo vehicle or platoon member, thus stimulating a reply, in strict rotation, so as to avoid message overlay. The vehicle receives information about its maximum speed along with other data as are for within-platoon control (speed of platoon leader). The vehicle responds with an account of its speed, distance to the vehicle ahead (if it is detectable), and lateral displacement. The fact of reply ensures that the communication equipment is in order. This information is passed to roadside state vectors (RSVs), one per vehicle in a block. RSVs are data records held in asynchronous data stores. These stores are accessible by the platoon level. Messages are sent to platoon level, if appropriate. Thus, the continuous independent check on vehicle control system behavior can be made.

The other part of the roadside regulatory system is the monitoring via the VPDs of all vehicles on the TL and on the instrumented part of the AL. Each VPD is monitored in sequence. Vehicles (including unconcerned vehicles) are tracked using the traces they leave on the line of VPDs. Vehicle position and measured speed are communicated to the RSV. Again, messages are sent to platoon level if there is an unusual feature, such as too little space ahead of a platoon. Thus, the rest of the information needed in order to monitor vehicle control systems is gathered. In addition, there is a special RSV for each gate. If it is set, passage through the gate is barred because a vehicle is present on the other side. These RSVs, too, are controlled as part of the logical process described earlier.

The design can now be built up, module by module. Initially the required normal behavior of vehicles, drivers, and system is defined. This behavior is largely determined by the fact that each vehicle must reach its destination. As each maneuver is proposed, the system must have a means of checking that it has been carried out. The design must contain an alternative safe procedure if the maneuver is not executed.

An example may make this clear. Figures 3 and 4 show part of the flow diagram starting with a solo vehicle that must resume manual control and complete its journey. To avoid hazard, the driver must indicate readiness to resume control. A message is sent. If the driver replies, control is passed, and the iterator ceases to communicate with the vehicle. All is well. But, according to the design method, the question is asked, "What if the driver does not reply?"

The TL is not continuous—the vehicle cannot remain in motion on it indefinitely. What should happen is not certain. A choice must be made by the designer. The present system tries to reinsert the vehicle into the AL. The decision point is the last on-gate. Here, there may be a gap for entry at the rear of the platoon the vehicle has just left. The VPD signal is checked to ensure safety before instructing the vehicle to

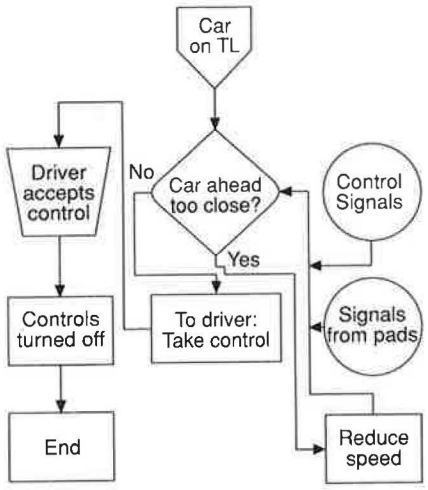

FIGURE 3 Normal exit.

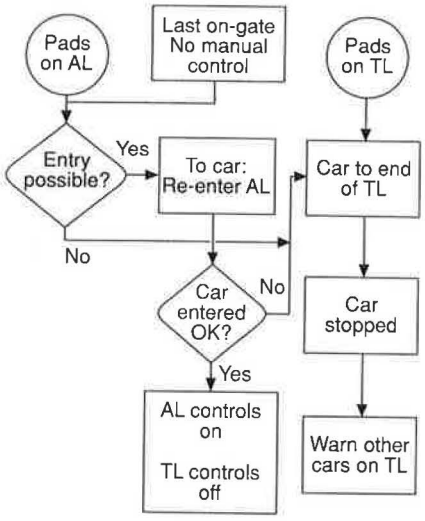

FIGURE 4 Driver does not take over.

enter. The VPDs are then checked again. (The case for which the vehicle enters is not followed further.) Entry may be unsafe because of vehicles on the AL. The gate may be closed because of operation in a degraded mode. The check may reveal that the vehicle did not enter when invited. Further checks are therefore needed to ensure that the vehicle comes to rest (or resumes manual control) before it reaches the end of the TL.

By proceeding in this manner throughout operation, the design has been completely specified. In this case, there was a choice to be made (what should be done with a vehicle whose driver fails to resume control?). A different system design would have resulted if some other choice had been made. This happens at a few other places in this system design.

DESIGN CHOICES

Each time a choice arises, an alternative choice will lead to an alternative design. The number of such choices is small, because in most cases it is quite clear what should be done if a maneuver is wrongly executed. Within the assumptions here, the only major choices are the preceding one and the following:

- Where should platoons be formed or dissipated? Platooning on the TL assists capacity, since more vehicles can pass a gate at once. However, on the TL the platoons are exposed to casualty-causing accidents involving unconcerned vehicles. These accidents would have happened anyway. The consequences are more serious because a platoon is involved. Large exit platoons may have difficulty in dissipating before the end of the TL is reached. Changing lanes in a platoon may place difficult demands on the vehicle-borne control systems and reduce reliability.
- In the design chosen, complexity is introduced by counting vehicles on the AL as they move from block to block. This may not be necessary: an intruder should be detected by the gate VPDs, and a lost vehicle should make its presence known long before it is missed at the end of a block. The redundancy introduced by counting may not, therefore, add to overall system safety.
- There are many possible ways of identifying the vehicle referred to in a message. In the design selected, the designer found it convenient to regard vehicles as passing through a series of modes as they passed tests at the chicane, joined platoons, entered the AL, and so on. This gave rise to a complex shifting ID system to which there are many alternatives.
- Further choices arise if some functions of the VPDs, such as checking on safe spacing from the platoon ahead, are transferred to the vehicle. This requires that a sensor can be constructed that has the ability to detect a vehicle in the same lane, distinguishing it from vehicles in other lanes.

FAULT-TREE ANALYSIS

In a fault-tree analysis, each hazard (a predecessor of catastrophe) is considered in turn. One asks, "How could this arise?" The answer will take on a form such as "If *A* happens, or *B* happens, or *C* happens, . . ." One then asks "How could *A* arise?" "If *AA* happens or *AB* happens, . . ." The process of identifying precursors continues. Mathematically, "*A* happens," "*B* happens," . . . are logical propositions, and "and," "or," and "not" are Boolean operators. Sooner or later one arrives at the point at which the proposition is one of the following:

- This can happen as a result of a single fault in a vehicle or other system component. In this case design error has been found.
- This implies that two simultaneous faults have occurred.
- There has been a computer or a communication error (the computers and communication equipment are assumed to be so redundant that this implies two simultaneous faults).
- The proposition is not possible (e.g., it involves reversal of gravity).
- The proposition implies that there has been an inadequacy in maintenance.

In each of the last four cases there is no breach of the safety criterion on this branch of the tree.

A fault tree clearly involves subjective elements. It is always possible that the investigator will fail to realize one of the ways in which a situation could arise. This becomes more likely when, as in this case, the investigator is the designer.

Nevertheless, in both specification and analysis, the process has been carried out with formal rigor. Besides its inclusion in some 20 pages of flowchart drawings, each module in the design (there are about 120) has been specified in a standard form. This form shares many features with the forms used for module specifications in formal-method computer languages such as Z or OBJ-3. The specification language used here, however, is not based on formal axioms. The complete formal specification is stated and discussed elsewhere (*4*). In the fault-tree analysis, similar rigor has been used—there are some 50 elements in the tree, and the arguments in each have been recorded precisely (*5*). Both reports are long and complicated, and no attempt is made to summarize them here.

RESULTS OF FAULT-TREE ANALYSIS

Four design faults were found:

1. On the uninstrumented part of the AL between gates, a following platoon may gain slightly on its predecessor. No mechanism is provided to correct this. In any one block the effect is trivial, but it could accumulate and cause a hazard.
2. Care is taken to check that a vehicle joining the AL does so only at the rear of a platoon or into a large gap. However, no check is made on the vehicle's speed when it enters the AL. If this is grossly mismatched with the platoon speeds, a hazard can arise.
3. If a vehicle develops a fault, it is detected and the driver is invited to resume manual control as soon as this can be made possible. No special precautions, however, are taken before it does so in order to keep other vehicles away from the danger a faulty vehicle presents. This can lead to hazards.
4. When a vehicle is released from a platoon or admitted to the TL on its way out of the system, its release is controlled so that its separation from the vehicle in front is safe. Controls also ensure that it is not moving much faster than its predecessor. Thereafter its distance from preceding vehicles is controlled to a safe spacing (though an unconcerned vehicle can always cut in). However, at the moment of release no check is made on its speed relative to its predecessor. This too can lead to hazards.

DISCUSSION OF RESULTS

The design described in this paper suffers from errors. They can readily be remedied should there be a serious intent to develop it. This is not very important, since there is no intent to develop this particular system. More important, at the system level considered here, it does appear to be possible to produce a design that avoids the hazards. This conclusion stands on the basis that controls and sensors can be instantiated that conform to what is specified.

The method of analysis chosen here is detailed, complete specification followed by fault-tree analysis. The example suggests that this approach is sufficient to ensure and verify conformity to safety criteria. The subsequent stage of a quantified hazard and risk analysis is, plausibly, also sufficient to ensure conformity to safety criteria of the whole system. There is more work to be done before these claims can be pronounced valid. That stage may be reached, however. These techniques could then become the basis for standards for design and evaluation procedures against stated safety criteria.

If regarded as an exemplar for standards, however, there are some serious flaws in the present demonstration. First, the designer and the analyst are the same individual. Proper management of system safety, as described in many guides in nonhighway fields (*8*), requires parallel and independent development of design and safety analysis.

Next, the design method should lead to a complete specification. No check has been made of this. In the work of Hsu et al., formal methods are used to demonstrate completeness for a part of the system (*7*). Whether these methods can be extended to the whole system, or even to the S-CS, is still to be investigated. But some independent validation of the completeness of the design concept is necessary before the present work can be regarded as exemplary.

It would be most satisfactory if verification and validation could be done by formal methods, so that completeness would be proven mathematically. If the fault tree also could be proved to be complete, it would be even better. However, this does not yet seem practical.

The choice of hazards constrains the number of possible system designs of an automated freeway. The constraint seems to be severe and the number, to be small. This parallels the earlier result (*1*), which was restricted to the physical layout.

ACKNOWLEDGMENTS

This work was performed as part of the PATH Program of the University of California, in cooperation with the state of California, Business, Transportation and Housing Agency, Department of Transportation, and the U.S. Department of Transportation, FHWA, and NHTSA. The author would also like to thank S. E. Shladover and P. Varaiya for helpful discussions.

REFERENCES

1. A. Hitchcock. Intelligent Vehicle-Highway System Safety: Problems of Specification and Hazard Analysis. In *Transportation Research Record 1318*, TRB, National Research Council, Washington, D.C., 1991.
2. P. Varaiya and S. E. Shladover. *A Sketch of an IVHS System Architecture*. Research Report UCB-ITS-PRR-91-3. PATH, University of California, Berkeley, 1991.
3. N. H. Roberts. Fault Tree Handbook. NUREG-0492. U.S. Nuclear Regulatory Commission, Springfield, Va., 1981.
4. A. Hitchcock. *A Specification of an Automated Freeway*. Research Report. PATH, University of California, Berkeley (in preparation).
5. A. Hitchcock. *Fault Tree Analysis of an Automated Freeway*. Research Report. PATH, University of California, Berkeley (in preparation).
6. S. E. Shladover. *Operation of Automated Guideway Transit Vehicles in Dynamically Reconfigured Platoons*. UMTA-MA-06-0085-79-1, 2 & 3. U.S. Department of Transportation, University of California, 1979.
7. A. Hsu, F. Eskafi, S. Sachs, and P. Varaiya. *The Design of Platoon Maneuver Protocols for IVHS*. Research Report UCB-ITS-PRR-91-6. PATH, University of California, Berkeley, 1991.
8. *Procedures for Safety-Critical Software*. DO 178A. Radiotechnical Commission for Aeronautics. Washington, D.C., 1986.

Publication of this paper sponsored by Task Force on Advanced Vehicle and Highway Technologies.

Abridgment

California INRAD Project: Demonstration of Low-Power Inductive Loop Radio Technology for Use in Traffic Operations

STEPHEN L. M. HOCKADAY, ALYPIOS E. CHATZIIOANOU, SAMUEL S. TAFF, AND WALT A. WINTER

Interest is growing at both the national and international level in using new technologies to alleviate congestion. The INRAD project examines one set of small-scale technologies working toward an incremental solution to improving the operation of the existing road network. The system uses loop detectors, originally installed in roadway pavement to perform standard vehicle detection, as antennas to exchange information between specially equipped vehicles and a traffic operations center (TOC). Two-way communication is performed using short message exchanges that can be updated from the TOC and the vehicle in real time.

In light of an anticipated increase in current highway congestion levels, new technologies are being sought for use in transportation to significantly improve the efficiency of existing and future highway facilities. The INRAD (Inductive Loop Radio) project at California Polytechnic State University (Cal Poly), San Luis Obispo, investigates the feasibility of incorporating existing sensor, computer, and communication technologies with the present highway system to provide two-way communication between the roadway and vehicles. INRAD has been developed in a manner that requires minor infrastructure changes and creates little disruption to the system.

Two-way vehicle-to-roadway communication is necessary in all intelligent vehicle-highway system IVHS concepts, including advanced traveler information systems (ATISs), advanced traffic management systems (ATMSs), and automatic vehicle control systems (AVCSs). Alternative technologies are being researched to provide this two-way communication. Network design standards must be developed before individual roadway-to-vehicle communication technologies can be effectively evaluated. Design standards will include specifications on the speed and volume of data transmission, one- or two-way communication requirements, and desired functions of the communication system. An effective way to develop such standards is through demonstration projects that reveal some of the networking and logistics intricacies and the potential of the technology being proposed.

S. L. M. Hockaday, A. E. Chatziioanou, S. S. Taff, Civil and Environmental Engineering Department, California Polytechnic State University, San Luis Obispo, Calif. 93407. W. A. Winter, Division of New Technology Materials and Research, California Department of Transportation, 5900 Folsom Boulevard, Sacramento, Calif. 95819.

In the INRAD project, messages are transferred when an equipped car passes over an inductive loop. The nature of the transmission technology, low-frequency radio waves over short inductive loops (used as antennas), will allow limited information exchange. The system design is not complicated by massive data transfers to and from the traffic operations center (TOC), and little additional load is placed on the network already operating for traffic detection purposes. The negative aspect, however, is that limited space on the channel of communication between the vehicles and the roadway constrains the kind of information exchanges that can occur.

Driver information and navigation systems have been developed in recent years and are becoming commercially available. However, the INRAD project is one of the few systems attempting to provide real-time, two-way communication between vehicles and a centralized control center. The Road/Automobile Communication System (RACS) (*1*), which is being developed in Japan, uses microwave technology to give drivers location and real-time traffic information. This is done through an infrastructure of roadside beacons, on-board devices, and a central facility. Advantages of this system include high broadcast speed, reasonable cost, and possible two-way communication.

ALI-SCOUT, a European project, uses infrared transmitters and receivers to link on-board displays with roadside beacons that are linked to a central control facility. This system was designed to be capable of two-way communication; however, it is not apparent that drivers can select and send information of their choice to the control center. Infrared communication provides very high transmission rates (between 0.5 and 1 M/sec). This system also requires optical connectivity for transmitters mounted on traffic lights and overpasses to send route guidance information to properly equipped vehicles (*2*,*3*).

DEMONSTRATION AND RESULTS

In March 1992, the first inductive radio demonstration was held in Los Angeles, California. The demonstration was designed to evaluate the effectiveness of two-way short-range radio communications between vehicles and the roadway. Participants of the demonstration were able to view the op-

eration of the TOC as it communicated with vehicles traveling on the freeway. They were also allowed to road test an INRAD-equipped vehicle to experience the process of receiving and sending messages via the in-vehicle display. Information exchanges were recorded on disk in order to evaluate the feasibility of the applications suggested in the last part of the paper.

The two problem areas that surfaced as needing more development efforts are information management and TOC design. More sophisticated integration of the different components is needed, and the central control computer interface needs to be made more flexible. Overall, the project successfully demonstrated the potential of inductive radio technology in supplying two-way vehicle-to-road communication.

SYSTEM COMPONENTS

The INRAD project included designing and developing electronic hardware and software to support radio communications and highway operations. INRAD software and hardware are divided into four components (Figure 1) that work together communicating through radio transmissions or telephone lines.

In-Vehicle Components

The hardware in test vehicles includes an on-board computer, a liquid crystal display (LCD), and an antenna. The function

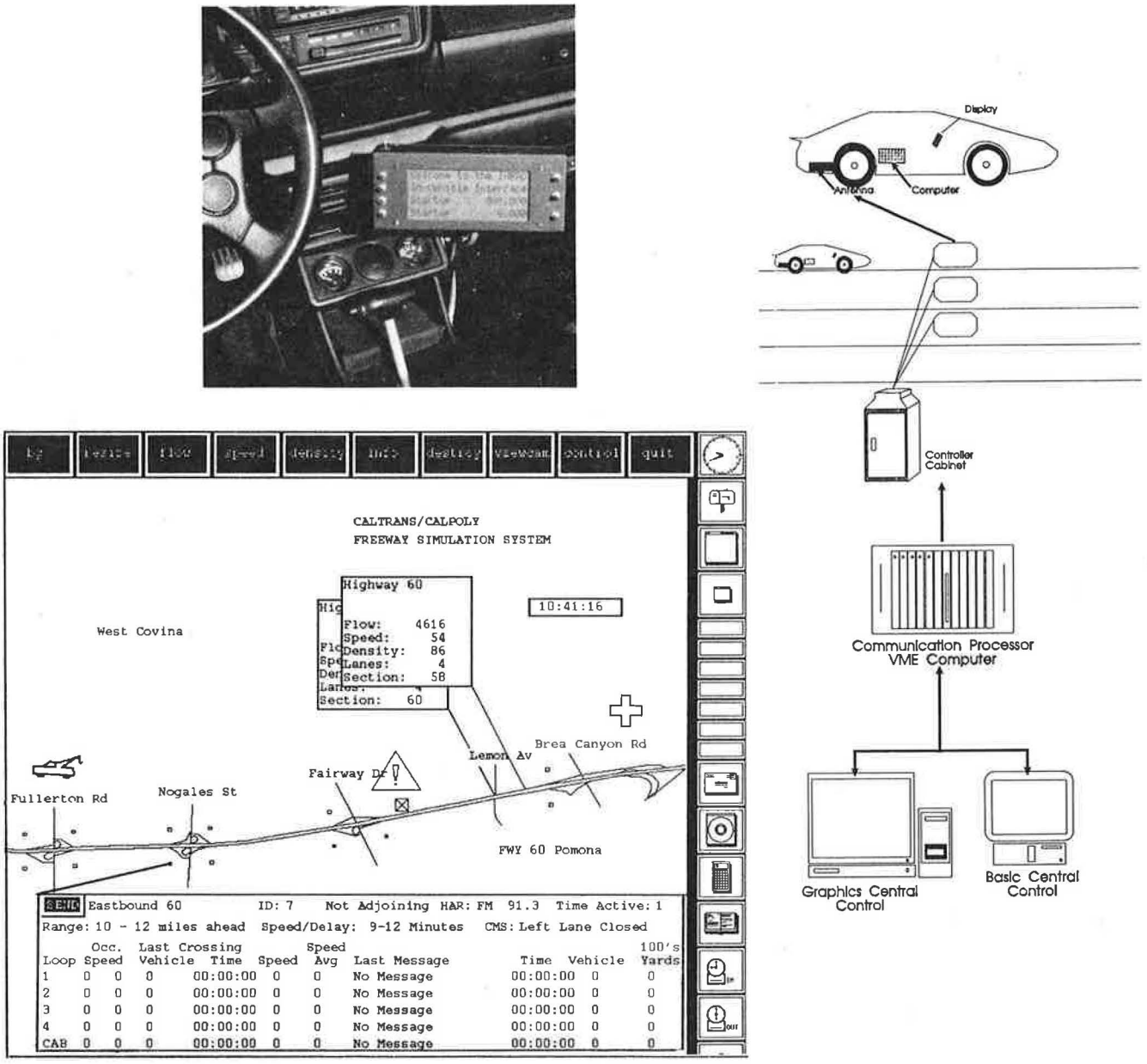

FIGURE 1 Components of INRAD.

of the computer is to post, on the display, messages received from the TOC and to enter messages sent to the TOC. The computer is connected to a radio transceiver that is attached to the rear bumper of the vehicle. The limited-range radio transceiver is able to communicate simple messages to and from an inductive loop as the vehicle crosses the loop. The messages are sent in the form of 6-byte digital code at a frequency of 375 kHz. Serial data are transferred at rates up to 9,600 baud (bit/sec). The system will allow one code to be transmitted in each direction at each loop crossing.

In-Cabinet Components

The roadside cabinets are equipped with a computer using an industry-standard STD bus. This computer controls the transmission of radio messages between vehicles and the communications processor (CP). The STD system uses a 2,400-baud phone line to communicate with the CP located at a central location.

Communications Processor

The CP is a VME bus computer that runs the OS-9 operating system. This computer manages all communication between the controllers and the central computers and distributes all the messages to their correct station. The CP also records all communication between the central control computers and the STD computers on disk or tape.

Central Control Computers

The central control (CC) computers allow the TOC operator to communicate with INRAD-equipped vehicles and monitor traffic data. Two different versions of CC computers have been developed. The first is the basic control computer (BCC) and is a PC system running the MS-DOS operating system. This version provides an inexpensive and simple screen-based user interface. The BCC had four display screens, which show all incoming and outgoing messages and the current average speed and occupancy of each section of highway. Another function of the BCC is to automatically analyze traffic data in order to issue speed and congestion warnings.

The second version of CC computers is the graphic control computer (GCC). This is a Sun workstation running the Unix operating system under the X Windows environment, which provides a sophisticated, user-friendly graphical interface operated with the keyboard and mouse device. The GCC displays INRAD loop and vehicle icons on a pictorial map of the freeway. All the information in the BCC can also be accessed by the GCC. The two CC computers can communicate with the CP via direct cables or, for longer distances, 2,400-baud telephone lines.

INRAD APPLICATIONS

The INRAD system has been designed to maximize the hardware and software capability and flexibility to send and receive pertinent real-time information to and from INRAD-equipped vehicles and the TOC or other central control facility. The user groups benefitted by the system would include commuters, highway maintenance crews, emergency services, transit and taxi services, commercial freight services, and private delivery firms. Possible uses are described in the following section.

Commercial Fleets

Commercial fleet use is an example of INRAD's automatic vehicle identification and location (AVI/AVL) capabilities. INRAD-equipped buses would automatically send identification and location information to central control each time they pass over an inductive loop. This allows central control to track the bus graphically, providing accurate, real-time monitoring of the bus fleet. A printout would also be produced at the central control site showing vehicle identification, location, date, time of day, and messages being transmitted or received. Included in the automatically generated information from the vehicle is space devoted to a message selected by the driver from menu items on the in-vehicle display. The driver could inform central control of remaining capacity, request emergency services, or receive new instructions from cental control.

The tracking and two-way communication provided by INRAD allow fleet controllers to make more demand responsive decisions. This type of technology can be tailored to benefit many various delivery services, local transit, and car rental companies.

Advisory System

Advisory system use would apply INRAD as a simple means of alerting drivers to various driving conditions ahead. Using INRAD as an advisory system provides real-time and very specific information on the immediate section of freeway being traveled. Advisory systems have proven to be effective in reducing accidents and congestion. The following alerts were successfully transmitted in the March 1992 INRAD demonstration.

HAR Alert

The most general alert is the highway advisory radio (HAR) alert. Drivers of INRAD-equipped vehicles could be alerted to the HAR alert by an audible "beep" when they pass over an inductive loop. The driver would then receive an alphanumeric message on the in-vehicle display; the message would advise the driver to tune the radio to a specific radio frequency on which is broadcast more detailed information about the section of freeway on which the driver is traveling or entering. The information may include details about freeway construction, fires, accidents, or other extraordinary conditions. The on-board interface uses the top two lines of the 20-character by four-line display panel. The lines read

DETAILED INFO ON HAR
TUNE TO *XX YYYYY*

where *XX* is either AM or FM and *YYYYY* is the frequency.

Speed-Decrease-Ahead Warning

Another advisory alert is the speed-decrease-ahead warning, which alerts drivers to a significant speed decrease in the next freeway section. The average speed at each set of loops is automatically determined and forwarded to the upstream loops by the BCC. The upstream loops then transmit this information to INRAD-equipped vehicles as they pass. A computer hookup to the speedometer in the vehicle records the vehicle's current speed and compares it to the speed of the next section. If the current speed is substantially higher than the average speed at the next set of loops, the driver is warned by a beep and a message. This gives the driver additional time to slow down and is particularly useful when adverse weather conditions impair vision.

The in-vehicle user interface uses the top two lines of the LCD panel. The message relays the following information:

> > > > > ALERT! < < < < <
XX MPH TRAFFIC AHEAD

where *XX* is the speed in the next section. This warning will be overridden only by a congestion-ahead warning.

Congestion-Ahead Warning

This message alerts drivers to delays on their freeway or an adjoining freeway up to 26 mi ahead of their present location. Information from standard loop detectors and current vehicle speeds received by INRAD-equipped vehicles can be analyzed to determine if significant delays are occurring. If so, the driver is alerted and shown the number of miles until congestion is encountered and the expected delay in minutes on the display. This accurate, real-time information about current conditions on the roadway enables drivers to make informed decisions affecting their travel route. The top three lines of the in-vehicle display shows the congestion-ahead warning. This message appears as follows:

AA–BB MINUTE DELAY
CC–DD MILES AHEAD
ON ADJOINING FREEWAY

where *AA–BB* shows the range of probable delay and *CC–DD* shows the range of distance before the congestion occurs.

Freeway Maintenance

The drivers of INRAD-equipped maintenance vehicles could send messages to central control concerning the road condition as they drive. Because the vehicle would be tracked by INRAD, the location where the notes apply would automatically be recorded along with the notes. INRAD can be used to calibrate and fine-tune other system status detection methods (speed calculation through occupancy, video image processing systems, etc.). Eventually, if enough vehicles are equipped with INRAD, it could be an excellent detection method itself.

Freeway Service Patrol

The use of roving service patrols to provide motorist aid is a proven strategy for reducing congestion in urban areas. The effectiveness of these service patrols can be greatly improved with the help of INRAD technology. The system would track service patrol vehicle location and allow for quicker and more efficient dispatching of the most appropriate vehicle to the emergency site. Information on system performance can also be obtained by monitoring the vehicle speeds and travel times of these roving patrol vehicles.

Dynamic Road Pricing

INRAD can be used to allocate user fees properly according to time and mileage on the system. Each vehicle would have its own identification and could be tracked by INRAD, updating a data base accordingly. INRAD could allow for dynamic congestion pricing to encourage drivers towards optimum operating conditions. A message would be posted on the in-vehicle display notifying the driver of a toll increase far enough in advance to allow proper decision making.

ACKNOWLEDGMENTS

The authors would like to acknowledge Les Kubel and Chris Dickey of the California Department of Transportation; Paul Lavallee, Scott Stewart, and Brian Delsey of the IBI Group; and Ron Nodder, Shirlee Cribb, Clinton Staley, Jim Heintz, Kimberley Mastako, Doug Modie, and Jeremy Gibson of California Polytechnic State University, San Luis Obispo, for their efforts on this project.

REFERENCES

1. K. Takada, T. Matsushita, and D. Fujita. Development of Road/Automobile Communication System (RACS). *Route Et Informatique*, March 1990.
2. R. K. Jurgen. Smart Cars and Highways Go Global. *IEEE Spectrum*, May 1991, pp 26–36.
3. R. von Tomkewitsch. Dynamic Route Guidance and Interactive Transport Management with ALI-SCOUT. *Vehicular Technology*, Vol. 40, No. 1, Feb. 1991.

INRAD research and development was performed under contract from the California Department of Transportation. This paper represents the views of the authors and should not be interpreted as either the policy or position of the California Department of Transportation. The authors bear sole responsibility for any errors or omissions. The paper does not constitute a standard, regulation, or specification. The mention of commercial products, their source, or their use is not to be construed as actual or implied endorsement of such products.

Publication of this paper sponsored by Committee on Communications.

Development of Prototype Knowledge-Based Expert System for Managing Congestion on Massachusetts Turnpike

ARTI GUPTA, VICTOR J. MASLANKA, AND GARY S. SPRING

A prototype knowledge-based expert system has been developed to assist in the management of nonrecurrent congestion. The system encompasses incident detection, verification, and response; it includes a real-time, dynamic network model for motorist diversion. A case-study simulation on the Massachusetts Turnpike illustrates the potential benefits to be derived from the system.

Urban traffic congestion is not a new problem. It antedates the motor vehicle and has been a continuing concern for much of this century. In recent years, several cities, including Phoenix, Atlanta, Houston, San Francisco, and Washington, have identified traffic congestion as their most serious regional problem (*1*). In fact, it is a serious national problem and continues to worsen. The reason is not hard to see: the number of cars owned in the United States increased by two-thirds between 1970 and 1987, and total annual vehicle miles traveled (VMT) increased from 1,120 billion to 1,910 billion during the same period (*2*), yet inflation-adjusted expenditures in 1987 were only 6 percent above 1970 levels. Since freeways account for only 3 percent of road mileage in urban areas but carry more than 30 percent of the total VMT (*3*), they are of particular interest in addressing the congestion problem.

Congestion may be classified as recurring or nonrecurring. Recurring congestion occurs when demand exceeds supply (usually during peak periods) on a regular basis. Common causes include lane drops, heavy volumes, poor geometrics, weaving sections, and so on. This type of congestion is often seen in cities during peak periods of travel when large numbers of work trips are being made. Nonrecurring congestion is characterized by unanticipated events such as accidents and disabled vehicles that cause a reduction in normal capacity. Given that these events, or incidents, are quasirandom in nature, they are difficult to predict and solutions to the problems they create are difficult to implement.

FHWA sponsored a study in 1986 to quantify the magnitude of the urban freeway congestion problem on a national scale (*4*). Estimates have been made for delay, excess fuel consumed, and user costs on the basis of assumed values for user time and wasted fuel, for both recurring and nonrecurring congestion. The results of the study show that there was a 30 percent increase in delay from 1984 to 1985 and predict an estimated fivefold increase by the year 2005 if no improvements are made, of which 70 percent will be due to nonrecurrent congestion. The congestion problem, as well as driver safety, could be greatly improved if these incidents were more efficiently managed.

It has long been acknowledged that we cannot "build" our way free from the congestion problem but must better manage existing facilities using transportation systems management (TSM) techniques. Freeway incident management has been used successfully for the past 30 years as a tool to reduce the impact of incidents in a number of urban areas throughout the United States. Programs of this kind have been in place in several states: California, Arizona, Washington, Illinois, Florida, Texas, and New York are examples.

FREEWAY INCIDENT MANAGEMENT

This paper explores the application of expert systems to freeway incident management (FIM) and proposes a methodology for developing an expert system to assist in incident management on the Massachusetts Turnpike. The generic incident management process has at its heart a traffic control center that monitors freeway operation for incidents. When an accident occurs, the controller responds appropriately—dispatching emergency vehicles and personnel, notifying appropriate agencies, alerting approaching motorists, and deciding whether to divert traffic and if so along which routes and for how long. From a traffic management viewpoint, freeway incidents should be removed as quickly and efficiently as possible. Additionally, freeway demand should be intercepted and diverted to other routes if the reduced roadway capacity during the incident is insufficient to satisfy demand and if practicable alternative routes exist. Success in achieving these goals results in increased freeway safety and decreased congestion and delay.

Incident management is a continuous process. The system implemented must be always available to detect and respond to incidents. Major components of the FIM process are detection, verification, response, and monitoring or feedback. Incident detection and verification require a freeway surveillance system. This system may be as sophisticated as the automatic detection systems (such as loop detectors and closed-circuit television) used in California (*5*) or as simple as police

A. Gupta, Department of Civil Engineering, University of Massachusetts, Marston Hall 235, Amherst, Mass. 01003. Current affiliation: JHK & Associates, Pasadena, Calif. 91101. V. J. Maslanka, DKS & Associates, Sacramento, Calif. 95814. G. S. Spring, Department of Civil Engineering, North Carolina A&T State University, Greensboro, N.C. 27411.

or traveler call-ins. The latter method is the default surveillance system on most roadways and consists of passing motorists, or police patrols, notifying local police agencies. When an incident is thought to have been detected, verification is necessary to screen out false alarms. This is usually accomplished through police patrol or closed-circuit television. Incident response encompasses a wide range of activities and is the basic requisite for a successful incident management system. Traditional response activities include the provision of medical services, fire agency response, hazardous material containment and cleanup, vehicle removal, and traffic control at the incident scene. Advanced response systems include motorist information systems, motorist diversion systems, and interconnections to other traffic control systems such as traffic signal systems along parallel routes and freeway ramp-metering systems.

RATIONALE AND SIGNIFICANCE

The large quantities of information flowing into the traffic control center, typically all at the same time, make responding quickly to freeway incidents extremely difficult. Decisions based on this copious, simultaneous information must be made quickly and accurately and must be disseminated just as quickly. This type of response requires traffic system managers who are well versed in handling emergencies and who are able to make decisions on the spot. Such managers have access to a tremendous amount of knowledge derived mainly from work experience in the area. Thus, the same problem faced in other application areas requiring special expertise must be faced here as well. Experts are rare and expensive, and it is often difficult to retain enough of them long enough to sustain effective operations. This means that valuable expertise is often available only sporadically and at significant cost to the user. It is for these reasons that expert systems offer such potential. Expert systems are computer programs designed to solve problems whose solutions require expertise. They attempt to use the knowledge of human experts to solve problems (*6*). Perhaps the most compelling reasons for using expert systems for FIM is their ability to use all available knowledge, consistently and without error or misjudgment—important considerations for real-time applications.

Within the context of incident management, the expert system is envisioned as a real-time, on-line computer system that will support the traffic system manager. The traffic system manager is traditionally a police agency representative responsible for incident management on a particular portion of the roadway network. This person is responsible for basic direction and coordination of all agencies involved in incident response. Without an expert system, the manager performs incident management duties, relying on knowledge, memory, past experience, and written guidelines. The computerized expert system supports the traffic system manager in the following ways:

1. The expert system can screen large volumes of data, alerting the manager only of data that appear to be abnormal. In this manner, the manager is protected from information overload and can devote his or her time to activities dealing with those data that suggest the existence of traffic system abnormalities. If properly programmed to screen traffic data, the expert system can detect abnormal fluctuations far more reliably than can a busy and fatigued human being.
2. The expert system provides consistency. The expert system will not vary in its response, as might different humans serving as traffic system managers, or even the same human under different working conditions. Of primary importance in this regard, the expert system will not forget important data or procedures.
3. The expert system provides an automated menu-driven procedure to guide the traffic system manager through the tasks of the job. Through interconnection with data bases and other computerized systems, the traffic system manager can work much more quickly and therefore effectively.
4. The expert system can be used as a training tool through its off-line use by both inexperienced and experienced managers through a wide range of hypothesized incidents.

It is important to note that the expert system is not intended to replace the traffic system manager. Each conclusion reached by the expert system can be accepted or rejected by the manager, as deemed appropriate. Additionally, the manager may review the logic that the expert system used to reach its conclusion. The nature of incidents that occur is so varied and unpredictable that the expert system may not be able to respond properly to unforeseen events.

As discussed previously, the response plan is the heart of incident management. Several automated FIM systems are currently under development. The FRED system places its main emphasis on the surveillance, verification, and decision support for on-site response strategies (*5*). Lakshminarayanan and Stephanedes focused almost exclusively on the on-site aspects of response (*7*). None of the current efforts, however, appears to have used network simulation techniques to assess the suitability of diversion plans.

The incident management process may be broken into two parts. One part requires judgment regarding appropriate response strategies based on incident severity level, such as dispatching emergency vehicles, notifying incident management teams, and determining response level and appropriate diversion routes from a large set of preplanned routes for the facility. This type of decision making (which in this case is based on type of incident, time of day, location—information used to determine expected durations, volume/capacity ratio, and so on) is a classic expert systems situation (*6*). The second component consists of essentially assigning vehicles to a network in real time and determining optimal paths from among the feasible routes recommended by the first part—clearly more amenable to a simulation program. Thus, two very different computer programming paradigms were used to reflect the special character of incident management problems: an expert system was built for the first, and a procedural data processing and network simulation program was built for the second.

The expert system developed herein uses a combination of an "Exsys" shell and a FORTRAN module. The Exsys shell is a layer of software developed using a generalized expert system package called EXSYS. This package can be run on any IBM PC, XT, AT, or compatible computer with 320K RAM. The expert system selects an appropriate response strategy and records basic information about an incident. Con-

trol is then passed to the FORTRAN module along with the basic information. This module contains the route-diversion algorithm. Each time the FORTRAN module identifies the optimum diversion strategy for a given incident condition, control is passsed back to Exsys, which displays the proposed strategy on screen for a dispatcher to review and implement if he or she agree with the recommended strategy.

CASE STUDY

The case study for this research is a section of the Massachusetts Turnpike. The turnpike is a 132-mi-long highway with 24 interchanges. It caters to both commuter and through traffic. It is well maintained and patrolled. A useful feature of the MassPike is its special emergency access points; highway authorities use them primarily for maintenance, snow removal, and other emergency purposes, and they are occasionally used for traffic diversion purposes also. The MassPike is a tolled facility. Thus, if traffic is diverted off the MassPike during an incident, there is a question of loss of revenue and the ability of toll plazas to handle extra traffic. Although we have identified these issues and recognize that they should eventually form an integral part of a route diversion strategy for a highway such as the MassPike, only toll lane capacities have been included in the present study. These issues need further research in collaboration with the MassPike and other highway authorities.

The section of the MassPike studied in detail is from Exit 9 to Exit 12, with incidents simulated in sections between Exits 10 and 11 and Exits 11 and 11A eastbound. These sections were chosen on the recommendation of MassPike staff, because there have been several serious incidents in the past in this area. For a map of the study section, refer to Figure 1. The study section passes through Sturbridge (Exit 9), Auburn (Exit 10), Millbury (Exit 11), Westborough-Hopkinton (Exit 11A), and Framingham (Exit 12). It intersects with I-84 at Exit 9 and connects to Route 20; with I-290, I-395, and Routes 12 and 20 at Exit 10; with Route 122 at Exit 11; with I-495 at Exist 11A; and with Route 9 at Exit 12. This section provides a variety of alternative routes, making the task of finding the best one challenging and interesting.

The volume and network data for the study network were obtained from the Massachusetts Turnpike Authority and the Massachusetts Department of Public Works. The turnpike authority also provided information on the incident management techniques currently used.

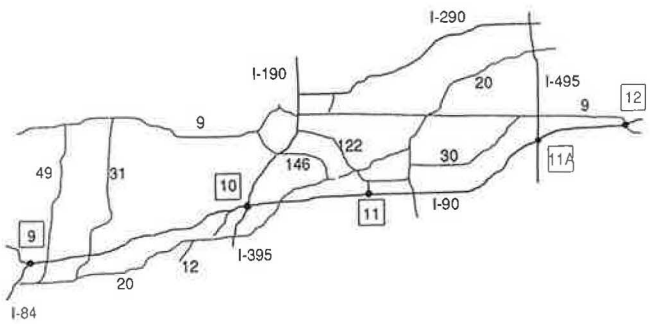

FIGURE 1 Massachusetts Turnpike, study section.

KNOWLEDGE-BASED INCIDENT MANAGEMENT SYSTEM

FIM is a multidisciplinary activity. Personnel from various agencies such as highway (traffic and maintenance), police, fire, environmental, and medical agencies need to work together in a coordinated manner. When an incident occurs, it must be decided who should be informed. In the expert system developed here, various incident situations and corresponding responses are included. It was designed as a comprehensive crisis management tool. For example, if there is a spill, depending on whether the spill is hazardous or not, special cleanup forces or regular cleanup forces should be informed. If the spill is a potential fire hazard, a fire crew should be put on alert or dispatched. If there are vehicles to be towed, a towing company should be called on to send trucks. Thus the expert system would prompt the operator for information about the incident and respond with the recommended action. The operator would receive the information from the police or another authorized person at the incident site. The operator would also act on the actions recommended by the expert system. The situations and responses have been formulated on the basis of what highway authorities do now or what they think should ideally be done, on the basis of their knowledge and experience.

Crisis Management

Expertise consists of knowledge about a domain, about how to use that knowledge, and about problem characteristics. Therefore, expert systems have three basic components: a knowledge base that contains heuristic knowledge (most often in the form of rules and facts) about the problem domain, an interpreter that contains reasoning methods (i.e., ways to process and use domain knowledge), and a data base that contains problem characteristics.

Knowledge Base

The process of coding the knowledge base consists of implementing a much more detailed version of the decision tree shown in Figures 2, 3, and 4 into a set of if-then-else combinations. The decision tree provides the conceptual framework of the problem into which details may be placed. Details, such as appropriate responses and incident severity level determinations, were taken from interviews with Massachusetts Turnpike Authority personnel. The system currently contains 36 if-then-else rules of thumb, some of which are necessarily site-specific. For simplicity, all site-specific information is stored in the body of the rules. However, the expert system shell used does provide a facility by which to separate site-specific information—thus allowing for system transferability. The rules for alternative route preplanning are based on flow and capacity constraints by time of day and location and the in-

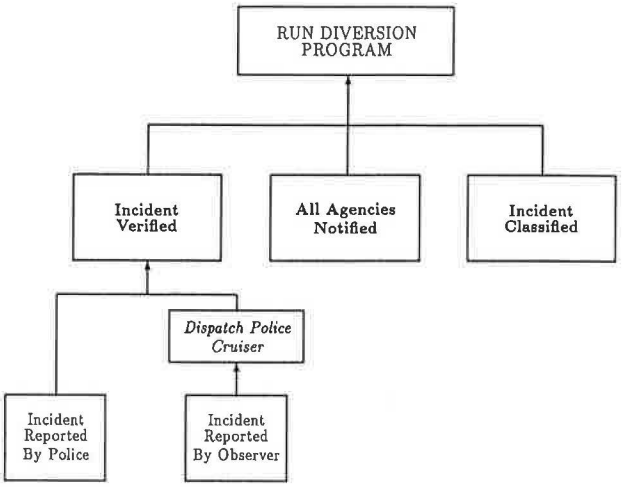

FIGURE 2 Components of knowledge base and incident verification phase.

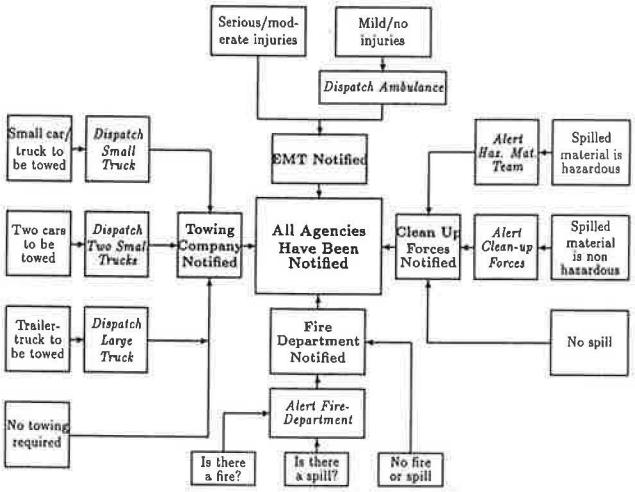

FIGURE 4 Incident notification phase.

cident's expected duration (in turn based on severity). The rules choose a set of alternative routes that are feasible, taking into account the dynamics of the parameters involved. The knowledge base has three primary parts:

- Incident detection and verfication—This requires a freeway surveillance system. The nature of this system may vary considerably. In the system developed here, the surveillance consists of notification by either passing motorists or police patrol. This method exists on many roadways. A more sophisticated system may rely on vehicle occupancy measurements, volume or speed data, or closed-circuit television. After an incident is detected, it needs to be verified to screen out false alarms. This is accomplished through police patrol or closed-circuit television, if the incident has been reported by a motorist or similar observer.
- Classification of the incident—This consists of information provided by police patrol on the incident characteristics, such as its time, location, and severity.
- Notification of the incident—This consists of notifying all agencies required to clear and manage the incident site, after the occurrence of the incident has been verified.

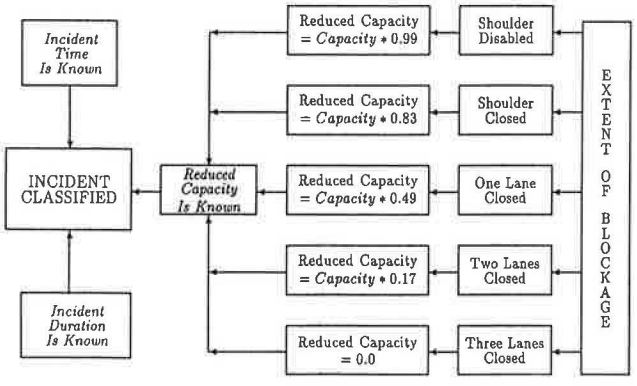

FIGURE 3 Incident classification phase.

After these three tasks have been accomplished, expert system control is passed to the diversion module with appropriate data to decide if diversion is required and to recommend diversion routes. Figure 2 depicts this process. The diversion module provides a diversion strategy.

Interpreter

The rule-based off-the-shelf interpreter EXSYS was chosen for this system because it provides a very simple programming environment while offering many useful utilities for the developer and the end user. The shell has all of the basic features necessary for effective implementation. It is relatively inexpensive and has a fairly friendly user interface. It provides a rudimentary explanation facility, allows what-if scenarios, interacts well with other external programs, and allows specification of uncertainty values. These qualities, along with the ability of the user to review the logic employed in arriving at a decision, make EXSYS an excellent prototyping tool.

Data Base

The system data base has several components: remotely sensed volume data, capacity data stored internally, and an active, constantly changing component that contains the current state of the problem—that is, which rules have been fired, which facts are true, and so on. Volume and capacity data for each link in the network are stored in ASCII data files containing temporal and spatial volume and capacity information. This is meant to simulate a real-time situation in which volume data are continuously fed into the computer system from field detectors. For each incident there exists a "best" diversion route to be followed depending on the location, severity level, and the time of day of the incident. Criteria used to determine incident severity level were expected incident duration and the volume/capacity (v/c) ratio of the affected freeway link.

At present, the freeway v/c ratio at any point is computed by the system using volume and free-flow capacity values

retrieved from the volume data file just described. However, the system is structured so that it supports traffic data acquisition hardware. It does not differentiate between data captured from a data file and data captured from, say, a loop detector. Similarly, system outputs could just as easily be displayed on roadside changeable message signs as on a computer screen.

Example

Consider an example to illustrate the operation of the expert system. We will first describe an incident and then go through the sequence of questions asked by the expert system, the responses selected by an operator, and the action recommended by the expert system.

Suppose a trailer truck containing combustible material overturns, colliding with a car near Mile Marker 95 in the eastbound direction on the MassPike at 9:00 a.m. Assume that the incident is reported by a driver on the scene. Let us further suppose that several people are hurt, some with serious injuries; that all MassPike eastbound lanes are closed; and that it is estimated that it will take approximately 2 hr to clear up the incident and restore the normal flow of traffic.

The sequence of questions and response is given in the following.

Incident Verification

- *Question*: Who is reporting the incident?
  1. Police
  2. Observer
- *Response*: 2 (observer)
- *Recommended action*: Dispatch police cruiser to verify the incident.

- Incident is verified.

Figure 2 depicts the steps in this phase.

Incident Classification

- *Question*: What is the approximate location of the incident (mile marker) and direction?
- *Operator*: 95, eastbound

- *Question*: What is the expected duration in minutes?
- *Operator*: 120

- *Question*: What is the extent of the blockage?
  1. Shoulder disabled
  2. Shoulder blocked
  3. One lane blocked
  4. Two lanes blocked
  5. Three lanes blocked
- *Operator*: 5 (three lanes blocked)

- Incident is classified.

Figure 3 shows the steps in this phase.

Incident Notification

- *Question*: What kind of injuries?
  1. Serious
  2. Moderate
  3. Mild
  4. None
- *Operator*: 1 (serious)
- *Recommended action*: Dispatch ambulance to Mile Marker 95 immediately.

- Emergency medical technician (EMT) is notified.

- *Question*: There is . . .
  1. Fire
  2. Spill
  3. Neither
- *Operator*: 2 (spill)

- *Question*: What is the classification of spilled material?
  1. Combustible
  2. Other hazardous
  3. Nonhazardous
- *Operator*: 1 and 2 (combustible and other hazardous)
- *Recommended action*: Alert fire department to stand by and inform hazardous material cleanup team.

- Fire department and cleanup forces are notified.

- *Question*: What types of vehicle are to be towed?
  1. Car
  2. Small truck
  3. Trailer truck
  4. None
- *Operator*: 1 and 3 (car and trailer truck)
- *Recommended action*: Dispatch one small and one large tow truck to Mile Marker 95 immediately.

- Towing company is notified.

This completes the notification phase, because all agencies have been notified (Figure 4). At this point, control and all incident information is passed to the FORTRAN module.

Traffic Diversion

When an incident reduces freeway capacity and causes congestion, the main concern of the transportation manager is to confine the problems due to the incident to that area itself, clear the incident as soon as possible, and return traffic flow to a normal condition. One of the main concerns is whether traffic should be diverted from the incident area. If it should be, How much traffic should be diverted? What alternative routes are appropriate? and When should the diversion end? The method for determining the appropriate alternative routes consists of two procedures: alternative route preplanning and real-time route diversion.

In the choice of alternative paths, several criteria can be applied. Travel time is one of the most widely used criteria, since it can be easily quantified and is of utmost concern to a motorist stuck in traffic. Here one needs to make a dis-

tinction between system- and user-optimal states. In the system-optimal state, travel time is minimized for the entire network; that is, the overall travel time for all the users is optimized. The term "user optimal" implies that each user tries to optimize travel time, which will not generally result in the system-optimal solution for transportation networks.

Alternative Route Preplanning

This step consists of developing preincident, detailed alternative route contingency plans for any location on the freeway system. As a first step, the freeway is divided into different sections and capacities are estimated. Then an inventory of freeways and all roadways that might serve as alternative routes for every section of freeway is completed. Information such as street widths, curvature, grades, pavement conditions, adjacent land uses, and weight restrictions is used to estimate the capacity of each link and to determine the suitability of each route. Special information, such as the presence of schools and special events, is taken into account in determining alternative routes. For example, during school opening and closing times, it may be desirable to avoid diverting traffic to school routes; during special events, such as sporting events, there may be no reserve capacity for diverted vehicles. This information can easily be stored in the knowledge base. Thus, we have a set of alternative routes for each section of the freeway with information about them for all hours of the day. Figure 1 depicts the MassPike study section divided into different links. There are in all 50 links and 38 alternative routes. The map shows 37 links; the rest of the links correspond to the MassPike links between the exists and the MassPike on- and off-ramps.

Real-Time Route Diversion

When a congestion-causing incident occurs, the expert system will act as an evaluator of the situation and help the system operator make decisions about traffic diversion. Real-time diversion implies assigning traffic to different routes by considering the prechosen alternative paths in a dynamic assignment modeling process. In this study, we have developed an algorithm to divert traffic from the MassPike by considering the conditions on the MassPike and on alternative routes and the subsequent effects of diversion strategy. Static and dynamic network models were developed to assign traffic to alternative paths during an incident. These network models for route diversion were tested extensively for incidents occurring at different times of the day, for various levels of lane closure, and for different incident durations. The purpose of the simulation was to test the output of the models for reasonableness and to increase insight into the problem. The test results show that MassPike exist ramps are the major bottlenecks when traffic is diverted off the MassPike. This can be attributed to the limited capacity of toll booths.

It was concluded that static models are not appropriate for modeling dynamic traffic events, such as incidents, because static models assume constant network characteristics over time. The use of static models to model incidents leads to impractical solutions, because of the inability of these models to include changes in volumes, demand, and capacities over time. In contrast, dynamic models can incorporate these changes and, therefore, provide better, more practical solutions.

The main input to the dynamic model is incident data and network data. Incident data include the time, duration, severity, and location of the incident; network data include volume, capacity, number of lanes, lane width, length, and free speed for each link in the entire network. The model operates on an IBM-compatible PC. The model output includes measures of effectiveness such as systemwide vehicle miles, vehicle hours, and queue size. The output also includes link flows, *v/c* ratios, and travel times at the end of each simulation period.

The optimization algorithm developed to select alternative routes is a heuristic procedure that has three objectives: minimization of travel time, reduction of congestion, and safety—that is, reduction of secondary incidents. Traffic is never diverted from the MassPike unless it reaches a certain level of congestion and extra capacity is available on alternative routes. The selection criterion for choosing alternative paths is a combination of *v/c* ratio and travel time. Briefly stated, the goal is to divert traffic to alternative routes only if all three of the following criteria are met:

1. MassPike will experience congestion if traffic is not diverted;
2. Uncongested alternative routes are available; and
3. Travel times on the alternative routes are less than the travel time on MassPike.

All alternative routes start from the MassPike exit (an exit or two upstream of the incident site) and return to the MassPike at an exit or two downstream of the incident site. During an incident, capacity and volume of the incident link are reduced. The volumes on the MassPike links downstream from the incident link are reduced by the difference between the MassPike demand volume and the reduced capacity of the incident link. Capacity on the incident link is reduced only for the anticipated duration of the incident. The model can also handle capacity changes at other times on other routes. The simulation is repeated once every minute until the queue at the incident time site becomes zero and the incident is cleared.

More than 100 simulations were conducted. Incidents were modeled between Exits 10 and 11A, with alternative routes extending from Exit 9 to Exit 12. Two- and three-lane closure incidents of various durations (30, 60, and 120 min) at different times of the day (7:00 a.m., noon, and 3:00 p.m.) were considered.

An analysis of the results obtained from the simulations shows that these results are reasonable and are consistent with the optimization objectives of the traffic diversion algorithm. For example, queue size increases as the duration of the incident increases, provided everything else remains constant. An incident of 30-min duration at 7:00 a.m. between Exit 10 and 11 would result in a queue of 222 vehicles if traffic were diverted from both Exit 9 and Exit 10. A similar incident of 60-min duration would lead to a queue of 455 vehicles. An incident of 120 vehicles would also lead to a queue of 455 vehicles, because the peak hour finishes at 8:00 a.m.; after

that, volume on the MassPike drops. As more vehicles are diverted off the MassPike, queue size decreases, but it takes longer for the system to return to normal. This insight into the system performance is very important if the system is to be extended for more than one incident in a given time period.

The use of "expert" algorithms in diverting traffic results in a reduction in congestion on the MassPike and systemwide savings in vehicle hours. Consider a two-lane closure incident of 30-min duration between Exits 10 and 11 at 7:00 a.m. If no diversion were exercised, a queue of 455 vehicles would be formed at the MassPike. On the other hand, if diversion were performed using the expert algorithm, queue length would reduce to 221 vehicles, and 8 vehicle-hr would be saved. A similar incident at noon would result in a queue of 111 vehicles. The diversion would reduce the queue length to zero and result in a savings of 20 vehicle-hr.

The algorithm never diverts traffic from more than one exit unless there is enough traffic at the incident link to suggest queue formation. For example, for a two-lane closure incident at noon, traffic is never diverted from Exit 9 because no queues are anticipated. Thus, the algorithm works consistently with the objective of not diverting traffic on longer paths, unless necessary.

CONCLUSION

Given the nature of the incident management problem, which involves many interacting agencies and the utilization of preselected incident response plans, the use of a knowledge-based expert system as a support tool for the traffic system manager is highly recommended. The application to the Massachusetts Turnpike test case provides support for the suitability of this approach. For the system to become operational, more testing and research are required. It would be useful to test the system on an actual section of the Massachusetts Turnpike rather than in simulation. This system could be used as a starting point for a turnpike-wide incident management system. The application of such an expert system would require a sophisticated communication system to disseminate information; the acquisition of such equipment should be considered. The dynamic network model for motorist diversion was tested successfully without requiring the input of information that is not readily available, such as origin-destination data. Using the discretization and incremental assignment of diverted motorists, the model can provide results in real time. However, as with any effective alternative route information or diversion system, effective implementation presumes real-time information about traffic conditions on the alternative routes. Therefore, an effort should also be made to collect such data.

The challenges in this area of research include modeling motorist response to information and considering more than one incident at a time. In the first area, it is important to realize that many of the modeling assumptions generally accepted by the transportation community as part of metropolitan transportation planning do not apply. We cannot assume that motorists will perform as we would like. The importance of motorist response has long been recognized as an issue that requires considerably more research before it can be incorporated as a meaningful parameter in any route diversion model. In the second area, the expert system approach appears promising because of its flexibility and adaptation to the problem at hand.

REFERENCES

1. M. J. Rothenberg. *Urban Congestion in the United States: What Does the Future Hold?* Publication IR-040. ITE, 1986.
2. D. K. Willis. Intelligent Vehicle/Highway Systems a Summary of Activities, Worldwide. *Proc., 1st International Conference on Applications in Transportation Engineering*, ASCE, San Diego, Calif., Feb. 1989.
3. *Selected Highway Statistics and Charts, 1986*. Report DOT-PL-88-003. FHWA, U.S. Department of Transportation, Oct. 1987.
4. J. A. Lindley. *Quantification of Urban Freeway Congestion and Analysis of Remedial Measures*. Report DOT-RD-87-052. FHWA, U.S. Department of Transportation, Oct. 1986.
5. S. G. Ritchie and N. A. Prosser. Real-Time Expert System Approach to Freeway Incident Management. In *Transporation Research Record 1320*, National Research Council, Washington, D.C., 1991.
6. F. Hayes-Roth, D. A. Waterman, and D. B. Lenat. *Building Expert Systems*. Addison-Wesley Publishing Company, Inc., Reading, Mass., 1983.
7. N. M. Lakshminarayanan and Y. J. Stephanedes. Expert System for Strategic Response to Freeway Incidents. *Proc., 1st International Conference on Applications in Transportation Engineering*, ASCE, San Diego, Calif., Feb. 1989.

Publication of this paper sponsored by Committee on Freeway Operations.

Artificial Intelligence–Based System Representation and Search Procedures for Transit Route Network Design

M. HADI BAAJ AND HANI S. MAHMASSANI

An artificial intelligence (AI)–based representation of transportation networks is described. Such representation facilitates the development of efficient AI search algorithms that make up the bulk of the computational effort involved in the design and analysis of transportation networks. The novel representation is demonstrated, as are its advantages as implemented in the AI search algorithms developed for the design and analysis of a particular type of transportation network, namely, transit bus routes networks.

The purpose of this paper is to describe an artificial intelligence (AI)–based representation of transportation networks. Such representation facilitates the development of efficient AI search algorithms that make up the bulk of the computational effort involved in the design and analysis of transportation networks. We demonstrate the novel representation and its advantages as implemented in the AI search algorithms developed for the design and analysis of a particular type of transportation network, namely, transit bus route networks. From an implementation perspective, using AI search techniques offers the advantage of representing the transit network design problem (TNDP) and carrying out a search efficiently using the "list" data structure of Lisp (List Programming), a so-called fifth-generation computer language (*1*).

TNDP

Several authors have studied the TNDP (*2*). In the TNDP, one seeks to determine a configuration, consisting of a set of transit routes and associated frequencies, that achieves some desired objective, subject to the constraints of the problem. Mathematical formulations of the TNDP have been concerned primarily with minimizing an overall cost measure, generally a combination of user costs and operator costs. The former is often captured by the total travel time incurred by users in the network, whereas a proxy for operator costs is the total number of buses required for a particular configuration. Feasibility constraints may include, but are not limited to, (a) minimum operating frequencies on all or selected routes (policy headways, where applicable), (b) a maximum load factor on any bus route, and (c) a maximum allowable bus fleet size.

M. H. Baaj, Department of Civil Engineering, Arizona State University, Tempe, Ariz. 85287. H. S. Mahmassani, Department of Civil Engineering, University of Texas, Austin, Tex. 78712.

Most existing formulations can be viewed as variants of the following mathematical program:

Minimize

$$\left\{c_1\left[\sum_{j=1}^{n}\sum_{i=1}^{n} d_{ij}t_{ij}\right] + c_2\left[\sum_{\text{all } k \in \text{SR}} f_k T_k\right]\right\} \tag{1}$$

Subject to

$$\text{Frequency feasibility: } f_k \geq f_{\min} \qquad \text{for all } k \in \text{SR} \tag{2}$$

$$\text{Load factor constraint: } \text{LF}_k = \frac{(Q_k)_{\max}}{f_k \text{CAP}} \leq \text{LF}_{\max}$$

$$\text{for all } k \in \text{SR} \tag{3}$$

$$\text{Fleet size constraint: } \sum_{\text{all } k \in \text{SR}} N_k = \left[\sum_{\text{all } k \in \text{SR}} f_k T_k\right]$$

$$\leq W \qquad \text{for all } k \in \text{SR} \tag{4}$$

where

d_{ij} = demand between nodes i and j;
t_{ij} = total travel time between i and j = $t_{\text{invtt},ij} + t_{\text{wt},ij} + t_{\text{tt},ij}$;
$t_{\text{invtt},ij}$ = in-vehicle travel time between nodes i and j;
$t_{\text{wt},ij}$ = waiting time incurred while traveling between nodes i and j;
$t_{\text{tt},ij}$ = transfer time incurred while traveling between nodes i and j;
N_k = number of buses operating on route k; $N_k = f_k T_k$;
f_k = frequency of buses operating on route k;
$f_{\min}$ = minimum frequency of buses operating on any route;
T_k = round trip time of route k;
W = fleet size available for operation on the route network;
LF_k = load factor of route k;
$(Q_k)_{\max}$ = maximum flow occurring on any link of route k;
CAP = seating capacity of buses operating on the network's routes;
SR = set of transit routes; and
c_1, c_2 = weights reflecting the relative importance of the two cost components.

Our proposed solution approach is hybrid in nature in that it provides a framework to incorporate the knowledge and expertise of transit network planners, efficient search techniques using AI tools, and some algorithmic procedures developed by others or adapted from related problems in vehicle routing. The three major components in the proposed approach are a route generation design algorithm (RGA) that generates different sets of routes corresponding to different trade-offs among the principal objectives; an analysis procedure (TRUST) that computes an array of network-, route-, and node-level descriptors as well as the frequencies of buses necessary on all routes to maintain load factors under a prespecified maximum; and a route improvement algorithm (RIA) that considers each set of routes and uses the result of the analysis procedure to generate an improved set of routes (*3*).

The RGA is a design algorithm that configures, for a given set of nodes connected by a road network and demand matrix, sets of routes that correspond to different trade-offs between the user and operator costs. It queries the user for the minimum percentage of the total demand that is to be satisfied directly (i.e., without transfers) and the percentage of the total demand that is to be satisfied with no more than two transfers. It searches the demand matrix for high-demand node pairs and selects them as seeds for the initial set of skeletons. These skeletons are expanded to routes by way of different node selection and insertion strategies.

The knowledge and expertise of transit planners is implemented in the different routines in the form of constraints on search and within the different node selection and insertion strategies. Different targets for the demand satisfaction and different insertion strategies result in different sets of routes with different user and operator costs. RGA relies on algorithmic procedures such as the *k*-shortest-paths algorithm (*4*) and on the selective application of the transit planners' knowledge and expertise to guide the search.

Once sets of routes are generated, an analysis procedure called TRUST (Transit Routes Analyst) is called to evaluate those alternative transit network route configurations. TRUST computes a variety of performance measures reflecting the quality of service and costs experienced by the users and the resources required by the operator. An essential feature of TRUST is the computation of service quality measures in terms of the fraction of trips with different number of transfers. Also important are the summary measures of transfer activity by route and by node.

After RGA has generated and TRUST has evaluated the sets of routes, the RIA is called to improve each of the generated sets. These modifications can be classified into two groups of actions: actions on the transit system coverage level, and actions on the route structure level. The first goal that RIA was designed to achieve is that of making the sets of routes generated by RGA economically and operationally feasible. RIA considers two modifications: discontinuing these low ridership routes and joining these routes or their nodes with other medium- to high-ridership routes. The second goal of RIA is to demonstrate and test existing improvement procedures. The modifications that RIA considers are route splitting and branch exchange heuristics (whereby branches of different routes are exchanged to form new routes so as to reduce transfers at the intersection nodes).

AI-BASED TRANSIT NETWORK REPRESENTATION

According to Rich and Knight (*5*), a good system representing knowledge in a particular domain should possess the following properties:

1. Representational adequacy: the ability to represent all kinds of knowledge that are needed in that domain.
2. Inferential adequacy: the ability to manipulate the representational structures in such a way as to derive new structures corresponding to new knowledge inferred from old.
3. Inferential efficiency: the ability to incorporate into the knowledge structure additional information that can be used to focus the attention of the inference mechanisms in the most promising directions.
4. Acquisitional efficiency: the ability to acquire new information easily. The simplest case involves direct insertion, by a person, of new knowledge into the data base.

Our Lisp-based representation of the TNDP offers advantages over conventional languages such as FORTRAN, C, or Pascal in terms of knowledge representation and search processing. We implement an object-oriented hierarchical structural representation: nodes are connected by links, which in turn are traversed by routes. The routes combine to form the paths by which a node pair's demand is assigned. The set of routes and their associated bus frequencies define the transit network object.

The transit network data representation lends itself conveniently to the "list" data structure representation of Lisp, which in turn supports the kind of path search strategies of interest in this application. This can be illustrated by the following:

- The network connectivity can be conveniently represented in a descriptive language such as Lisp: to each network node, one associates a set (or, in Lisp, a list) of neighboring nodes as well as the trip time (cost) associated with the nodes. Thus, the list {2[(1 11.4)(3 2.9)(6 8.0)]} indicates that one can travel from Node 2 to Node 1 in 11.4 min, to Node 3 in 2.9 min, and to Node 6 in 8 min. In addition, with each node object one associates properties whose values are useful to the design and analysis procedures. Such properties include the demand originating at a given node (and the percentage assigned by the network under design), the demand destined to a given node, and the number of trips transferring at a given node (all indicators of the node's relative importance).
- A route can be represented as a list of nodes, thus Route r25 is defined by the list of nodes (18 11 10 9 8 12 14). Associated with the route object are the list of flows on the routes links and the number of buses deployed to maintain the load factor below a minimum value that is prespecified by the user.
- The search techniques that are specific to the TNDP can be readily programmed in Lisp. In such techniques, a feasible path connecting two network nodes can be represented as a list. Thus, the list [(r1 9 16)(r8 16 21)] implies that one can travel from Node 9 to Node 21 by boarding Route r1 from Node 9 to Node 16 and Route r8 from Node 16 to Node 21 (i.e., Node 16 is a transfer node).

DESIGN PHILOSOPHY OF AI SEARCH ALGORITHMS

Figure 1 presents the principal computational trade-offs that should be considered in the design of AI search techniques. As Nilsson argued, the computational costs of any AI search algorithm can be separated into two major components: the rule application costs and the control costs (*6*). A completely uninformed control system is characterized by a small control strategy cost because arbitrary rule selection need not depend on costly computations. However, such a strategy results in high rule application costs because generally it must try a large number of rules to find a solution. To inform a control system completely about the problem typically involves a high-cost control strategy, in terms of the storage and computations required. However, such a completely informed control strategy results in minimal rule application costs, for they guide the search directly to a solution.

The efficiency of an AI search technique is thus directly tied to achieving a proper balance between both computational cost components. This relies on "informedness," or the amount of knowledge and information that the rule-selecting computations possess about the problem at hand. Optimum search efficiency is usually obtained from control strategies that are less than completely informed. Thus, all three major components of our AI-based solution approach constitute a testing ground for selective application of knowledge.

EFFICIENCY OF AI-BASED DESIGN AND ANALYSIS SEARCH PROCEDURES

The principal motivation for using Lisp (or, more generally, a fifth-generation language) lies in the nature of the computational activity taking place in our solution approach, which consists of searching and screening paths in a graph. It has been common wisdom in transportation network applications to avoid any form of path enumeration. Thus, most existing

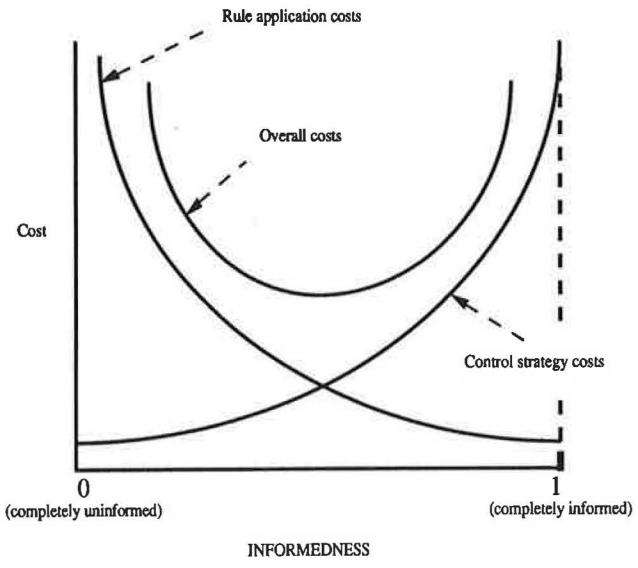

FIGURE 1 Computational costs of AI production system.

assignment procedures are limited to shortest-path constructs. However, other programming paradigms and advances in computing hardware and software can greatly facilitate some degree of path search and enumeration, which is justified by the added realism that it could allow into the resulting procedure.

Taylor describes simple programs written in Prolog, another fifth-generation language, to solve different route selection problems (*7*). His examples underscore the brevity of code as well as the relative ease of programming with fifth-generation languages. At the basis of these programs are some general "predicates" (Prolog meta-statements) that test for set membership, generate the intersection or union of any two lists as well as the complement of one list in another, sort a list of objects according to some numerical property, or append a new element to a set. Such meta-statements define the necessary condition for the required solution, thus relieving the program developer from worrying about the elemental computing and housekeeping chores, as would be the case with conventional programming languages such as FORTRAN, C, and Pascal.

An example of the use of such predicates in one AI algorithm is the way TRUST assigns a given node pair's demand to the transit network generated by RGA. For a given node pair (i,j), the list of routes passing through node i and the list of routes passing through node j (denoted by SR1 and SR2, respectively) are assembled by calling the "routes-passing-by" procedure for both nodes. If either of the lists are empty, then at least one of the two nodes (i or j) is not served by any transit route, and the demand d_{ij} cannot be assigned. If both lists are not empty, than a call is made to procedure "assign-0-transfer?," which checks whether the demand can be assigned directly (i.e., without transfers). This is possible only if the intersection list (of the two lists SR1 and SR2), which is the subset of all routes that have both nodes i and j on their node list, is not empty. When this is the case, the intersection list is passed on to the procedure "decide-0," whose function is to distribute the demand d_{ij} among the acceptable routes.

If the "assign-0-transfer?" procedure is unable to assign the demand between i and j directly (with no transfers), a call to procedure "assign-1-transfer?" is made. The latter checks whether the trip can be completed with one transfer. This check is carried out by examining, for every possible combination of a route member of List SR1 (say, R1) and another of List SR2 (say, R2), the intersection list of the list of nodes of R1 and R2. If the intersection list is not empty, then its contents are possible transfer nodes between R1 and R2. For example, if the intersection list is (tf1 tf2), TRUST forms two possible paths for the assignment of demand between nodes i and j: [(R1i tf1)(R2 tf1 j)] and [(R1 i tf2)(R2 tf2 j)]. The first indicates that a possible path from i to j consists of boarding Route R1's bus at i and staying on it until tf1 (Transfer Node 1) is reached. There the passenger should transfer to Route R2 and travel on it until the destination node j. For each possible path involving one transfer, an estimate of the total travel time is calculated. All paths between i and j whose total travel times exceed the minimum possible value by more than a specified threshold (say, 10 percent, as selected here) are rejected. Procedure "decide-1" subsequently distributes d_{ij} among the paths that have passed the filtering process.

Similarly, if it is determined that d_{ij} cannot be assigned with at most one transfer, procedure "assign-2-transfer?" is called to search for paths involving two transfers. If no such path is found, the demand between i and j will remain unsatisfied. In other words, it is assumed that a passenger will simply not consider boarding the transit buses to accomplish a trip that requires three or more transfers. Hence, TRUST avoids searching for paths that reach destination j with three or more transfers, thereby avoiding an otherwise considerable amount of meaningless search and keeping the execution time within tolerable limits.

The "assign-2-transfer?" procedure searches for all paths with exactly two transfers between given nodes i and j. The search consists of finding a route that passes through neither i nor j but that shares a node with a route passing through i (i.e., with a member of SR1) and another through node j (i.e., with a member of SR2). List SR3 of routes that pass through neither i nor j is obtained as the complement (in SR, the set of all routes) of the union list of the previously generated lists, SR1 and SR2. For a trip to require exactly two transfers between origin i and destination j, the first route, R1, must pass by node i (hence, R1 $\in$ SR1); the second route, R3, must pass by some node other than i or j (hence, R3 $\in$ SR3); and the third route, R2, must pass by node j (hence, R2 $\in$ SR2). Thus, three DO loops are executed: the outer one on SR3, the inner one on SR1, and the innermost one on SR2. A route from each of the above three sets is selected, and the following test is performed: if the "list-of-nodes" of the route from SR3 (say, R3) intersects both the "list-of-nodes" of the route from SR1 (say, R1), and the "list-of-nodes" of the route from SR2 (say, R2), then a possible two-transfer path is defined. For example, if the intersection of the "list-of-nodes" of R3 and R1 is (tf1 tf2) and that of the "list-of-nodes" of R3 and R2 is (tf3), then TRUST defines two possible paths: [(R1 i tf1)(R3 tf1 tf3)(R2 tf3 j)] and [(R1 i tf2)(R3 tf2 tf3)(R2 tf3 j)]. This is repeated until all possible combinations of triplets (R3 $\in$ SR3, R1 $\in$ SR1, R2 $\in$ SR2) are checked and listed.

On the negative side, Lisp, like most higher-level languages, may experience relatively slower computational performance when it comes to mathematical computations (as opposed to symbolic manipulations). However, in our particular application, the tests conducted and reported by Baaj indicate that the AI search algorithms perform satisfactorily within reasonable execution times (*3*).

CONCLUSIONS

In this paper we described an AI-based representation of transportation networks. Such representation facilitated the development of efficient AI search algorithms that composed the bulk of the computational effort involved in the design and analysis of transportation networks. AI search techniques offer the advantage of representing the TNDP and carrying out search efficiently using the list data structure of Lisp. Such representation was essential in the success of our AI-based hybrid solution approach proposed for the solution of the TNDP. Our hybrid approach consisted of (a) AI heuristics for transit route generation and improvement, (b) a transit network evaluation model, and (c) the systematic use of context-specific knowledge to guide the search techniques. Results of computational testing of our AI-based solution approach on a benchmark transit network and on data generated for the transit network of the city of Austin, Texas, were promising (*3*). Further testing remains to be done on different transit networks and their corresponding transit demand matrices. In addition, we seek to investigate the merits of applying an AI-based representation and solution approach in other transportation network design problems.

REFERENCES

1. P. H. Winston and B. K. P. Horn. *Lisp*, 3rd ed. Addison-Wesley Publishing Company, Inc., Reading, Mass., 1989.
2. M. H. Baaj and H. S. Mahmassani. TRUST: A Lisp Program for the Analysis of Transit Route Configurations. In *Transportation Research Record 1283*, TRB, National Research Council, Washington, D.C., 1990, pp. 125–135.
3. M. H. Baaj. *The Transit Network Design Problem: An AI-Based Approach*. Ph.D. thesis. University of Texas, Austin, 1990.
4. D. R. Shier. On Algorithms for Finding the K Shortest Paths in a Network. *Networks*, Vol. 9, 1979, pp. 195–214.
5. E. Rich and K. Knight. *Artificial Intelligence*. McGraw-Hill, Inc., New York, N.Y., 1991.
6. N. J. Nilsson. *Principles of Artificial Intelligence*. Tioga Publishing Company, Palo Alto, Calif., 1980.
7. M. P. A. Taylor. *Knowledge-Based Systems for Transport Network Analysis: A Fifth-Generation Perspective on Transport Network Problems*. Department of Civil Engineering, Monash University, Victoria, Australia, 1989.

Publication of this paper sponsored by Committee on Expert Systems.

Evaluation of Artificial Neural Network Applications in Transportation Engineering

ARDESHIR FAGHRI AND JIUYI HUA

The increased interest in artificial neural networks (ANNs) seen in government and private research as well as business and industry has included relatively little activity in transportation engineering. The position that ANNs, as a branch of artificial intelligence, hold in the transportation engineering field is discussed, including the differences between ANNs and biological neural networks and expert systems, respectively. The characteristics of ANNs in different fields are discussed and summarized, and their potential applications in transportation engineering are explored. A case study of trip generation forecasting using one traditional method and two ANN models is presented to show the application potential of ANNs in transportation engineering. The results of each method are compared and analyzed, and it is concluded that the potential for using ANNs to enhance both software and hardware in transportation engineering applications is high, even in comparison with expert systems and other types of artificial intelligence technique.

Artificial neural networks (ANNs) have proven to be an important development in a variety of problem solving areas. Increasing research activity in ANN applications has been accompanied by equally rapid growth in the commercial mainstream use of ANNs. However, there is relatively little research or practical application of ANNs taking place in the field of transportation engineering. This paper summarizes the characteristics of ANNs, evaluates the applicability of ANNs to transportation engineering, and explores the interface of ANN techniques and different transportation engineering problems.

DEFINITION OF ANNs

The human brain consists of 10 billion to 500 billion neurons. A cell body, an axon, and dendrites make up a biological neuron such as the one shown in Figure 1 (*top*). The connections between the neurons are called synapses, and each neuron is connected to 100 to 10,000 other neurons. A neuron executes a very simple task: when presented with a stimulus, it emits an output into other neurons connected to it via the synapses (*1*,*2*).

Artificial neurons (also called processing units or processing elements) mimic the functions of biological neurons by adding the inputs presented to them and computing the total value as an output with a transfer function. Figure 1 (*bottom*) shows a simple example of an artificial neuron. The artificial neuron also connects to other artificial neurons as the biological neuron does. The strength of the connections is called weight.

An ANN is a system composed of artificial neurons and artificial synapses that simulates the activities of the biological neural network. The ANNs can be single layer or multilayer, depending on their structure. In a single-layer ANN, all the processing units of the ANN take inputs from the outside of the network and their outputs go to the outside of the network; otherwise, it is a multilayer ANN. The principle that can approximately compute any reasonable function in an ANN is called architecture. The weights are adjustable, the programming for adjusting the weights is called training, and the training effect is called learning. The learning can be done either by being given weights computed from a set of training data or by automatically adjusting the weights according to some criterion.

A general definition for an ANN can be given as "a computing system made up of a number of simple, highly interconnected processing elements that process information by dynamic state response to external inputs" (*3*).

Differences Between ANNs and Biological Neural Networks

Although ANNs attempt to simulate real neural networks, they operate differently in many ways. The primary differences between ANNs and biological neural networks follow.

1. The processing speed is different. Cycle time is the time taken to process a single piece of information from input to output. The effective cycle time of a biological neuron is about 10 to 100 msec; the effective cycle time of an advanced computer's CPU is in the nanosecond range (*1*,*2*).
2. There are more than 100 kinds of biological neuron. ANNs contain only a few kinds of processing unit.
3. The "computations" (chemical reactions) in biological neual networks occur not only in neurons, but also in dendrites and synapses (*2*).
4. ANNs seldom have more than a few hundred processing units; a brain has 10 billion to 500 billion neurons.
5. The human brain has stronger error removal capability than an ANN. An upside-down letter may cause a lot of error in an ANN recognition system, but the human brain can recognize it easily.
6. Knowledge in ANNs is replaceable, but knowledge in biological neural networks is adaptable (*2*).

Department of Civil Engineering, University of Delaware, Newark, Del. 19716.

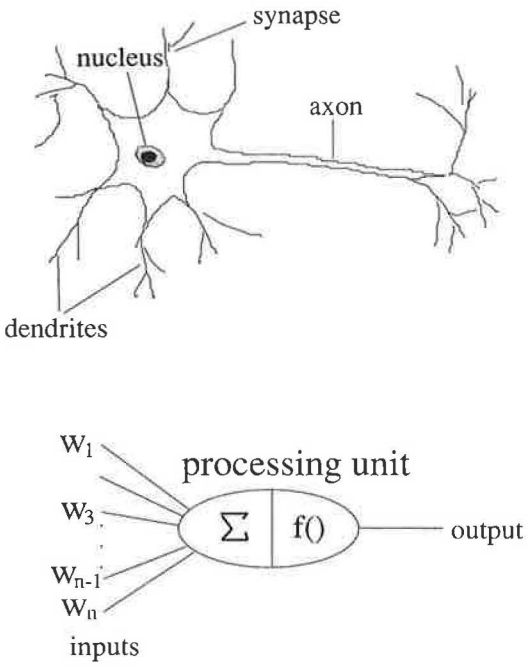

FIGURE 1 Schematic drawing of typical biological neuron and artificial neuron.

Differences Between ANNs and Expert Systems

As a part of artificial intelligence, ANNs possess some similarities with expert systems, such as storing knowledge and having learning processes in their operation. However, each has its own characteristics. The relationship between ANNs and expert systems is one not of replacement but of partnership. The differences between ANNs and expert systems follow.

1. ANNs and expert systems differ in the method of developing an intelligent system. The expert system approach uses a domain expert to identify the explicit heuristics used to solve problems, whereas the ANN approach assumes the problem-solving steps are to be derived without direct attention as to how a human actually performs the task. Thus, expert systems try to figure out how the human mind is working, and ANNs mimic the most primitive mechanisms of the brain and allow the external input and output to designate the proper functioning (*4*).
2. ANNs are flexible with knowledge. The knowledge stored in expert systems is restricted to the human knowledge domain. In contrast, ANNs can be trained with some data acquired in the past for the particular problems to be solved; the data may not even to be a type of knowledge. For some problems, it is not necessary for humans to understand the knowledge about the data; the ANNs give answers according to their own internal criteria.
3. ANNs have different ways of learning. The learning of expert systems is a procedure designed to store some knowledge in the system. The learning of ANNs is to adjust the strengths of the connections between the processing units. The knowledge in an ANN is more like a function than a content. Some ANNs could have a self-learning capability.
4. ANNs have different ways of computing. The results in expert systems depend on how the knowledge has been represented. The computation method used by an ANN is determined by its architecture, and the results depend on how the network is structured. In a multilayer ANN, if the number of hidden units is changed, the accuracy of the results will be different.
5. Once an ANN is developed, no more programming is required; the only requirement is to feed data to the ANN and train it. However, in expert systems, programming may be required if additional knowledge is to be introduced to the system.

ANNs and expert systems can cooperate with each other, and for some problems they overlap in use. Once integrated, it is expected that they will enable artificial intelligence (also called AI) to cover a much wider variety of applications.

Applications of ANNs

The special characteristics make ANNs especially useful in a variety of applications. An ANN can provide an approach that is closer to human perception and recognition than traditional methods. In situations in which input is noisy or incomplete, ANNs can still produce reasonable results.

Although ANNs have not been widely explored in transportation engineering, their unique properties indicate great potential. We summarize two applications in the following.

Self-Organizing Traffic Control System

Nakatsuji developed an optimizing splits traffic control system using a four-layer ANN (*5*). The development is based on two assumptions: (a) cycle length is common over the road network and does not vary with time, and (b) there are no offsets between adjacent intersections. The purpose of this project was to estimate optimal splits of signal phases using ANN technology. The inputs to the ANN are control variables, that is, split lengths of signal phases and the traffic volumes on inflow links; the outputs from the ANN are the measures of effectiveness such as queue lengths and the performance index. Through a case study, Nakatsuji reported that his system was effective in adjusting the synaptic weights in the training process and was able to improve the convergence into global minimum, and the solutions achieved were in accord with analytical ones.

Intelligent System for Automated Pavement Evaluation

The intelligent system for automated pavement evaluation was developed by Ritchie et al. (*6*). The focus in this research was the development of an advanced sensor-processing capability using ANN technology to determine the type, severity, and extent of distresses from digitized video image representations of the pavement surface acquired in real time. A three-layer ANN was used in this system. The results of the

initial case study presented in this paper clearly show the potential for application of ANNs for distress classification of pavement images as part of the proposed innovative noncontact intelligent system.

CLASSIFICATION OF ANNs

In the operation of ANNs, two issues should be addressed: (a) How are the processing units and the interconnection configured? that is, What is the structure of the ANN? and (b) How will weight values be assigned to the interconnections? that is, What are the ANN's learning rules? Here, the general classification of architectures and the learning of ANNs are presented.

The types of ANN architecture can be divided into four categories.

1. Mapping ANNs—Using a transfer function, mapping ANNs compute the sum of all the products of corresponding inputs and weights to make the outputs, that is,

$$Y = f(X,W)$$

where

Y = outputs,
X = inputs,
W = weights, and
f = transfer function.

2. Recurrent ANNs—In recurrent ANNs, some (or all) of the outputs are connected to the inputs. The outputs of the ANNs are the function of the inputs, the weights, and some (or all) outputs, that is,

$$Y = f(Y,X,W)$$

or

$$y(t + 1) = f[Y(t), X(t), W(t)]$$

where t denotes the time.

3. Temporal ANNs—Temporal ANNs compute the rates of the changes of their outputs as a function of the outputs, the inputs and the weights. In mathematical notation,

$$dY/dt = f(Y,X,W)$$

4. Hybrid ANNs—Hybrid ANNs integrate different kinds of learning into one network.

The learning of ANNs is generally classified as

- Supervised learning—ANNs are trained on a set of input-output pairs. The weights are adjusted to minimize error of the outputs. Another set of input-output pairs, called testing data, is provided to test the effects of training.
- Self-organizing learning—The network is trained on a set of inputs. No guidance is presented to the network about what it is supposed to learn. The ANN adjusts the weights to meet its own built-in criterion.
- Reinforcement learning—ANNs are trained on a set of inputs. The target values are not provided for learning; instead, error signals of the output are given to the ANNs. This process is analogous to reward or punishment.

Mapping ANNs

Of the four kinds of ANN, mapping ANNs have the most models. Basically, for mapping ANNs, if Y_i denotes the ith output, X_j denotes the jth input ($j = 0,1, \ldots, n$) and W_{ij} denotes the corresponding weight, then we have

$$Y_i = g\left(\sum_{j=0 \text{ to } n} W_{ij} X_j\right)$$

where g is the transfer function. Generally, g will be one of the following types:

- Linear function:

$$g(x) = x$$

- Threshold function:

$$g(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1(\text{or } 0) & \text{else} \end{cases}$$

- Sigmoid function:

$$g(x) = \tan h(x) \text{ or } g(x) = \frac{1}{1 + e^{-x}}$$

X_1 to X_n are the external inputs. X_0 is the artificial or the bias input, and it is added to simplify the network implementation. Without it the formula for the network would be

$$Y_i = g\left(\sum_{j=1 \text{ to } n} W_{ij}X_j - \text{threshold}_j\right)$$

The major models of mapping ANNs are given in the following.

Linear Associator

The linear associator is one of the earliest basic mapping ANNs; it was invented by several people during the period 1968 through 1972. It can learn at most L input-output vector pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)$. When one of these input vectors, say x_k, is entered in the network, the output vector y should be y_k. When a vector $x_k + \varepsilon$ (close to x_k) is entered into the network, the output vector should be $y_k + \delta$ (close to y_k)(*1*).

Learning Matrix

The learning matrix is a crossbar, heteroassociative, nearest-neighbor classifier. It was applied to problems such as highly

distorted handwritten chracters and diagnoses of mechanical failures to reduce downtime of machines.

ADALINE and MADALINE

ADALINE (Adaptive Linear Element) is described as a combinatorial logical circuit that accepts several inputs and produces one output. MADALINE (Multiple ADALINE) is a more comprehensive network consisting of many ADALINEs. Both ADALINE and MADALINE are commonly used models of mapping ANNs, and generally they operate with a least mean square error-correcting learning rule. The applications of ADALINE and MADALINE have been developed in control, pattern recognition, image processing, noise cancellation, and antenna systems.

Back Propagation

Back propagation is one of the best-known ANNs. The typical back-propagation ANN always has at least one hidden layer. There is no theoretical limit on the number of hidden layers, but typically there are no more than two. During the learning process, the error information is propagated back from the output layer through the network to the first hidden layer. Back propagation is a very powerful technique for constructing nonlinear transfer functions between a number of continuously valued inputs and one or more continuously valued outputs. This property leads to many interesting applications. Back-propagation ANNs have been used in solving many real-world problems such as image processing, speech processing, optimization, prediction, diagnostics, control, signal processing, noise filtering, and forecasting (*1*,*2*,*7*).

Self-Organizing Mapping

The self-organizing mapping ANNs can be used to sort items into appropriate categories. One of the most famous of these ANNs is the Kohonen layer, which was developed in Finland by Kohonen of Helsinki University of Technology between 1979 and 1982. The Kohonen layer basically implements a clustering algorithm to the network. In the operation, only one unit fires and takes a value of 1; the others will be 0. This process is accomplished by a winner-takes-all strategy. Self-organizing mapping can be easily adapted to handle categorization problems (*7*).

Adaptive Resonance Theory

The adaptive resonance theory (ART) was developed by Carpenter and Grossberg of Boston University in 1987. It includes three implementations: ART1 for binary inputs, ART2 for continuous-valued inputs, and ART3, which is the refinement of ART2. An ART network can classify and recognize input patterns without the presence of an omniscient teacher; that is, no instructor tells the network to which category each particular stimulus input belongs (*7*).

Recurrent ANNs

In recurrent ANNs the outputs are fed back to the network as a part of their inputs. This process is also described as a recurrent connection. The major models of recurrent ANNs are summarized in the following.

Hopfield

The Hopfield network was proposed in 1982 by Hopfield of the Biophysics Division of Bell Laboratories. It is a primary example of a recurrent network. The Hopfield network acts as an associative memory, that is, it passes through a sequence of several patterns and chooses one that most closely resembles the input pattern as the output. The stable patterns into which the network settles are called attractors, and the possible states of the network are called the configuration space. The sets of the states that eventually transform into the same attractor are called basins of attraction.

The Hopfield networks have been used not only in the common areas of ANNs such as image processing, signal processing, pattern matching, and prediction, but also to solve some classical combinatorial problems such as the "traveling salesperson problem." Its ability in optimization should also be highlighted.

Brain-State-in-a-Box

Compared with the Hopfield network, whose processing units are allowed to take only binary values, the brain-state-in-a-box (BSB) network has processing units that are allowed to take on any value from -1 to $+1$. BSB has been used in pattern classification, diagnostics, knowledge processing, image processing, and psychological experiments (*2*). It is reported that one of the BSB applications, the instant physician system developed by Anderson of Brown University in 1985, is performing surprisingly well (*1*).

Bidirectional Associative Memory

The bidirectional associative memory (BAM) network was developed in 1987 by Koskonow of the University of Southern California. It allows associations between two arbitrary patterns. A BAM network consists of two layers, with every unit in one layer connected to every unit in the other layer. BAM applications in image processing, control, resource allocation, and optical/electrooptical have been reported.

Boltzmann Machine

The Boltzmann machine was developed in 1984 by Hinton of the University of Toronto and colleagues. This network is a trainable, stochastic version of the Hopfield network. Hidden units constitute an important part of the architecture. Boltzmann machines have many applications, such as image processing, speech processing, temporal processing, prediction,

optimization, diagnostics, character recognition, knowledge processing, and signal processing (*2*).

Recurrent Back Propagation

Recurrent back-propagation ANNs can recognize dynamic patterns whose input vectors change with time; that is, a sequence of input vectors X(0), X(1), . . . , X(t_{stop} − 1) are represented to the network. A typical recurrent back-propagation ANN consists of two layers, a functional layer and a register layer. It handles a set of time-dependent input-output data instead of static input-output of back-propagation ANNs.

Temporal ANNs

The operation of temporal ANNs can be represented by a differential equation, and the output and input vectors of the network are changed with time. Fewer models of temporal ANNs have been developed. Generally speaking, temporal ANNs are suitable for dealing with the dynamic types of problems.

Hybrid ANNs

Hybrid ANNs combine supervised and unsupervised learning in one network. We summarize two models.

Counter Propagation

The counter-propagation network is a typical hybrid ANN initially proposed by Hecht-Nielsen. Counter-propagation networks consist of a linear mapping layer on top of a Kohonen layer. The Kohonen layer is trained in the usual way, and the linear mapping layer is trained by a simplified version of delta rule that is called ouster learning. Counter-propagation networks often run 10 to 100 times as fast as back-propagation networks when they are applicable to a situation.

GMDH

The GMDH ANN combines simple nonlinear processing units into an effective multilayer network. The most important feature of this kind of ANN is that it includes a procedure for building a near-optimal network.

In short, compared with traditional computation methods, ANNs behave as if they are depending on some kind of "intuitional reaction." They are concerned not with the principles of their operation but with the effects of their behaviors. This makes ANNs special and superior in solving certain problems, especially for recognition, control, image processing, optimization, and diagnostics. The structures of ANNs seem to be problem-dependent—for different types of problems, different structures should be used. In Table 1 we present application problems divided into 11 categories and rate the performances of 13 major ANN models in these categories.

APPLICATIONS OF ANNs IN TRANSPORTATION ENGINEERING

Transportation problems can be divided into five categories: planning, operation and control, administration and finance, construction and maintenance, and design. The applications are identified and the suitabilities of several major ANN models in different transportation engineering domains are discussed.

Planning

Transportation planning identifies a set of actions that will lead to the achievement of given objectives. The prediction of traffic conditions is a basic problem of transportation planning. Two of the most common traffic planning problems, trip generation and the origin-destination (O-D) distribution, are identified as suitable for the ADALINE and back-propagation ANN models, respectively.

Trip Generation

Trip Generation forecasting analysis predicts the zonal amount of traffic. The primary approach used for trip generation forecasting now is the regression method, which uses the statistical data of the past to establish a mathematical model used to compute the number of traffic trips required in the future. The ADALINE is identified as being suitable for this problem. The weights between the inputs and the outputs could be considered the regression coefficients in regression analysis (linear case). The training data sets are taken from past survey data. The output of the network can be a variable or a vector. If we take only one output, the model predicts the value of trip generation at a specific time in the future. If we take a sequence of outputs, it predicts the values of the trip generation in a time series in the future. This idea is also expected to be used in traffic attraction forecasting.

O-D Distribution

The trip generation problem described traffic volume production for a specific area. The O-D distribution describes the traffic volume between specific areas (called zones). The O-D forecasting is implemented according to the data sets of traffic generation and traffic attraction of the zones. A number of O-D forecasting methods have been developed; basically, they are a traffic O-D distribution pattern estimation model that predicts the distribution pattern on the basis of three factors: the future traffic generation, the future attraction (or one of them), and the present distribution pattern. The back-propagation ANN is identified as suitable for O-D forecasting. The future zonal trip generations and attractions can be identified as the external inputs of the network, and the outputs of the network represent future O-D distribution. Through proper training, which is a process of minimizing the total

TABLE 1 Application Evaluation of ANN Models

Category \ Model	Mada./ Ada.	ART	BAM	Hopf.	Boltz. Mach.	Back Prop.	BSB	Count. Prop.	Lin. Asso.	Learn Matr.	Koho.	GMDH
Recognition	●			●	●	●		○				
Control	⊠	⊠	⊠	⊠		●					○	
Forecasting/ Prediction	●					●						
Classification		●				●	⊠	⊠			●	
Diagnosis		⊠				⊠	●					
Optimization				●	⊠	●						●
Noise Filtering	⊠					●						
Image Processing	⊠	⊠	⊠	⊠	⊠	⊠	⊠					
Association			⊠	●	○		○		○	⊠		
Decision Making		⊠						●				
Temporal Processing	⊠	⊠				●						

Key:
- ● Strong Applicability
- ⊠ Moderate Applicability
- ○ Applicable
- ▭ Difficult to Evaluate

error of outputs, the noise in past O-D patterns can be removed. Our prototype development of an O-D forecasting model using a back-propagation network showed reliable feasibility.

Operation and Control

Operation and control of transportation deal with problems such as traffic congestion diagnosis and control, hazardous material transportation, air traffic control, and ground traffic signal timing control. Two applications are identified in this area.

Traffic Pattern Recognition System

Because of the difficulty of dynamic traffic assignment modeling, the most commonly used coordinated area traffic control systems are based on time interval–dependent control on the assumption that in one time interval the network traffic state (called the traffic pattern) is static. Obviously, the current method has two major defects. The first is that there is no proper criterion to identify the traffic patterns—that is, if the traffic pattern is changed we need a criterion that will identify the new traffic pattern and measure how much the new pattern differs from the old one so that a set of optimized control parameters can be presented. The second defect is that the tolerance of the traffic pattern changes—that is, when the traffic pattern does not change too much until it matches another set of control parameters, we usually do not want to change the control parameters because the changes may bring some traffic disorder. Against these two defects, the traffic pattern recognition system is proposed. At present, the ANN model used in this system is ART1, but it will soon be replaced with ART2. The ART properties of parallel process and tolerance adjustability can provide more reasonable traffic pattern recognition results than the traditional methods can.

Automatic Vehicle Identification

The inefficiency of toll booths has attracted the attention of many researchers. Several systems that may replace the functions of toll booths have been developed in the past few years. An automatic vehicle identification (AVI) system can provide the convenience of electronic vehicle identification, through the use of subscribed purchase transponders, as vehicles move through a toll plaza or checkpoint without stopping or, in some cases, even slowing down. The system automatically charges or debits tolls from the drivers' accounts. In 1988, such a system was set up on the Dallas (Texas) North Tollway, and 15,000 tags were issued. Three AVI toll roads being built in Orange County, California, reportedly will feature the most extensive use of AVI in the United States (*8*).

The most important advantage of the license plate recognition system is that no special devices are required for the vehicles passing through. In this system, the ANN is used to complete the task of character recognition. Several conventional computing methods are used to complete the statistical data processing and calculate the amount of toll fee for each motor vehicle that passes the checkpoint. This system is under development, and the character recognition software has been successfully completed. Another important advantage of this system is that it will be very helpful for the O-D survey.

Administration and Finance

The area of administration and finance deals with the sources and distribution of the money used for financing transportation systems and their improvement. Topics include innovative transportation financing techniques; the effect of deregulation on different modes, pricing, user payments, and cost; and the prioritization, allocation, and distribution of funds. Here, potential applications of ANNs in the innovative transportation financing field are described.

Many state and local transportation agencies have used innovative transportation financing techniques to cope with financial problems. Their efforts to attract private funding, along with specific case studies, may be found in published sources. A back-propagation ANN can be trained to learn the experiences of many state and local transportation agencies that have succeeded in generating funds by using innovative transportation financing techniques. The input to the network consists of such parameters as the transportation agency's size, financial needs, institutional elements, and the mode (i.e., transit, highway, pedestrian) for which financing is requested. The output includes the most appropriate financing techniques for the transportation agency to adopt and the amount of money that will potentially be generated.

Construction and Maintenance

Transportation construction involves all aspects of the construction process, including preparing the surface, placing the pavement materials, and preparing the roadway for use by traffic. Transportation maintenance starts some time after the construction and involves all the required work and procedures to keep the facility in working order. Under this topic the pavement traffic mark recognition system using appropriate ANN techniques is proposed.

Pavement traffic marks can inform motorists, guide traffic flow, and ensure traffic safety. The damage sustained by traffic marks painted on the roadway surface varies depending on the traffic that travels over them as well as the climatic conditions. Investigating the existence and the degree of damage to traffic marks is usually expensive and disruptive of the normal flow of traffic. The system currently under development uses a hybrid ANN (consisting of a Hopfield ANN and two multilayer back-propagation ANNs) to recognize the damaged traffic mark and classify it into one of several categories that represent the different degrees of damage.

Design

In transportation, design includes the visible elements of the transportation facility. It deals with factors such as the grade line or profile, horizontal and vertical alignments, pavement of roadway, terminals, and stations. The input to the design process includes such parameters as the forecasted design demand rate, type of demand, and environmental conditions such as weather and temperature. The output consists of the specific numerical or the range of values of the design elements. ANNs can be used for a variety of design problems in transportation. The ANN knowledge-storing ability is especially useful in the development of design supporting systems. Through proper training of the networks, the models can learn to provide suitable output results when provided with input values. For human experience–based design supporting systems, the most suitable ANN models are ADALINE, MADALINE, and back propagation.

In brief, the transportation applicabilities of ANNs in image processing, control, optimization, and forecasting are especially impressive. For some problems, such as traffic pattern recognition, ANNs obviously appear to be stronger than traditional methods. Some ANNs' unique abilities, such as pattern classification, filled some blanks in transportation engineering such as the traffic mark classification system. Table 2 shows the ANN models appropriate to transportation engineering applications.

CASE STUDY

As a demonstration of the applicability of ANNs in transportation engineering, we aimed trip generation forecasting, which is one of the most common transportation planning problems, using the ADALINE and the back-propagation ANN models.

Definition of Problem

Past studies have identified the fact that human traveling activities are related to some socioeconomic indexes such as income, population, and employment. For a specific area (or zone), the trip production can be conjectured from these indexes. The purpose of trip generation analysis is to develop a mathematical model that can be used to forecast the traffic

TABLE 2 Related ANN Models to Transportation Applications

TRANS. DOM.	APPLICATIONS	NEURAL NETWORK MODELS							
		BALZ. MACH.	ADA./ MADA.	BACK PRO.	HOP. NET.	ART 1	ART2	Cauchy MACH.	NEO COG.
PLANNING	TRIP GEN.		X						
	OD DISTR.			X					
OPERA. and CONTR.	SIG. TIMING OPT.			X				X	
	TRAF. PATTR. CLASSIFIC.				X	X	X		
MANAG. and FINAN.	TSP OPT.	X			X				
	LICENSE NO. REC.			X					X
CONSTR. and MAINT.	PAV. DISTRESS DETECTION			X					
	TRAF. MARK CLASSIF.			X	X				
DESIGN	DESIGN SUPPORT SYSTEM		X	X					

trip production of the forecasting zone according to a number of socioeconomic indexes. For example, two indexes, zonal households and zonal population, are used in one major Canadian transport planning study to predict the zonal peak-hour trip production (*9*).

zonal peak-hour
trips produced = 0.3036 (zonal households)
+ 0.5638 (zonal population)

Current Solution

The previous approaches for trip generation forecasting include two main methods: regression analysis and category analysis. The majority of trip generation studies performed to date have used multiple regression analysis to develop the prediction equations for the trips generated by various types of land use. Generally, the least squares method is used to determine the constant c and the coefficients.

Description of ANN Models

An ADALINE ANN was used in this case study. This very simple ANN model has only four input units and one processing unit. One of the obvious differences between ADALINE and the regression method is the handling of the optimization of the weights and the coefficients. The regression method pursues the coefficients that will produce the minimum error on the surveyed data, which can be considered the training data sets for the ADALINE model. The training of ADALINE pursues the best value of the weights that will

TABLE 3 Observed Data of Trips and Index

No.of single family houses x1	Permanent single family residents x2	Number of multi-family dwelling units x3	Number of families with single auto ownership x4	trip rate y		
48	116	130	89	234		
71	126	24	48	132		
47	84	26	37	108		
26	46	10	18	49		
221	477	32	127	412		Regression data
114	275	41	78	247		
41	73	25	25	69		
167	360	31	99	320		
397	916	37	217	741	Training data	
264	748	23	144	549		
269	763	38	154	586		
171	485	43	107	388		
179	507	26	103	384		
233	661	8	121	471		
185	524	26	103	396		
252	714	48	147	554		
275	779	72	170	627		
175	470	58	114	400		
128	343	67	96	321		
83	179	15	49	159		
401	1073	148	269	934		
52	93	20	36	99		
26	46	9	18	53	Testing data	
83	179	15	49	158		
3	5	0	2	4		
4	8	2	3	8		
416	1146	468	433	1273		
8	14	0	4	12		
30	65	11	22	64	Forecasting data	
22	39	5	14	53		
41	99	173	107	271		
124	121	51	88	240		

allow the model to obtain good results on the testing data sets, but not on the training data sets. Even if a set of weights will allow the model to perform well on the training data sets, unless those values of the weights will allow the model to reach the approximate error minimum on the testing data sets, those weights are not considered good. In fact, this phenomenon is called over training. An ADALINE process such as this is expected to achieve better results than the regression method does. The belief in the training of ADALINE described here is also a general characteristic of ANNs.

A back-propagation model is also applied. This is a two-layer model with four input units to match the number of indexes and two processing units in the hidden layer. Unlike the usual way, which uses the sigmoid function, a linear transfer function is used in this model to match the ADALINE so that we can compare the performance of the two models. The other purpose is to examine its operation when a linear function is used. The learning of back propagation is done in the usual manner, that is, by using a delta learning rule.

Computation

The data sets are shown in Table 3. The first 20 sets of the samples were used as the training data, and the next 7 sets of samples were used as the testing data for ANN models. The other five data sets of samples are used to verify the effects of the three methods, and hence we call them forecasting data. Both the training and testing data were put together for the regression analysis.

In the training of the ADALINE model, the mean squared error (*MSE*) is used to indicate how the training is going. When iterations were equal to 10,000, we reached the approximate minimum of which the $MSE = 0.000195$, and $MSE = 0.000053$. The training of the back-propagation model went much faster than the ADALINE. At iteration equal to 2,500, the minimum *MSE* on the testing data was detected at 0.00004242. The *MSE* on the training data at iteration equal to 2,500 was 0.000005035.

Figure 2 shows the forecasting results obtained using the three methods on the forecasting data, respectively. We can see that both of the ANN models forecasted values closer to the actually observed values than the regression does. The *MSE*s of these three methods are shown in Figure 3.

Discussion of Results

The results obtained indicate that ANN models work better than the linear regression method in this case. The results of ADALINE are slightly closer to the observed values than those of back propagation. This is considered to be the result of using floating data type in all the computations and then the stacked errors caused this small difference. However, the ADALINE took 10,000 iterations whereas the back propagation took only 2,500 iterations to reach the total error min-

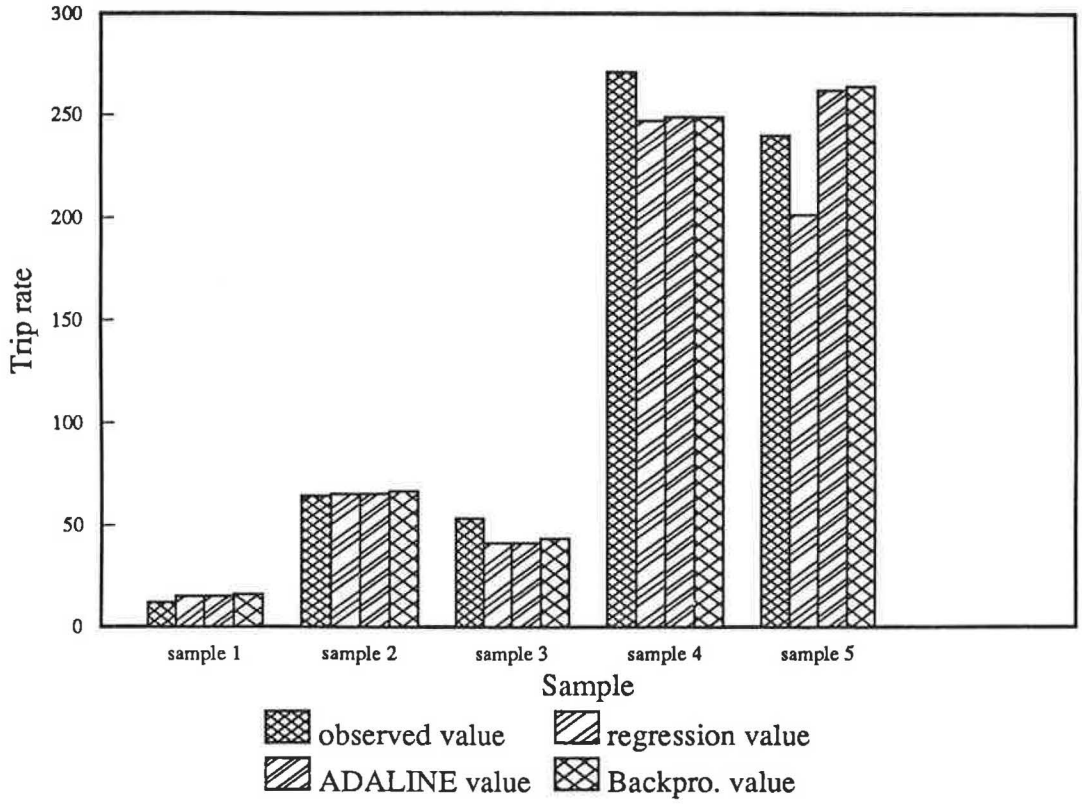

FIGURE 2 Results of regression, ADALINE, and back propagation.

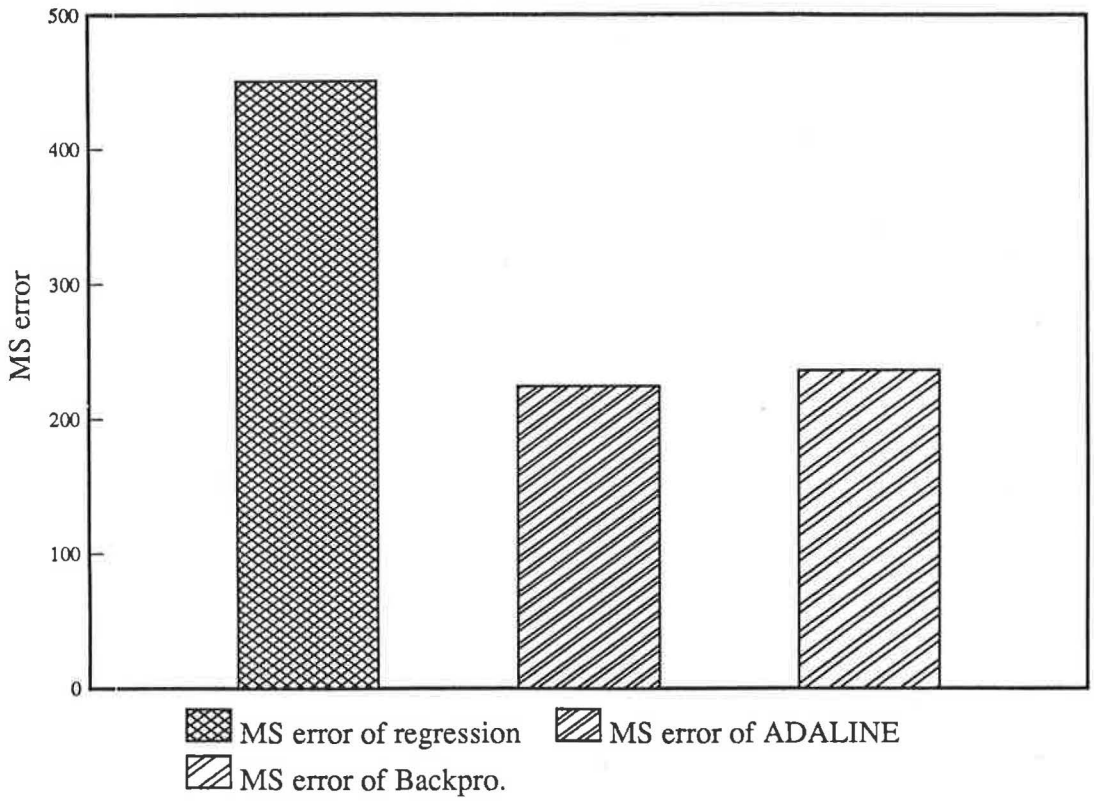

FIGURE 3 Comparison of *MSE* of regression, ADALINE, and back propagation.

imum on the testing data sets in the training. The training of the back-propagation model is much more efficient.

SUMMARY AND CONCLUSIONS

The fundamental concepts of ANNs were introduced, and the basic differences between ANNs and biological neural networks and expect systems were presented. On the basis of their unique architecture and learning techniques, ANNs were classified into four categories, and under each category, several ANN paradigms were explained. A matrix for relating each ANN paradigm to 11 attributes, including classification, pattern recognition, image processing and diagnosis, was then presented. On the basis of this matrix, the applicability of different ANN techniques for solving transportation problems was evaluated. Finally, a case study demonstrating the applications of two powerful ANN techniques—back propagation and ADALINE—in solving the trip generation problem was demonstrated. The results outperformed those obtained by conventional regression methods.

The strong applicability and suitability of different ANN techniques were demonstrated in this paper. Many of the example problems presented in the transportation application section have already been developed or are currently under development by the authors. ANNs are envisioned to become powerful tools not only for transportation, but also for enhancing the current state of such contemporary issues as artificial intelligence applications and intelligent vehicle-highway system technologies.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support for this project provided by a University of Delaware Research Foundation grant.

REFERENCES

1. R. Hecht-Nielsen. *Neurocomputing*. Addison-Wesley Publishing Company, Inc., Reading, Mass., 1990.
2. P. K. Simpson. *Artificial Neural System*. Pergamon Press, Inc., New York, N.Y., 1990.
3. M. Caudill. Neural Networks Primer Part 1. *AI Expert*, Dec. 1987, pp. 46–51.
4. B. L. Eliot. Neural Networks: A Comparison with Expert Systems. *Expert Systems Trends*, Winter 1990, pp. 72–75.
5. T. Nakatsuji and T. Kaku. Development of a Self-Organizing Traffic Control System Using Neural Network Models. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991.
6. S. G. Ritchie, M. Kaseko, and B. Bavarian. Development of an Intelligent System for Automated Pavement Evaluation. In *Transportation Research Record 1324*, TRB, National Research Council, Washington, D.C., 1991.
7. *Neural Computing Software Manual*. NeuralWare, Inc., Pittsburgh, Pa., 1991.
8. R. L. Hartje. Tomorrow's Toll Road. *Civil Engineering*, Feb. 1991, pp. 60–61.
9. B. G. Hutchinson. *Principles of Urban Transportation Planning*. Scripta Book Company, 1974.

Publication of this paper sponsored by Committee on Expert Systems.

Validation of an Expert System: A Case Study

MICHAEL J. DEMETSKY

The steps taken to verify and validate a prototype expert system (TRANZ) for traffic control through highway work zones are described. The prototype is viewed as an applied statement of the system requirements and a focus for the development of a complete knowledge base. The tasks used in the validation included (a) revisiting the experts who assisted in developing the program, (b) selectively distributing validation copies of TRANZ, (c) identifying problems and decisions that interface with and affect the work zone traffic control problem, and (d) conducting a validation workshop. The workshop was found to be the most effective way to review the system because the results represent a group consensus, whereas the prior tasks encompassed only individual inputs obtained in isolated instances. The informal request for reviews from users was completely ineffective. The workshop results were interpreted as general comments on the overall concept of TRANZ and specific programming modifications that resulted from erroneous recommendations by TRANZ. The general comments indicated that the attendees would not completely rely on TRANZ for finding traffic control solutions for work zones. However, they did indicate a willingness to work with TRANZ in the field and during instructional programs and thus to bring it along slowly while carefully validating it. The conclusions show the relatively long time needed to continue to validate and update an expert system until it correctly addresses a good majority of cases. The specific changes recommended by the panel demonstrated how the basic prototype can be expanded through experience. In the case of TRANZ, the knowledge base was expanded by adding new rules. The strategy used of asking experts to use the system can quickly expand the set of problems that a system addresses via direct expert input. Overall, the validation process must address both the accuracy and the completeness of the system.

A critical stage in developing an expert system is verifying and validating it as an acceptable piece of applied software. It must be proved that the system is an accurate and useful representation of knowledge. At present,"bits and pieces of a verification and validation methodology currently exist, but have not been assembled and standardized due to the many applications, design paradigms, development approaches, and the stage of development and fragmentation of the industry"(*1*).

Verification has come to deal with the program text development, which is simplified when a shell such as EXSYS is used. Validation "is a determination that the completed program performs the functions in the requirements specification and is usable for the intended purposes" (*2*).

This paper focuses on the validation of TRANZ, which is an expert system for traffic control in highway work zones (*3*,*4*). It describes the steps in the validation process and the relevant findings.

The function of TRANZ is to specify appropriate traffic handling strategies for a specified highway work zone operation. The objectives of traffic control in maintenance work zones are (a) protecting the freeway users and the work force, (b) moving the maximum volume of traffic (minimization of delay), and (c) providing efficiency and economy in work procedures. TRANZ achieves these objectives by interpreting construction control and roadway factors to recommend appropriate traffic controls.

BACKGROUND

In recent years, researchers have been investigating and developing knowledge-based expert systems for transportation engineering applications. These computer programs, however, are not ready for commerical use. The state of the art is a range of prototypes that typically require extensive testing and refinement before they are acceptable for regular use. These systems typically are used only by the developer or sponsoring agency. These prototypes are an applied statement of the system requirements and provide a focus for the ultimate development of a complete knowledge base (*1*). The prototypes provide a framework for a continuous process of working with experts in supplying the required knowledge.

Factors that limit the utility of these prototypes are (a) the scope of the problem domain addressed by a prototype in relation to the appropriate scope of issues influencing decisions (this conflicts with the notion that the narrower the scope, the better the expert system), (b) the inability of the system to combine rules or other logic representations suitable for simple situations into more complex relationships suitable for more complex situations, and (c) the accuracy of the representation of the decision process of the expert(s).

This paper describes the steps taken after a prototype expert system was developed to validate the system as a decision aid to engineers.

OVERVIEW OF TRANZ

Figure 1 illustrates a comprehensive formulation of traffic management tasks for highway work zones. This model describes the work zone scenario and indicates how the traffic volume through the work zone is derived after demand management strategies have been implemented. The TRANZ module, in this case, reflects the traffic control plan for the

Virginia Transportation Research Council, Box 3817, University Station, Charlottesville, Va. 22903.

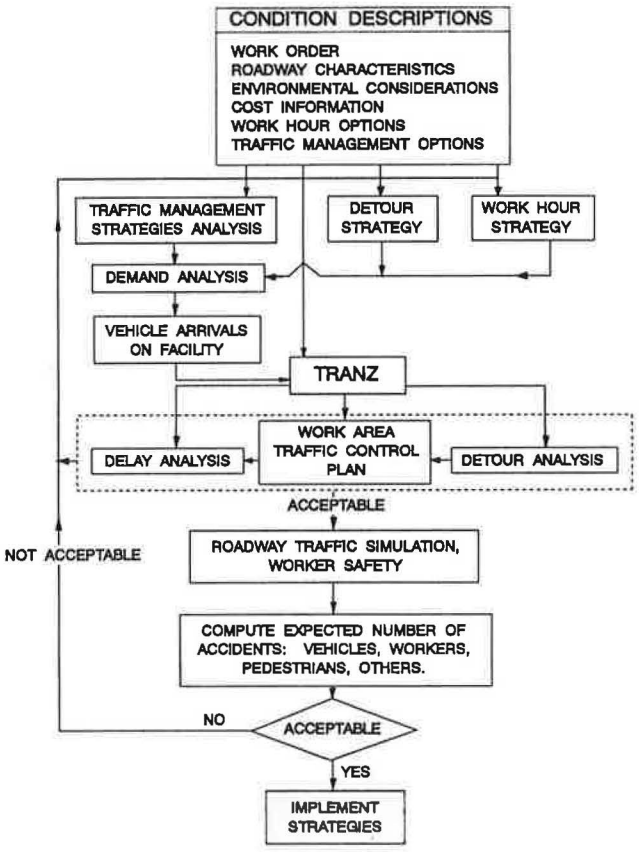

FIGURE 1 Traffic management tasks for highway work zones.

work area plus detour and delay analysis to examine the adequacy of the work zone to handle the generated traffic. Figure 2 illustrates the micro decision framework of TRANZ for which the rule base was developed. This figure shows the range of control options (barrier, cones, drums, signs, vehicles, etc.) and associated conditional variables [roadway type, location of work, average daily traffic (ADT), etc.] that interact to produce control requirements. All of the rectangular boxes in Figure 2 show places at which a second-level decision analysis takes place and a further decision tree can be established.

The *Virginia Work Area Protection Manual* (WAPM) offers a selection of typical traffic control plans for certain work activities (*5*). These plans provided an initial collection of cases for the prototype system. The decision tree in Figure 2 provides a framework for extending the WAPM knowledge box by working with experts who have field experience.

VALIDATION PROCEDURE

The overall validation plan for TRANZ actually began before the publication and release of the prototype. The prototype that was distributed for limited validation by the public is an improved version of the original prototype. The actual validation began with the assignment of a second software engineer to the project. The knowledge engineer of the first prototype was a transportation engineer who used an expert system shell (EXSYS) to code the knowledge base into rules. The second knowledge engineer was a computer scientist who took a more mechanical view of the system and improved its efficiency by an informal validation that included correcting logical errors, modifying rule specifications and structure, and enhancing the user interface.

The validation consisted of the following:

1. Revisiting the experts who helped develop the system to receive comments about the current prototype, and revising the prototype as warranted.
2. Selectively distributing copies of TRANZ with documentation and a validation form to Virginia Department of Transportation (VDOT) personnel and others requesting it. Users were asked to apply the system and report their observations on the forms. The intent was to compile an exhaustive list of case applications to enhance the knowledge base. If necessary, respondents were to be contacted by telephone to clarify the data. The information sought included the appropriate input data for TRANZ and a description of the traffic control plan as implemented.
3. Performing research on related problems that interface with and affect the work zone traffic control problem—This would allow expansion of the system so that it encompassed a broader and more complete decision problem than the prototype covered. Issues considered included queuing and delays to traffic, traffic diversion strategies (facility demand management), detour alternatives, time of day for the work effort (including nighttime), delineation of traffic lanes through work zones, and worker safety (including applications of new technology).
4. Conducting a workshop with approximately eight experts on traffic management through highway work zones to finalize the initial validation of TRANZ for field applications.

The version of TRANZ used in Task 4 is a version that reflects the results of the first three tasks. Most of this paper will focus on the workshop (Task 4) because it was there that the status of the system as an aid to transportation practitioners was tested. The results represent a group consensus, whereas the prior tasks encompassed only individual inputs obtained in isolated instances. However, summary statements about the other tasks are provided.

EXPERT REVIEWS

The first step in the formal validation of TRANZ required the participation of the knowledge engineer, an expert, and two users. The first user was a novice safety engineer who was interested in using TRANZ in an office environment for consultation and learning. The second user was interested in seeing TRANZ used in seminars and short courses in freeway work zone safety.

There were two tests. The first applied to TRANZ to 11 problems that were used for a short course sponsored by the Virginia Transportation Research Council. To these 11 textbook problems, TRANZ gave correct solutions in all cases. This could be expected because the manuals were used in the development of the knowledge base, and no judgment was required beyond that given in the WAPM (*5*). The second test applied TRANZ to six actual problems that were provided

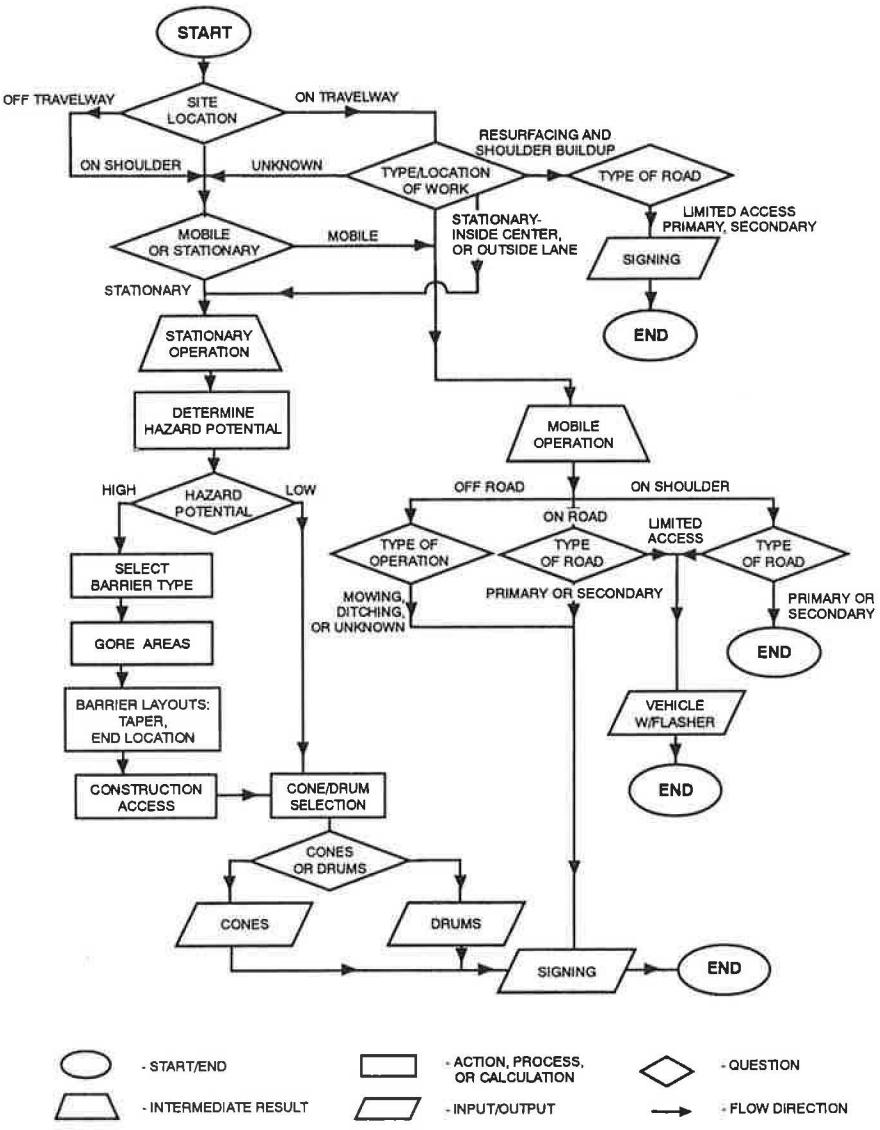

FIGURE 2 TRANZ decision framework.

by the Staunton District of VDOT. In this test, TRANZ and the expert disagreed in four of the six cases.

Bridge Deck Maintenance

Problem

A bridge deck operation is being conducted on a two-lane, two-way secondary road with ADT of 5,000 and an anticipated operating speed of 45 mph. This operation is being conducted off of the travelway with the accident factor P greater than 0.5 (P is the expected number of run-off-road accidents that will happen in the particular hazard time, T).

Results

In this case, the expert used temporary concrete traffic barriers but agreed that other devices recommended by TRANZ would also be applicable. However, the expert pointed out the fact that TRANZ did not recommend any signs such as Construction Ahead or Reduce Speed Ahead.

Pavement Milling and Plant Mix Operation

Problem

A pavement milling operation on one lane of a four-lane Interstate highway.

Results

In this case, the expert agreed with all TRANZ's recommendations but thought that the recommended values should have been 9 on a scale of 0 (impossible) to 10 (certain), and TRANZ failed to recommend the necessary signs for this project.

Pipe Placement

Problem

This pipe replacement project is being conducted on a two-lane two-way secondary road. About 1,200 vehicles with an average speed of 45 mph traverse the site. The value of the accident factor P is greater than 0.5, and the operation is being conducted off of the shoulder.

Results

The expert completely agreed with TRANZ's recommended devices but mentioned TRANZ's failure to recommend the appropriate signs.

Excavation Project

Problem

Excavation activity with equipment near the roadway is being conducted on I-81 off of the shoulder. The operating speed of vehicles passing through the zone is about 60 mph, and the ADT is more than 15,000. The P value is greater than 0.5.

Results

The expert completely agreed with TRANZ's recommended device but mentioned TRANZ's failure to recommend the appropriate signs.

Mowing Operation

Problem

Mowing operation is being conducted on the inside shoulder of I-81. More than 150,000 vehicles are expected to pass through the work zone at an average speed of 60 mph. The P value is greater than 0.5 for this project.

Results

The expert completely agreed with TRANZ's recommendation.

Pavement Patching

Problem

Pavement patching on the inside lane of I-81. More than 15,000 vehicles are expected to pass through the work zone with an average speed of 60 mph. The P value is greater than 0.5 for this project.

Results

The expert completely agreed with TRANZ's recommendation.

The results of these six problems directed the knowledge engineer to modify TRANZ so that the system's recommendations would agree with the expert. Essentially, the disagreements were associated with the lack of signage in TRANZ's solutions. This is an easy modification to make.

This test was relatively simple in comparison with the evaluations that were accomplished in the workshop phase. However, these tests were beneficial in the early testing of TRANZ because significant improvements resulted from them.

DISTRIBUTION OF PROTOTYPE

One of the reasons EXSYS (*6*) was selected to develop TRANZ was that FHWA has a license for the run-time version. This made it possible for a limited number of copies of TRANZ to be distributed to potential users including members of state DOTs outside Virginia. Because TRANZ followed procedures used in Virginia and because procedures for directing traffic through or around highway reconstruction zones differ among states, the Virginia prototype was directly useful for practice only in Virginia. It would need to be modified for use in other states, notably California and New York. A form for documenting the problems to which TRANZ was applied was included in the distribution. Very few cases, however, were recorded on the forms and returned. It was concluded that this approach was unrealistic, and the workshop strategy was initiated to meet the objectives of both tasks.

PROBLEM INTERFACES WITH TRANZ

With given data on the job and roadway environment (condition descriptions), various options for traffic management are available. Any option will use information obtained from a series of appropriate analyses. TRANZ focuses on the traffic controls of the affected facility, but additional considerations are usually relevant. For example, the design of transit schemes for diverting traffic is not included in TRANZ. However, TRANZ does interface with detour considerations and construction work-hour choices. These three strategies reduce overall traffic on the facility during reconstruction. Given the final demand estimate and the condition descriptions that have been prepared exogenous to TRANZ, the system then defines the appropriate traffic control plan, aids in the evaluation of the adequacy of any proposed detour, and calls the QUEWZ program (*7*) to compute delay on the facility. If any components of the traffic management plan are inadequate, the analyst must go back and alter the demand plans to arrive at an acceptable plan. Once an acceptable strategy or traffic management plan is formulated, the safety of the traffic flow through the work zone should be evaluated. The capability to assess safety does not exist in the current TRANZ, but a simulation model similar to the QUEWZ program could be coupled with it to perform this task.

EVALUATION WORKSHOP

Eight transportation engineers from different divisions of VDOT were invited to attend a workshop for validating TRANZ. Five representatives from the traffic engineering, location and design,and construction divisions and three representatives from the district engineers offices participated. The Richmond Division Office of FHWA was also represented.

Before the workshop, copies of a notebook and a TRANZ disk were sent to the attendees. They were to familiarize themselves with TRANZ and the issues to be addressed. The attendees were requested to document applications (as in Task 2) for discussion at the meeting.

The results of the workshop were first stated in terms of both comments received and in terms of specific tasks that rendered a validated version of TRANZ.

General Comments

1. "The TRANZ expert system should be a very beneficial tool for engineers dealing with the development of traffic control plans . . . it should not replace the important aspect of engineering judgment and the knowledge and experience gained from the individuals who are responsible for the traffic control devices in the field. In other words, you should generally know what the answer will be before applying TRANZ." (This indicates a lack of confidence in TRANZ in that it should be used only by experts themselves as a check on plans. This concern should be lessened as experience with TRANZ is gained.)
2. "One . . . use [of TRANZ] could be to allow students in a classroom situation to solve some problems using the manual and some using TRANZ, or a combination of the two. Feedback from the instructor could then be an asset in considering further upgrades of TRANZ." (This is a necessary stage if TRANZ is to become a reliable tool.)
3. "Once the existing logic is refined, field-tested, and the bugs worked out, one desirable feature for consideration . . . would be the inclusion of graphics in both the screen as well as printed output." (This could be accomplished with the integration of a laser disk with TRANZ.)
4. "When the program is revised, it is recommended that the new software be provided to the districts for a further evaluation. This should be a good test for the system."
5. "Traffic engineering and location and design strongly support the program and we will be glad to assist in any way to implement the project."

These comments indicate that the attendees would not completely rely on TRANZ for controlling traffic in work zones. However, they also indicate a willingness and desire to work with TRANZ in the field and during instructional programs and thus to bring it along slowly while carefully validating it. These conclusions show the relatively long time needed to continue to validate and update an expert system until it correctly addresses a good majority of cases.

TRANZ deals with a very complex and open-ended problem. It will require a long period of testing and revision before it becomes a complete knowledge system. This clearly requires a continuing effort toward maintaining the expert system, which is not normal for software developed and used only at the state level; but it is the norm for many federally supported software packages such as HCM and NETSIM, which are distributed and supported through McTRANS. Maintenance of an open-ended software package such as TRANZ is much more critical and consumes many more resources than maintenance for a conventional algorithmic program. Accordingly, this issue must be addressed when a state DOT plans for the development of an expert system.

Programming Modifications

Specific recommendations for improvements in the overall program derived from applications of TRANZ to real problems by experts included the following:

1. When the program was run, there was some uncertainty in selecting the option of whether or not the work crew was exposed to traffic. This operation should be given further explanation to ensure consistent application, preferably on the screen where the question is displayed. For example, "Does this mean before traffic control devices are installed?" or "Does this mean the crew will be working in the lane(s) of travel?"
2. Selecting Resurfacing and Shoulder Buildup as the type of operation on an Interstate resurfacing job does not give the desired level of work zone protection, but selecting Stationary Operation gave the desired result. The resurfacing option should be further explained so that the user will know up front when to select this option and precisely what it means.
3. In certain situations, it is not clear if the recommended devices are alternatives or are to be used in conjunction with one another. For example, on an Interstate lane closure, Barrier A, Temporary Concrete Parapet, Temporary Concrete Traffic Barrier Drums, and Temporary Asphalt Median all had a value of 9 on the output screen. Discussions at the workshop revealed that drums would be used in conjunction with the physical lane closure device but that a Type A device would not be used in conjunction with a Temporary Asphalt Median. Devices that are alternatives to one another should be clearly shown on the output.
4. Several recommended signs were indicated only by the code in the manual such as W20-7A or W4-2. It would be desirable to have a short verbal description with each recommendation.
5. The inclusion of QUEWZ is a favorable option. Perhaps the TRANZ manual could include the title page and summary pages ii and iii from QUEWZ as information for the operator who is unfamiliar with its use.
6. TRANZ should include quantities for the recommended traffic control devices.
7. New accident data for different classes of roads are used by the department to replace the single bar graph listed in the Work Zone Safety Short course Notebook for run-off-roadway accidents. Also, new values for different road systems by ADT levels are recommended to be used in the following formula:

$$p = f * t * l$$

where

p = expected number of run-off-the-road accidents,
f = accident frequency factor,
t = particular hazard time, and
l = length of hazardous fixed object (mi).

These changes should be made in TRANZ.

8. In the question-and-answer process of TRANZ, the following adjustments should be made: (a) define "Barrier A" in the answers as "concrete"; (b) define "Barrier B" in the answers as "guardrail"; (c) define terms "on travelway" and "off travelway," and (d) "travelway" perhaps should read "edge of pavement."

These are the key statements made by the panel concerning changes that should be made to clarify the TRANZ question-and-answer process. Appropriate corrections were made in the revised version of TRANZ.

Errors in Recommendations of TRANZ

TRANZ gave incorrect solutions to the following problems, according to the experts:

1. *Problem*: Interstate facility, ADT = 20,300 vehicles per day (VPD), operating speed = 68 mph; replacing existing concrete pavement, 2-mi segments, four lanes, close one lane while work is under way in the adjacent lane. *Solution*: TRANZ provides different traffic control devices for the inside lane and outside lane.WAPM 6-79 is used for the inside lane and WAPM 6-83 for the outside lane. In this problem, WAPM 6-83 should have been used for both inside and outside lanes. The only difference should be the messages on the signs. The WAPM does not provide a typical drawing for both the inside and outside lanes for the same type of construction. In other words, if the typical drawing indicates the inside lane, then the same drawing would apply to the outside with word changes on the signs.

2. *Problem*: Four-lane divided primary route, ADT = 40,000 VPD, operating speed = 35 mph; excavation of a 10-ft vertical drop trench in the median 3 ft from the edge of pavement, 15 ft wide, and 30 ft long. *Solution*: TRANZ indicates channelizing devices (Group 2 drums) as the solution. It appears that signing should also be included in this solution and in solutions to similar problems.

3. *Problem*: Limited-access roadway; on travelway, resurface and shoulder build-up operation. *Solution*: The sign layout provided is wrong. It should be the same as for a mowing operation.

4. *Problem*: Drop inlet existing in the median of a four-lane limited-access roadway; 30-day work operation is stationary and off the travelway; work crew is not exposed to traffic; operating speed = 55 mph; ADT = 37,500 VPD; hazard length = 0.3 mi. *Solution*: A minimum sign layout should be given that includes Roadwork Ahead and End Roadwork.

5. *Problem*: Stationary work off the travelway is being conducted for 120 days on a four-lane limited-access highway; the work crew is exposed to traffic; operating speed = 65 mph; ADT = 37,500 VPD. *Solution*: A barrier is specified by TRANZ. A sign layout should also be displayed. Also, the distance from the traveled roadway should be considered. On some Interstates, this work could be within 25 ft of the traveled roadway.

6. *Problem*: Stationary work off the travelway is being conducted for 120 days on a four-lane primary highway; the work crew is exposed to traffic; variable operating speed = 55 mph; ADT = 30,000 VPD. *Solution*: Same as Solution 5 except on a primary highway. This work could be behind the ditch line but within 10 to 15 ft of the traveled roadway.

7. *Problem*: A stationary operation between the travelway and ditch line is being conducted on a four-lane primary highway for 120 days; the work crew is exposed to traffic; operating speed = 55 mph; ADT = 30,000 VPD; median width = 50 ft. *Solution*: Distance from the roadway should be a factor. This work could be anywhere from 1 to 20 ft from the edge of the pavement. If it were 1 ft, lights would be needed.

8. *Problem*: A stationary operation between the travelway and the ditch line is being conducted on a four-lane limited-access highway for 120 days; the work crew is exposed to traffic; operating speed = 65 mph; ADT = 37,500 VPD. *Solution*: When work is on an Interstate or divided primary, the left shoulder also needs to be considered. The program now assumes everything is on the right.

9. *Problem*: A stationary operation is being conducted off of the travelway on a four-lane Interstate highway for 120 days; a nonremovable fixed object near the travelway exists for 2.5 mi; the work crew is not exposed to traffic; operating speed = 65 mph; ADT = 43,130 VPD. *Solution*: If a barrier is specified, there should be a minimum sign layout of Road Work Ahead and End of Roadwork.

10. *Problem*: The work is a deck replacement on the inside lane of a four-lane limited-access highway; the work is to be done between 8:00 a.m. and 4:30 p.m. for 4 months; the length of the work zone is 300 ft; the work crew is exposed to traffic; operating speed = 65 mph; ADT = 35,000 VPD. *Solution*: The solution provides devices that are alternatives to one another, but it does not indicate what they are. The temporary asphalt median recommended is not likely to be used on an Interstate highway.

11. *Problem*: Resurfacing job on the travelway of the outside shoulder on a four-lane limited-access highway; the length of the work zone is 10,000 ft; the stationary work crew is exposed to traffic; the duration is 90 days; work is conducted between 8:00 a.m. and 4:30 p.m.; operating speed = 65 mph; ADT = 29,569 VPD; if barriers are used, access openings to the construction site will be used by work vehicles entering the main traffic flow. *Solution*: TRANZ provided an incorrect solution, according to the experts. The solution should include advanced construction signs, taper lane closure, drums or cones, 72-in. concrete barriers, and a flashing arrow. Selecting "stationary operation in the outside lane" gave the correct solution.

12. *Problem*: A one-way deck operation on Route 60, which is a two-lane undivided primary highway, between 8:00 a.m. and 4:30 p.m. for 120 days; the length is 400 ft for a stationary operation where the work crew is exposed to traffic; no access through the barrier is required; gore areas are not present; operating speed = 55 mph; ADT = 3,255 VPD. *Solution*: TRANZ specified a flagger, but in the actual case reviewed,

a temporary traffic signal was used; the solution gave Barrier B as an option, but these are not used on a bridge deck.

These changes resulting from the experts' use of TRANZ demonstrate how the basic prototype was expanded through testing. Specifically, the knowledge base was expanded by adding rules. Thus, this evaluation expanded and reexamined the knowledge acquisition process. A large rule base presents a complex set of possible outcomes based on the application of relevant rules. Incorrect recommendations are expected from prototype systems. The strategy used here can quickly expand the set of problems that a system addresses by way of direct expert input.

QUALIFICATIONS

At this time, TRANZ can be recommended as either a check on a plan for work zone traffic control in Virginia or as a first formulation of such a plan. If it is used as a first formulation, experienced engineers should continue to verify the recommendations from TRANZ until the user community is comfortable with the accuracy of TRANZ. In either case, TRANZ should replace at least one expert in the process and provide savings in time and costs. TRANZ can be used in short courses or as a pseudotutor for individuals who wish to become familiar with Virginia's WAPM. Because the WAPM's problems and solutions have been validated for TRANZ, it can help novices in learning to use it or substitute for it.

Since the purpose of the project from which TRANZ was developed was to demonstrate expert system applications in transportation engineering, the completion of TRANZ as a validated professional tool was beyond the scope of the effort. The present version of TRANZ meets the study objective by providing a case study that demonstrates

1. The development of an expert system, which incorporates standard engineering procedures with expert judgments and interpretations;
2. The programming complexities of combining rules, calculations,and external programs in a complete decision support system;
3. The need to identify the role of the expert system in a broad system decision framework (i.e., its interaction with other decisions and the assumptions governing the scope of the system);
4. The identification of inappropriate (voluntary) and appropriate (controlled) validation procedures; and
5. The identification of the continuing maintenance and support requirements necessary for an expert system to remain relevant.

CONCLUSIONS AND RECOMMENDATIONS

The results of this study reveal that a complex expert system must be patiently developed until it behaves as an "educated expert." The evaluation of TRANZ revealed that it is difficult to quickly develop a system that will accurately handle all possible permutations of a problem.

Widespread distribution of a prototype expert system with a request that users validate it resulted in little feedback.This pointed to the need for a structured validation process.

This case indicates that the application of a prototype expert system, rather than completely solving a problem, may actually show a need to expand the scope of the decision problem. Accordingly, the system design can be expanded to interface with complementary decisions. In this case, traffic control strategies are seen to be related to other traffic management strategies including detours, work-hour choices, and transit diversions.

Finally, an evaluation workshop should be seriously considered in the evaluation of any expert system. The workshop led to an expanded knowledge base rather than to corrections of TRANZ's logic. Accordingly, the validation process must focus on both the accuracy and the completeness of the expert system.

Overall, the effort in validation of an expert system that was discussed herein prompted the following summary recommendations for agency policy regarding expert system validations in the future.

1. Personnel and resources need to be budgeted for 3 to 4 years beyond the development of a prototype to continue to evaluate, support, and update the system.
2. Once a focused prototype is developed, it must be slowly modified; it cannot be accelerated.
3. The validation process should use the workshop as a focus to establish a core group of supporters to continue to work with the developer to apply the system and include it in appropriate training programs.

REFERENCES

1. J. R. Geissman and R. D. Schultz. Verification and Validation on Expert Systems. *AI Expert*, 3, Vol. No. 2, 1988.
2. D. Barnett, C. Jackson, and J. Wentworth. *Developing Expert Systems*. Report FHWA-T5-88-022. Office of Implementation, FHWA, McLean Va., 1988.
3. A. Faghri and M. J. Demetsky. *A Demonstration of Expert System Applications in Transportation Engineering, Volume II, TRANZ: A Prototype Expert System for Traffic Control in Highway Work Zones*. Report FHWA/VA-89/R5. Virginia Transportation Research Council, Charlottesville, Va., 1988.
4. A. Faghri and M. J. Demetsky. Expert System for Traffic Control in Work Zones. *Journal of Transportation Engineering*, ASCE, Vol. 116, No. 6, 1990, pp. 759–769.
5. *Virginia Work Area Protection Manual*. Virginia Department of Transportation, Richmond, 1987.
6. *Expert System Development Software*. EXSYS, Inc., Albuquerque, N.M., 1986.
7. J. L. Memmott and C. L. Dudek. *A Model to Calculate the Road User Costs at Work Zones*. Report FHWA/TX-83/20+292-1. Texas Transportation Institute, College Station, 1982.

Publication of this paper sponsored by Committee on Expert Systems.

Model for Optimum Deployment of Emergency Repair Trucks: Application in Electric Utility Industry

K. G. Zografos, C. Douligeris, and L. Chaoxi

The uninterrupted supply of electricity is an important criterion for measuring the quality of service in the electric utility industry. An effective method for reducing service unavailability is to reduce the response time of emergency repair trucks (ERTs). An integrated methodological framework for the optimum deployment of ERTs is presented. The proposed methodology is based on an iterative procedure that involves the partition of a given geographical area into service territories and the subsequent simulation of the emergency repair operations within each service territory. A real-world problem is used to demonstrate the applicability of the proposed methodology. The results of the case study suggest that increasing the number of available ERTs from one to two decreases the response time by 64 percent, whereas an increase from two to three can provide only an additional 8 percent reduction. The reduction is mainly attributed to reduced dispatch time. Another significant observation is the fact that during heavy load, the nearest-neighbor dispatching policy provides better performance than the first-come, first-served policy. The proposed model is used by an electric utility company as a tool for determining the required number of ERTs and their service territories in such a way as to achieve predetermined service unavailability objectives.

Managers of emergency repair operations in electric utility companies frequently face decisions related to the optimum deployment of their emergency repair fleets. The major objective of emergency repair operations is that of minimizing service unavailability (*1*,*2*). Service unavailability can be decreased by reducing (a) the frequency of power interruptions, (b) the number of customers affected per interruption, and (c) the duration of the interruption. The first two approaches are related to the design properties and maintenance policies of the power distribution network. The third approach relates to the deployment of emergency repair trucks (ERTs).

This paper is motivated by the problem of deploying the ERTs of a large utility company. The main focus is the reduction of the duration of electric power interruptions through the optimum deployment of ERTs. In particular, this paper will address the issue of determining the optimum number of ERTs and their service territories in such a way as to achieve a predetermined threshold value of power restoration time, balance the workload of ERTs, and provide uniform level of service to customers.

K. G. Zografos, Department of Civil and Architectural Engineering, University of Miami, Coral Gables, Fla. 33124. C. Douligeris, Department of Electrical and Computer Engineering, University of Miami, Coral Gables, Fla. 33124. L. Chaoxi, Transportation Laboratory, University of Miami, Coral Gables, Fla. 33124.

PROPERTIES AND CHARACTERISTICS OF ERT DEPLOYMENT PROBLEM

The properties and characteristics of the ERT deployment problem reflect the operations of a large electric utility company. This utility provided to us data covering its ERT deployment operations from July 1, 1988, to July 1, 1989. Although the particular values of the data represent the operating characteristics of this company, the general emerging patterns are typical of the emergency repair operations of any large utility company in the United States (*3*).

The demand for emergency repair services is created from incidents causing power interruptions that occur randomly in time and space. When a power interruption occurs, one or more customers calls a service center to report service unavailability. On the basis of the customers' calls, a work order (ticket) is created by a computerized system, in this application called Trouble Call Management System (TCMS). This ticket is assigned to an emergency repair truck for further investigation and repair. The ERTs are mobile servers and at the time of dispatch can be located anywhere in a designated repair district.

The time elapsed between the arrival of a service call and the power restoration is called service restoration time (SRT). For analysis purposes the SRT is divided into the following components: ticket creation time (T_0), dispatch (T_1), travel time (T_2), and repair time (T_3).

T_0 is equal to the time interval between the placement of the call and the generation of the ticket. T_0 is controlled by the TCMS and is almost constant with an average value of about 10 min. T_0 does not depend on the spatial and temporal distribution of calls (*3*,*4*).

T_1 is equal to the elapsed time between the creation of a ticket and the assignment of the ticket to the first available ERT. T_1 depends on the workload assigned to each ERT.

T_2 is defined as the elapsed time between the ticket assignment and the arrival of the ERT at the scene of the incident. T_2 is a function of the shape and the size of the service area and the travel speed of the ERT (*3*,*5*).

T_3 is defined as the time interval between an ERT's arrival at the scene of the incident and the power restoration. The duration of the repair time depends on the type of incident, that is, severity of the problem, type of failing equipment, day versus night repair, adverse versus favorable weather conditions, and the training and expertise of the repair personnel (*3*).

The proposed model will consider the reduction of the duration of power interruptions by studying the effect of redistricting and ERT deployment policies on the reduction of T_1 and T_2.

In its general form, the ERT deployment problem can be defined as follows:

> Given the temporal, spatial, and priority distribution of power interruptions, define the number, area of responsibility location, and dispatching policy of emergency repair trucks so as to achieve a set of predetermined service objectives.

PREVIOUS RELATED WORK

The deployment of ERTs falls into a general category of dynamic vehicle routing problems that possess the following characteristics:

- Probabilistic demand over time and space,
- Probabilistic distribution of service times,
- Mobile servers, and
- Minimization of total system time (waiting, travel, and service time).

The problem of the optimum deployment of emergency response units has been extensively covered in the literature. Particular emphasis has been paid to the allocation and dispatching of police patrol units, fire engines, and ambulances in an urban environment (*5,6*).

Common to all these problems is the fact that demands vary stochastically, both temporally and spatially. Several methods have been used for solving this problem. We can categorize the models as queueing, travel time, geometric, and simulation models (*7*).

Queueing models have been developed to determine the number of units required to be on duty in order to achieve a threshold value of dispatch delay. In queueing models, calls for an emergency unit arriving at a service center are placed in queue and served on a first-come, first-served basis, or according to a priority system, or in more complex models according to the dispatching policy at hand—an example of which is a model's considering a minimum and a maximum number of units required to serve a specific call (*8,9*). Although these models provide a clear insight in the queueing phenomena of spatially distributed servers, they cannot be applied to solve the previously defined ERT deployment problem, because they assume predetermined districts and they require multiple-server dispatching and preventive patrolling.

Travel time models are more appropriate when the degree of congestion in the system is low and the travel time dominates the system response time (*7*). In the development of travel time models, the fact that units travel from a depot or from a place in the district to the place where they are needed is taken into account. In particular, the geography of the area, the travel speeds in the appropriate directions, and the location of the units are considered (*10,11*). Although these formulations are closely related to the ERT deployment problem, they cannot be applied directly for its solution, because they are limited by their first-come, first-served dispatching policy and the requirement that servers return to their depots after servicing the incident.

In geometric models the spatial distribution of the calls combined with the location of the units and the corresponding traveling speeds is taken into account to validate the empirical data (*12,13*). These models are not dynamic and stochastic in their formulation.

Simulation models have been used to study the effect of proposed administrative, organizational, and technological changes on the performance of emergency response systems. In simulation models a more realistic picture of the events that take place can be achieved since many factors can be taken into account and their effect on the performance of the system can be studied (*6*). Existing simulation models are problem-specific, and none of them has been developed in the context of emergency services in the electric utility industry.

A variety of objectives has been proposed for the optimum deployment of emergency service restoration units. Minimization of response time, equal workload for all units, uniform performance in all service territories, and minimum average delay have been used to evaluate the performance of various emergency response systems (*7*).

Essential to the optimal allocation of the emergency units is a districting method that considers the distribution of the calls for service in time and space (*15*).

The location of mobile servers within a service area is also a problem related to the design of emergency response systems. The location of the server affects its response time, which is the dominant component of the total system time under light workload conditions (*3,14*). Location models for the deployment of emergency vehicles have been proposed by Charnes and Storbeck (*16*), Daskin and Stern (*17*), Toregas et al. (*18*), and Brandeau et al. (*19*). These models deal only with the travel time aspects of the emergency response system, assuming that a service unit is always available to respond to a call in its area of responsibility. Location models considering the queueing aspects of emergency response systems have also been proposed in the literature (*20,21*).

Most of the existing models have been motivated by applications that have substantial organizational, operational, and technical differences from the emergency repair operation of electric utility companies. For instance, in the electric utility industry environment there is no preventive patrolling as in police operations (*8,9*), there is no multiple-vehicle dispatching as in fire and police operations (*5*), and the service unit does not necessarily return to its home base after an incident as in medical emergency systems and fire protection operations (*12*).

These differences preclude the wholesale application of existing emergency response models to the electric utility service restoration operations (*22*). Therefore, a model that reflects the service restoration operations of electric utility companies should be developed.

PROPOSED METHODOLOGICAL FRAMEWORK

The problem of the optimum deployment of ERTs is a complex resource allocation problem that presents serious ana-

lytical difficulties for its solution. For methodological reasons the problem can be decomposed into two interrelated nested subproblems: designating response areas (districting) and determining the number of ERTs required to be on duty in each district per shift.

Given the complexity and magnitude of real-world problems—large and dispersed geographical areas and the temporal, spatial, and priority distribution of calls—it is not possible to develop an exact optimization algorithm for the service restoration problem. Therefore, an iterative procedure combining optimization and computer simulation is proposed as a solution. The basic modules and the logic of the proposed method are shown in Figure 1.

The inputs of the proposed procedure are (a) the arrival rate of repair calls, (b) the travel time between geographic entities (atoms) of the study area, and (c) the number of ERTs that should be available per shift (N).

The initial module of the proposed methodology involves the solution of the districting problem. The objective of the districting problem is to partition the area under study into areas of primary responsibility, that is, service areas, so as to achieve some level or combination of levels of service. Analysis of the existing operations (*22*) led to the conclusion that the entire area should be partitioned into homogeneous service areas according to work load of the ERTs and area covered by each unit. The basic assumption behind the consideration of such criteria is that service areas with "similar" characteristics should have similar performance.

After the solution of the districting problem, the method proceeds by simulating the service restoration operations in each of the generated service areas. At this stage, a comparison of the performance of the service restoration operations is performed by comparing the simulated values with preestablished target values. If the performance of the system is not satisfactory, the number of ERTs (N) is increased by one and the entire process is repeated. The process stops when the preestablished threshold values are achieved.

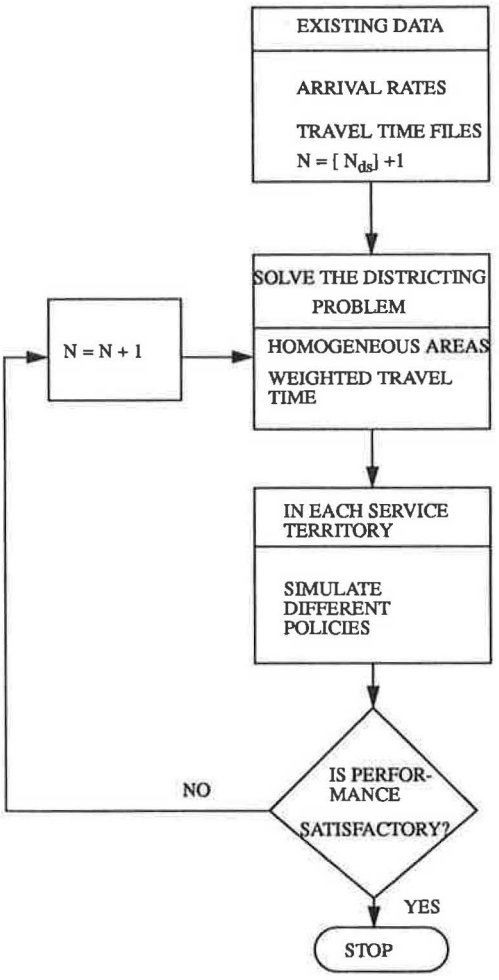

FIGURE 1 Proposed solution approach.

Model for Designing Emergency Repair Districts

In its general form, the districting problem can be expressed as follows:

> Given an area consisting of a number of elementary spatial units (atoms) with a given level of activity, determine nonoverlapping contiguous clusters of atoms (districts) that "optimize" a set of objectives.

The districting problem has been used extensively in the literature in order to design political districts (*15*,*23*,*24*), school districts (*25*–*27*), residential refuse collection districts (*28*), sales territories (*29*, p.469; *30*,*31*), and inspection and repair territories (*32*,*33*).

Two basic approaches have been used for the solution of the districting problem: implicit enumeration and heuristic algorithms. The first approach formulates the districting problem as a 0-1 optimization problem and uses an implicit enumeration technique for its solution (*15*). This method involves two stages: the generation of feasible districts and the selection of the optimum districting pattern. Although this method is mathematically rigorous, its computational burden is very high. The second approach has been suggested in the literature for the solution of large-scale districting problems (*22*,*32*,*34*). This approach is based on a generalized formulation of the transportation problem.

A modification of the latter approach is used for the mathematical formulation of the emergency repair districting problem. The following notation should be introduced before the mathematical presentation of the model:

N = number of centers or service territories;
M = number of atoms;
I = set of service territories $I = \{1,2,\ldots,N\}$;
J = set of all atoms $J = \{1,2\ldots,M\}$;
X_{ij} = workload of the jth atom assigned to the ith center;
t_{ij} = separation (travel time) of service territory center (i) from the center of atom j;
A_j = area of atom j;
AT_i = area of service territory i;
$\overline{A}$ = average area of a service territory;
P_j = workload of atom j;

PT_i = workload of service territory i;
$\overline{P}$ = average workload of a service territory;
λ_j = number of calls originating at atom j;
$\overline{S}_j$ = average repair time of calls originating at atom j;
α_1 = maximum allowable percentage deviation of workload of a service territory from the average service territory workload $\overline{P}$, $0 \leq \alpha_1 \leq 1$;
α_2 = maximum allowable percentage deviation of the area of service territory from the average area of a service territory $\overline{A}$, $0 \leq \alpha_2 \leq 1$; and
m_{ij} = M_0 (very large number) if atom j belongs to an enclave and service territory i is not the neighboring area of the enclave, and 1 otherwise.

Since the following relations hold,

$$P_j = \lambda_j \overline{S}_j$$

$$PT_i = \sum_{j \in J} X_{ij}$$

$$\overline{P} = \sum_{i=1}^{N} \frac{PT_i}{N}$$

$$AT_i = \sum_{j=1}^{M} A_j \frac{X_{ij}}{P_j}$$

$$\overline{A} = \sum_{i=1}^{N} \frac{AT_i}{N}$$

the mathematical expression of the model can be written as follows:

Minimize

$$F = \sum_{i \in I} \sum_{j \in J} X_{ij} t_{ij} \tag{1}$$

Such that

$$\sum_{i \in 1} X_{ij} = P_j \qquad j \in J \tag{2}$$

$$(1 - \alpha_1)\overline{P} \leq \sum_{j \in J} X_{ij} \leq (1 + \alpha_1)\overline{P} \qquad i \in I \tag{3}$$

$$(1 - \alpha_2)\overline{A} \leq AT_i \leq (1 + \alpha_2)\overline{A} \qquad i \in I \tag{4}$$

$$X_{ij} \geq 0 \tag{5}$$

The districting module uses a static and aggregate measure of workload and assumes that this workload is concentrated at the centroid of each atom. The shortest travel time t_{ij} is used as a separation measure between the center of a service territory i and the center of an atom j. The demand of an atom is calculated as the product of the number of repair calls λ_j and the average repair time of each call.

Equation 1 is the objective function of the model, and it expresses the minimization of the weighted travel time. Equation 2 requires that the summation of the demand originated from an atom j and assigned to all centers i is equal to the total demand generated at atom j. Equation 3 suggests that the total workload assigned to each center should be within a given threshold value (α_1 percent) from the average workload of the entire region. Equation 4 requires that the size of each service territory should be within a given threshold value (α_2 percent) from the average. Finally, Equation 5 gives the nonnegativity requirements for the workload.

The solution of the problem described by Equations 1 through 5 provides the assignment of atoms to the center of each territory. At this point it should be noted that the territories generated by the solution of Problems 1 through 5 are based on the initial selection of the centers. A heuristic iterative algorithm based on the procedures proposed by Maranzana (*34*) is used for the solution of the problems. This algorithm is summarized as follows:

Step 1

Determine the number of required service territories (N). If this is the initial iteration of the algorithm, then (N) is determined through the minimum number of required service territories (N_{ds}) corresponding to service day (d) and shift (s). Otherwise, the number (N) used in the previous iteration of the algorithm is increased by one (i.e., $N = N + 1$).

Step 2

Select an initial set of N atoms to be the centers of the N service areas.

Step 3

Use the formulation of the linear program 1–5 to find assignments of atoms to service territory centers. An atom may be split into more than one part, and each part may be assigned to a different service territory.

Step 4

Within each service territory that results from Step 3, find the center that minimizes the weighted travel cost. The following formulas can be used to determine a possible initial set of coordinates of the adjusted centers $(\overline{x}_i, \overline{y}_i)$ (note that $(x_i^{\text{ref}}, y_i^{\text{ref}})$ is the reference point for calculations of distances in service territory i):

$$\overline{x}_i = \frac{\sum_{j \in J} |x_i^{\text{ref}} - x_j| X_{ij}}{\sum_{j \in J} X_{ij}} \quad \forall i \in I \tag{6}$$

$$\overline{y}_i = \frac{\sum_{j \in J} |y_i^{\text{ref}} - y_j| X_{ij}}{\sum_{j \in J} X_{ij}} \quad \forall i \in I \tag{7}$$

Given the fact that the $(\overline{x}_i, \overline{y}_i)$ coordinates calculated by Equations 6 and 7 are optimum for the Euclidean distance metric and not for the Manhattan metric (movement is only allowed

on the x- and y-axes) used here for the calculation of the travel matrix, a search procedure is employed to determine the optimum location of the centers for the next iteration of the algorithm (*35*). The search procedure starts from the best between the previous center and $(\bar{x}_i, \bar{y}_i)$ and stops when no better center can be found (*22*).

Step 5

Check if there is an enclave. If an enclave exists solve Problem 2–5 with the following objective function:

Minimize

$$F = \sum_{i \in I} \sum_{j \in J} x_{ij} t_{ij} m_{ij} \qquad (1')$$

and go to Step 3. If an enclave does not exist, go to Step 6.

Step 6

For each service territory center i, check if the summation of the difference of the x- and y-coordinates between two successive iterations differs more than a small value ε_1 and the positions of all the centers between two successive iterations differs more than a small value ε_2. If yes: go to Step 3, otherwise stop.

Apparently, Steps 1 through 6 describe a heuristic algorithm, since the centers of the districts are not known in advance. A FORTRAN code has been written for the implementation of the described algorithm (*22*). The code uses the IMSL library for the solution of the linear program 1–5 in Step 3.

After the districting problem has been solved for a given number of service areas, the performance of the emergency repair operations within each generated district is evaluated using the simulation procedure described in the next subsection.

Simulation of Emergency Repair Operations

The proposed simulation module simulates the emergency repair operations within any service territory designed by the districting model. After the calls have been generated, the ERTs are dispatched to the locations of the incidents. At this stage of the simulation module, each truck can serve only its own territory, and there is no interdispatching allowed between truck areas. Even though this might appear to be a constraint, it has the main benefit of the truck driver's familiarity with the underlying transportation network and failing equipment that results in reduced travel and repair times.

Two alternative dispatching policies are simulated by this module. The first policy uses a first-come, first-served dispatching rule, and the second policy uses a nearest-neighbor (NN) dispatching rule.

CASE STUDY

A case study related to the emergency repair operations of a large electric utility company is presented to demonstrate the applicability of the proposed methodology. The service area of the case study covers a geographic area of approximately 65 mi^2. Workload data for the area under consideration were obtained from the TCMS data base for a 1-year period and for the operations of the second shift (3:00 p.m.–11:00 p.m.). Atoms having an area of 1 mi^2 were used as the unit of analysis.

Given the Manhattan-like structure of the underlying transportation network, the Manhattan metric was used to calculate the distances between points of interest in the study area. Travel speed data were obtained through personal interviews with ERT operators.

The proposed methodological framework was applied to evaluate the performance of the emergency repair operations for three numbers of ERTs and two dispatching policies.

Concerning the balancing of the workload and area, values of $\alpha_1 = 0.10$ and $\alpha_2 = 0.20$ were used. Complete balancing of area and workload is almost always impossible. Tighter balancing of the area frequently leads to enclaves and irregular shapes of the service territories.

The results of the case study are presented in graphical form in Figures 2 and 3 for two and three ERTs, respectively. Figure 4 shows the existing partition of the study area for $N = 3$. Each square of the background grid is 1 mi^2. The numbers in the axes relate to the longitude and latitude values of the atoms as they are used by the company under study. When a single ERT was considered, the entire area constituted the service territory.

The actual percentage deviation from the average workload and area and the associated workload and area for each scenario and the existing partitioning are summarized in Table 1. The improvements compared with the existing partition can also be seen.

The results of the districting scenario were used as inputs to the simulation module (i.e., the polygons describing each service territory). Two alternative dispatching policies were simulated for the three districting scenarios and the current configuration. Four hundred days of operation were simulated for each scenario. The results of these simulations are shown in Figures 5 and 6.

CONCLUDING REMARKS

An integrated methodological framework for the optimum deployment of emergency repair fleets in the electric utility

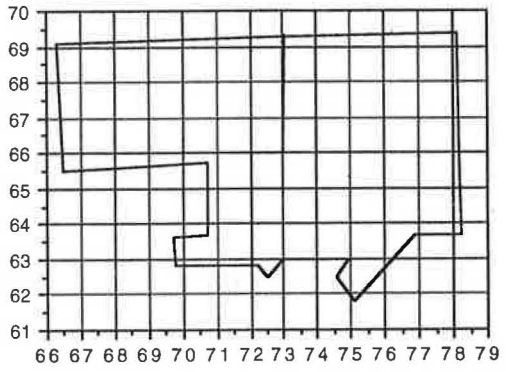

FIGURE 2 Service territories for two ERTs.

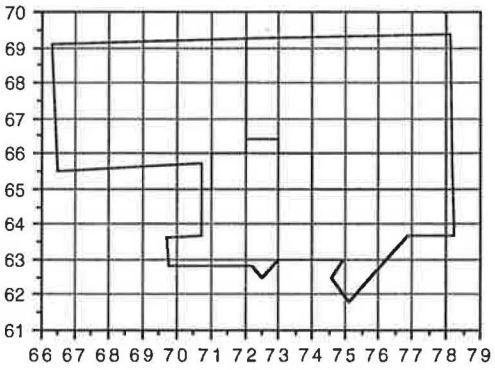

FIGURE 3 Service territories for three ERTs.

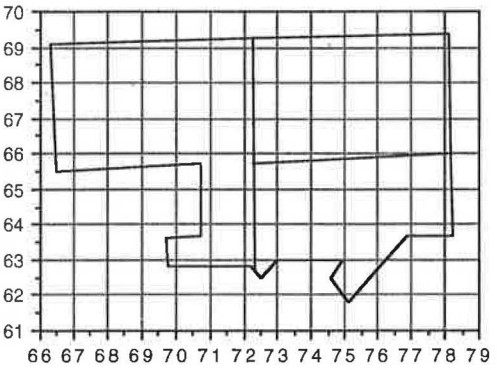

FIGURE 4 Existing service territories.

TABLE 1 Results of Districting

	Service Territory 1		Service Territory 2		Service Territory 3			
	α_1	α_2	α_1	α_2	α_1	α_2	$\overline{P}$	$\overline{A}$
N=2	-0.02	-0.14	0.02	0.14			16687	32.5
N=3	0.03	-0.07	-0.10	0.14	0.07	-0.07	11091	21.6
Existing	0.44	0.20	-0.41	0.01	-0.03	-0.21	11091	21.6

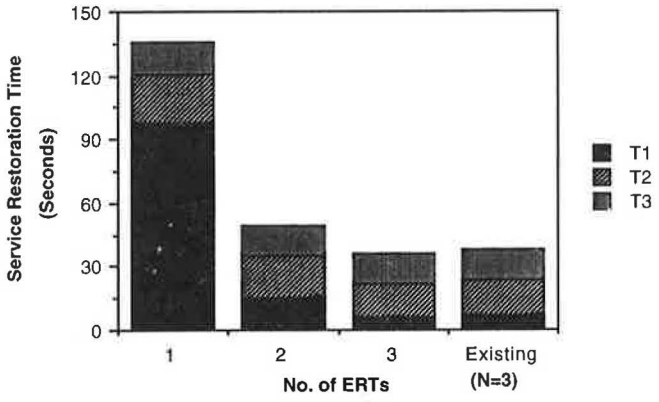

FIGURE 5 Results of simulation by first-come, first-served policy.

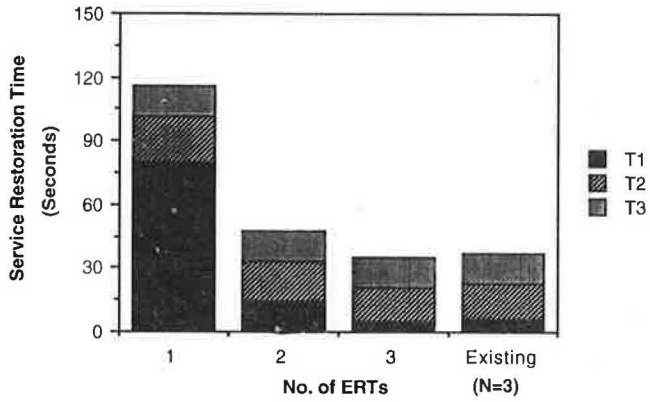

FIGURE 6 Results of simulation by nearest-neighbor policy.

industry has been presented. The proposed model takes into account the structural, operational, and technical characteristics of the emergency repair services of the electric utility industry and expands on work done in the area of emergency response operations.

A real-world case study was used to illustrate the applicability of the proposed methodology. The results of the case study suggest that the service restoration time relies heavily on the number of available ERTs. A decrease in service restoration time of about 64 percent was observed when the number of ERTs was increased from one to two. The decrease was substantial but not as significant when the number of ERTs was increased from two to three. Therefore, the marginal benefit of adding more servers diminishes after the second server. Most of the reduction observed in the service restoration time was due to the reduction of the dispatch time component.

Another interesting observation of the derived results is the difference in the performance of the ERTs between the two dispatching policies. In particular, under heavy workload (see Figures 5 and 6 for $N = 1$) the nearest-neighbor policy yields better performance of the system in terms of service restoration time than the first-come, first-served policy. This result is in agreement with recent analytical results dealing with the solution of the dynamic traveling repairman problem with rectangle service territories and uniform load distribution (*14*). Under light load (see Figures 5 and 6 for $N = 2$ and $N = 3$) the differences are not substantial.

The comparison of the existing partition with the proposed partition for $N = 3$ shows more balance in terms of area and workload partition. The generation of more homogeneous in terms of workload and area of responsibility truck areas has resulted in better system performance without increasing the number of required servers. The proposed partition results in a reduction of 2.5 min in the response time, which is attributed to the reduction of the average travel time due to the balancing of the areas and the consequent reduction of the dispatch time. Better improvements than the reported ones will be observed during periods of high demand since this is the time queues build up and balancing the workload has a profound impact on the performance of the system. This balancing also contributes to better working relationships between the truck operators and provides better response times especially under heavy load conditions.

With regards to the models implementation we can say that the proposed model provided the utility company with a tool for the efficient deployment of the fleet of service restoration trucks. Given the fact that the designation of service territories is a strategic decision for the utility companies that is established for a period of at least 5 years, computational requirements of the problem are not a burden for its use and implementation. Stated otherwise, designating the service restoration territories does not require real-time decision making and consequently the computational requirements of the districting model do not make the application of the model impractical. As a final note we would like to point out that during the final implementation of the proposed methodology, the results provided by the districting model were slightly modified with the help of the system operators to produce shapes of districts that reflect the real world operational conditions and constraints.

ACKNOWLEDGMENTS

This work has been supported by the Florida Power and Light Company. The authors would like to acknowledge with thanks the assistance of Jeff Mitchell, John Haupt, and Mannie Miranda.

REFERENCES

1. D. Perlstein. Automatic Vehicle Location Systems: A Tool for Computer Aided Dispatch Systems of the Future. *Proc., 1st Vehicle Navigation and Information Systems Conference*, Toronto, Canada, 1989, pp. 186–193.
2. K. G. Zografos, C. Douligeris, L. Chaoxi, and P. Tsoumpas. *Modeling Issues Related to the Service Restoration Problem*. Technical Memorandum CEN-90-2. Department of Civil and Architectural Engineering, University of Miami, Coral Gables, Fla., 1990.
3. K. G. Zografos, P. Tsoumpas,and C. Douligeris. *An Evaluation of the FP&L's Service Restoration Operations: The Experience of the Southern Division*. Technical Memorandum CEN-90-1. Department of Civil and Architectural Engineering, University of Miami, Coral Gables, Fla. 1990.
4. K. G. Zografos, C. Douligeris, J. Haupt, and J. Jordan. A Methodological Framework for Evaluating On-Board Computer Technology in Emergency Dispatch Operations. *Proc., International Conference on Vehicle Navigation and Information Systems*, Dearborn, Mich., Oct. 1991.
5. R. C. Larson. A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research*, Vol. 1, 1974, pp. 67–95.
6. J. Chaiken. *Estimating Numbers of Fire Engines Needed in New York City Fire Divisions*. Report R-508-NYC. Rand Institute, New York, N.Y., 1971.
7. J. M. Chaiken and R. C. Larson. Methods for Allocating Urban Engineering Units—A Survey. *Management Science*, Vol. 19, No. 4, Part 4, 1972, pp. 110–130.
8. L. Green. A Multiple Dispatch Queueing Model of Police Patrol Operations. *Management Science*, Vol. 30, No. 6, 1984, pp. 653–664.
9. L. Green and P. Kolesar. A Comparison of the Multiple Dispatch and M/M/C Priority Queueing Models for Police Patrol. *Management Science*, Vol. 30, No. 6, 1984, pp. 665–670.
10. O. Berman and R. Larson. The Median Problem with Congestion. *Computers and Operations Research*, Vol. 9, No. 2, 1982, pp. 119–126.
11. O. Berman and A. Odoni. Location Mobile Servers on a Network with Markovian Properties. *Networks*, Vol. 12, 1982, pp. 73–86.
12. J. M. Hogg. The Siting of Fire Stations. *Operational Research Quarterly*, Vol. 19, 1968, pp. 275–287.
13. F. Mannering and W. Kilareski. The Common Structure of Geo-Based Data for Roadway Information Systems. *ITE Journal*, 1986, pp. 43–49.
14. D. Bertsimas. *The Probabilistic Vehicle Routing Problem*. Working Paper 2067-88. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 1988.
15. R. S. Garfinkel and G. L. Nemhauser. Optimal Political Districting by Implicit Enumeration Techniques. *Management Science*, Vol. 16, No. 8, 1970, pp. 495–508.
16. A. Charnes and J. Storbeck. A Goal Programming Model for the Siting of Multilevel EMS Systems. *Socio-Economic Planning Science*, Vol. 14, 1980, pp. 166–161.
17. M. Daskin and E. Stern. A Hierarchical Objective Set Covering Model for Emergency Medical Service Deployment. *Transportation Science*, Vol. 15, No. 2, 1981, pp. 137–152.
18. C. Toregas, C. Revelle, R. Swain, and L. Bergman. The Location of Emergency Service Facilities. *Operations Research*, Vol. 12, 1971, pp. 1366–1373.
19. M. L. Brandeau and R. C. Larson. Extending and Applying the Hypercube Queueing Model to Deploy Ambulances in Boston. *TIMS Studies in Management Sciences*, Vol. 22, 1986, pp. 121–153.
20. R. Batta, R. C. Larson, and A. R. Odoni. A Single Server Priority Queueing Location Model. *Networks*, Vol. 18, 1988, pp. 87–103.
21. S. Chiu, O. Berman, and R. Larson. Location a Mobile Server Queueing Facility on a Tree Network. *Management Science*, Vol. 31, No. 6, 1985, pp. 764–772.
22. K. G. Zografos, C. Douligeris, and L. Chaoxi. *The Solution of the Districting Problem*. Technical Memorandum CEN-91-1. Department of Civil and Architectural Engineering, University of Miami, Coral Gables, Fla. 1991.
23. I. M. L. Robertson. The Delimitation of Local Government Electoral Areas in Scotland: A Semi-Automated Approach. *Journal of the Operational Research Society*, Vol. 33, 1982, pp. 517–525.
24. B. Nygreen. European Assembly Constituents for Wales—Comparing of Methods for Solving a Political Districting Problem. *Mathematical Programming*, Vol. 42, 1988, pp. 159–169.
25. C. A. Holloway, D. A. Wehrung, M. P. Zeitlin, and R. T. Nelson. An Interactive Procedure for the School Boundary Problem with Declining Enrollment. *Operations Research*, Vol. 23, No. 2, 1975, pp. 191–206.
26. J. Bovet. Simple Heuristics for the School Assignment Problem. *Journal of the Operational Research Society*, Vol. 33, 1982, pp. 695–703.
27. J. A. Ferland and G. Guennette. Decision Support System for the School Districting Problem. *Operations Research*, Vol. 38, No. 1, 1990, pp. 15–21.
28. G. N. Berlin. Method for Delineating Districts of Varying Shape. *Transportation Engineering Journal*, ASCE, Vol. 102, No. TE4, 1976, pp. 805–819.
29. J. B. Clooman. A Note on the Compactness of Sales Territories. *Management Science*, Vol. 19, Part I, 1977.
30. S. W. Hess and S. A. Samuels. Experiences with a Sales Districting Model: Criteria and Implication. *Management Science*, Vol. 18, Part 2, 1971, pp. 41–54.
31. R. G. Sharker et al. Sales Territory Design: An Integrated Approach. *Management Science*, Vol. 22, 1975, pp. 309–320.
32. P. G. Marlin. Application of the Transportation Model to a Large-Scale Districting Problem. *Computers and Operations Research*, Vol. 8, 1981, pp. 83–96.
33. M. Segal and D. B. Weinberger. Turfing. *Operations Research*, Vol. 25, No. 3, 1977, pp. 367–386.
34. F. Maranzana. On the Location of Supply Points for Minimizing Transport Cost. *Operations Research Quarterly*, Vol. 15, 1964.
35. R. F. Love, J. G. Morris, and G. O. Wesolowsky. *Facilities Location: Models and Methods*, North Holland, New York, N.Y., 1988.
36. K. G. Zografos, C. Douligeris, and L. Chaoxi. *A Model for the Optimum Deployment of Service Restoration Units: Phase 1 Report*. Technical Memorandum CEN-90-4. Department of Civil and Architectural Engineering, University of Miami, Coral Gables, Fla.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Attribute Importance in Supply of Aeromedical Service

MARK R. MCCORD, OSCAR FRANZESE, AND XIAO DUAN SUN

Multiattribute utility theory is used to investigate a supplier's value of offering aeromedical service. Using joint probability functions over net revenue, publicity, and medical benefit dimensions to capture the operating performance of the service and a multiattribute utility function with random parameters to capture the supplier's preferences, it is found that providing service is preferred to shutting down the program for all of the 1,000 sets of utility parameters generated. Using the analysis developed, however, it is evident that no one dimension is sufficient to justify service when the cost of providing the service is considered. The revenue dimension comes closest, but the roughly 50 percent chance of suffering financial losses and the strong aversion to these losses lead to the conclusion that revenues alone are not sufficient to continue operations. When using the analysis to look at pairs of dimensions, it appears that the revenue-medical benefit pair is sufficient to justify service for fewer than half of the 1,000 sets of utility parameters and that the publicity dimension is extremely important in motivating the supplier to provide service. The results are interpreted to form a working hypothesis that suppliers must either believe that flying emergency missions provides important publicity value to the sponsoring hospitals or be ensured of better financial security if they are to continue to provide this emergency medical service.

The number of aeromedical programs has been increasing since after the Korean War; in 1989 there were 200 programs in the United States (*1*). Aeromedical programs receive requests for immediate service, and aircraft fly medical crews from bases (usually hospitals or airports) to patients at accident scenes or, more frequently, at outlying hospitals with inadequate medical facilities. The patients are then secured in the aircraft and flown to hospitals in larger cities. Emergency care is provided by the medical flight team on the return flight, and speed in reaching the patients and flying them to the destination hospital is critical.

According to a recent study, Columbus, Ohio, ranked in the top five U.S. cities in terms of numbers of requests for emergency aeromedical service (*2*, p.12). Two helicopter-based programs serve this area: Skymed, which is a consortium of the Ohio State University hospitals and Children's Hospital in Columbus, and Lifeflight, an older program based at Grant Hospital in Columbus. These programs simultaneously compete and cooperate. They compete for "discretionary" patients who will be flown to the university hospitals if Skymed transports them and to Grant Hospital if Lifeflight transports them. In times when hospitals are competing for patients to generate the associated revenues, having access to these discretionary patients is thought to be important. The programs cooperate in that some patients must go to specific hospitals (e.g., burn patients are treated almost exclusively at university hospitals), and the program contacted first will transport such patients to the necessary hospital, even if it is the sponsoring hospital of the other program. In return, the transporting program receives only the flight revenues collected, which are generally small compared with the net inpatient revenue. The environment is also cooperative in that if one program receives a request that it cannot serve because its helicopters are flying other missions, it turns the request over to the competing program.

The authors have presented a multiobjective analysis designed to help the director of Skymed determine the desirability of leasing additional helicopters to expand the fleet size of his program (*3*). Currently, Skymed operates one helicopter 24 hr/day. The results of the analysis, which was based on multiattribute utility theory (*4–6*), showed that expansion of fleet size was warranted, according to the primary dimensions of performance supplied: net revenues, publicity to the sponsoring hospitals, and medical benefits. Some novel features developed there showed, however, that the increased value associated with these performance dimensions was small enough that less technical aspects might govern the ultimate decision, and we could not argue strongly for expansion from one full-time helicopter.

In this paper, we look more closely at the individual performance dimensions in the context of deciding whether to supply aeromedical service or not. Later, it is shown that the probability distributions over the multicommodity bundles of revenues, publicity, and medical benefits are such that existing service is preferred to the distribution associated with shutting down the program. Then it is asked whether any one of the dimensions taken by itself is enough to offset the costs associated with supplying service; the answer is no. It is seen that, although the medical benefits are considered important, neither the medical-revenue pair nor the medical-publicity pair can justify service when the costs of producing these benefits are considered. The implication is that cost recovery and publicity are critical to providing service.

To make the exposition easier, we shall sometimes talk about the results as if they are general, but we recognize that they are based on one helicopter program. They are also based on the preferences of one individual (the executive director), who, in a slightly different application, was faced with preparing an analysis for the true decision makers of the problem. Nevertheless, as we shall argue, these results can serve as a benchmark for, if not a representative sample of, the industry, and we discuss the implications of our findings there. Before presenting the new study, however, we review the data on which it is based.

Civil Engineering Department, Ohio State University, 2070 Neil Avenue, Columbus, Ohio 43210.

EVALUATION MODEL

The evaluation model is the expected utility model (*3,4,6,7*). In this procedure the analyst first determines a multiattribute vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ of the important measures or outcomes by which the alternatives are to be evaluated. A utility function $U(\mathbf{Y})$ is constructed which maps **Y**-vectors into real numbers between 0 and 1 representing preferences on an interval scale. Then, each alternative Xi is mapped into a probability mass function $P(\mathbf{Y}/Xi)$. (Although using continuous probability density functions is completely analogous, we use a discrete approximation to these functions to simplify exposition.) This gives the probabilities that the different combinations of **Y**-vectors will result if alternative Xi is implemented. Valuation, then, is over the probability distribution $P(\mathbf{Y}/Xi)$, and the axioms of the theory imply that the distribution should be valued according to the mathematical expectation of the utilities of the attribute vectors. Letting EU_i represent this expected utility valuation associated with alternative Xi, we have

$$EU_i = \sum_{\text{all } j} p[\mathbf{Y}(j)/Xi] * U[\mathbf{Y}(j)] \quad (1)$$

where the sum is taken over all possible attribute vectors j, given that alternative Xi is implemented, and $p[\mathbf{Y}(j)/Xi]$ is the probability that vector $\mathbf{Y}(j)$ obtains when Xi is implemented. This approach is arguably the most popular for multiattribute evaluation when uncertainties are present because of the behaviorally appealing assumptions on which it is based, the consistency with the assumptions of the methods designed to obtain probability and utility functions, and the use of performing the expectation operation in Equation 1 and determining the functions (*4–7*).

The attribute vector **Y** was determined, and the probability distributions $P(\mathbf{Y}/X)$ and utility functions $U(\mathbf{Y})$ were estimated for a study to assist Skymed in evaluating the desirability of adding helicopters to its fleet (*3*). We will briefly review the aspects that will be needed for the present study.

Attribute Vector

We had several discussions with the executive director of Skymed and reviewed the aeromedical literature over the course of several months while defining the problem and developing analytical models. During this time we iterated several times on the definition of our attribute vector **Y**. We eventually decided that although there were other considerations to the problem of deciding whether to add helicopters to the fleet (such as public perceptions and availability of funds in hospital programs), aeromedical operations primarily offer three types of benefit: (a) they produce revenues, from the flight charges and from the inpatient bills collected after treating a patient flown to the sponsoring hospital; (b) they offer publicity to the sponsoring hospital, which is important in times when there are empty beds and hospitals compete for patients; and (c) they provide medical benefits to the public by saving lives and decreasing morbidity. We also had to acknowledge that, since Lifeflight offers service in the area and since certain types of patients (e.g., burn patients) are flown to university hospitals even if Lifeflight transports them, some of the inpatient revenues associated with missions flown by Skymed would have been collected even if Skymed did not exist. Similarly, some of the medical benefits achieved by Skymed's transporting a patient would be achieved if Skymed shut down and the patients could be flown by Lifeflight. We handled these considerations by defining these attributes as the extra levels offered, where "extra" was taken to mean the amount of the attribute that would be generated if Skymed were to operate with a given number of helicopters in its fleet, minus that which would be generated if Skymed were to offer no service. On the other hand, the publicity associated with Skymed's operations would not be generated if Skymed did not operate, and we did not have to worry about the extra contribution in this case.

The costs associated with operations were primarily financial; they included the overhead costs of running the program, the fixed costs of leasing and staffing the helicopters, and the variable costs associated with flying the helicopter. They all would be avoided if Skymed did not exist and could, therefore, be subtracted from the extra gross revenues produced to form the extra net revenue generated. After many iterations, we decided that the number of patients flown by Skymed would serve as the best proxy variable to quantify the publicity dimension. We based our quantification of medical benefits in large part on pragmatic considerations of data availability. From an empirical study (*8*), we modeled the probability of helicopter transportation as being "essential" to a random helicopter patient's favorable outcome (probability .16), "helpful" to a random helicopter patient's favorable outcome (probability .10), or not contributing to the health state of a random helicopter patient (probability .74). From the descriptions of the categories (*8*) and discussions with the program director, we modeled the "essential" category as one in which the patient's life was saved because of the helicopter transport and the "helpful" category as one in which the patient would not have died without the transport but would have avoided a very bad outcome, such as losing a limb. We therefore quantified the medical attribute as a two-dimensional vector consisting of "extra lives saved" and "extra limbs saved."

In summary, then, our attribute vector **Y** was

$$\mathbf{Y} = [Y_1, Y_2, (Y_{3a}, Y_{3b})] \quad (2)$$

where

- Y_1 = extra net revenue to the university hospitals generated from Skymed's operations,
- Y_2 = number of patients carried by Skymed,
- Y_{3a} = extra number of lives saved from Skymed's operations, and
- Y_{3b} = extra number of debilitating injuries, such as losing a limb, avoided because of Skymed's operations.

Utility Function

We used the standard multiplicative utility function (*4,6*) to model preferences (*3*). Actually, as described there, we were helping the executive director synthesize the merits of the various expansion alternatives so that he could argue these

alternatives to the ultimate decision-making board. Therefore, he had to put himself in his boss's position when thinking through the preferences. Although he had no difficulty in doing this, the utility functions described in this section were elicited not from the decision makers directly responsible for the supply of service, but from their proxy. We believe these functions to be good first-order estimates of the ultimate decision-making board's preferences, however, especially because we model the parameters of these distributions as random variables. The multiplicative utility function $U_{\mathbf{Y}}$ can be written

$$U_{\mathbf{Y}}[Y_1, Y_2, (Y_{3a}, Y_{3b})] = k_1u_1 + k_2u_2 + k_3u_3 + K(k_1k_2u_1u_2 + k_1k_3u_1u_3 + k_2k_3u_2u_3) + K^2k_1k_2k_3u_1u_2u_3 \quad (3)$$

where

u_1, u_2, u_3 = unidimensional utilities associated with attribute levels $Y_1 = y_1$, $Y_2 = y_2$, and $(Y_{3a} = y_{3a}, Y_{3b} = y_{3b}$, respectively [i.e., $u_1 = u_1(y_1)$, $u_2 = u_2(y_2)$, $u_3 = u_3(y_{3a}, y_{3b})$];

k_1, k_2, k_3 = scaling constants, or "weights," of the financial, publicity, and medical dimensions, respectively; and

K = overall constant that ensures compatibility of the scales of individual dimensions with the multidimensional scale.

This multiplicative form can be shown to follow from certain properties of preferences, which appear to hold in many cases (*4,9*) and were shown to hold in our problem (*3*).

The utility function over the medical dimension u_3 was itself a two-attribute additive function:

$$u_3(y_{3a}, y_{3b}) = k_au_a(y_{3a}) + (1 - k_a)u_b(y_{3b}) \quad (4)$$

This additive function is a special case of the multiplicative function, and follows from testable behavioral properties (*4,6*) that were found to hold in our problem (*3*).

The unidimensional utility functions $u_i(y_i)$ are scaled so that the least and most preferred attribute levels considered—call them y_i^l and y_i^m, respectively—have utilities 0.0 and 1.0 (i.e., $u_i(y_i^l) = 0.0; u_i(y_i^m) = 1.0; i = 1, 2, 3a, 3b$). Given this scaling, one can see from Equations 3 and 4 that $u_{\mathbf{Y}}[y_1^m, y_2^l, (y_{3a}^l, y_{3b}^l)] = k_1$; $u_{\mathbf{Y}}[y_1^l, y_2^m, (y_{3a}^l, y_{3b}^l)] = k_2$; $u_{\mathbf{Y}}[y_1^l, y_2^l, (y_{3a}^m, y_{3b}^m)] = k_3$; and from Equation 4 that $u_3(y_{3a}^m, y_{3b}^l) = k_a$. That is, the scaling constant k_i represents the utility of the multiattribute vector $\mathbf{Y}$, when all attributes are at their least preferred levels, except attribute Yi, which is at its most preferred level. In this way, the scaling constants have a meaning consistent with the underlying theory, and this interpretation leads to operational ways of determining these values (*3,4*).

The single-attribute utility functions can also be determined from methods compatible with the underlying expected utility theory (*4,6,7*). We approximated the functions with the following specifications (*3*):

$$u_1(y_1) = a_1\left[\frac{y_1 + 6(10^6)}{6(10^6)}\right]^{b_1} \qquad -6(10^6) < y_1 \leq 0 \quad (5)$$

$$a_1 \qquad y_1 = 0 \quad (6)$$

$$a_1 + \frac{(1 - a_1)}{9(10^6)}y_1 \qquad 0 \leq y_1 \leq 9(10^6) \quad (7)$$

with

$$b_1 > 1.0 \quad (8)$$

$$0.7 < a_1 < 1.0 \quad (9)$$

$$u_2(y_2) = \left(\frac{y_2}{3{,}500}\right)^{b_2} \qquad 0 \leq y_2 \leq 3{,}500 \quad (10)$$

with

$$0 < b_2 < 1 \quad (11)$$

$$u_a(y_{3a}) = \frac{y_{3a}}{500} \qquad 0 \leq y_{3a} \leq 500 \quad (12)$$

$$u_b(y_{3b}) = \frac{y_{3b}}{300} \qquad 0 \leq y_{3b} \leq 300 \quad (13)$$

Because we had to determine least and most preferred levels for the attributes before completing our estimations of the probability mass functions, and because we wanted to allow for other options than those considered before (*3*), the least preferred level for the financial attribute [$-6(10^6)$\$] was lower than necessary, and the most preferred levels for the financial attribute [$+9(10^6)$\$], the publicity attribute (3,500 patients carried), and the medical subattributes (500 lives and 300 limbs saved) were higher than necessary. These "loose bounds" pose no problem, however, because the value of the scaling parameters k_i will vary with ranges of attributes used. This dependence on the attribute range is one reason that the scaling constants cannot be used by themselves to indicate attribute importance (*6*). (The least preferred levels of y_2, y_{3a}, and y_{3b} had natural levels of 0.)

The functional forms and constraints on the parameters also represented behavioral properties worthy of mention. The parameter a_1 represents the utility of breaking even in net revenue—that is, $U_1(0) = a_1$—relative to the utilities of \$6 million and +\$9 million. Constraint 9 represents that breaking even is extremely important. When it is combined with Constraints 5 through 8, one can see a strong decrease in marginal utility du_1/dy_1 once the break-even point is reached. Similarly, the convexity of u_1 in the net losses domain ($d^2u_1/dy_1^2 > 0$ for $y_1 < 0$) represents the increased importance of getting to zero, and the linearity in the net gains domain ($d^2u_1/dy_1^2 = 0$ for $y_1 > 0$) represents the constant marginal utility associated with increased revenues once the operation breaks even. These conditions came out of our general discussions of the relative value for revenues and were reflected in detailed and carefully designed utility assessments (*3*).

In determining the utility function for publicity, as quantified by the number of patients carried, we were careful to emphasize that the revenues and medical benefits were held constant, so that these indirect impacts of patients carried were not being valued in this function. Considering Equations 10 and

11, one notices decreasing marginal utility ($d^2u_2/dy_2^2 < 0$) for the number of patients carried, because this attribute contributes to the publicity dimension.

The linear utility functions ($d^2u_{3a}/dy_{3a}^2 = 0$) for lives saved represents that saving an extra life when 499 people, for example, have already been saved is just as important as when no one has been saved, and similarly for limbs saved. As described earlier (*3*), we were careful to frame the utility question to concentrate on medical benefits to the general public and to avoid valuing here the positive publicity associated with saving lives or limbs.

Once the values of the utility parameters are given, the utility functions are completely specified. We summarize these parameters as a vector $\boldsymbol{\Theta} = [a_1, b_1, b_2, k_1, k_2, k_3, k_a, K]$. From the literature (*10–14*) and extensive experience with preference modeling, we knew we could not get exact values for these parameters. We therefore used a variety of methods to determine bounds on the parameter values and modeled a_1, b_2, k_1, k_2, k_3, and k_a as being random variables from a triangular distribution. That is, denoting any of these parameters by Θ_i, the lower and upper bounds of the distributions by LB_i and UB_i, respectively, and the mode of the distribution by $\hat{\Theta}_i$, we had

$$f_{\Theta i}(\Theta_i) = 0 \qquad \Theta_i < LB_i$$

$$\frac{2(\Theta_i - LB_i)}{(UB_i - LB_i)(\hat{\Theta}_i - LB_i)'} \qquad LB_i \le \Theta_i \le \hat{\Theta}_i$$

$$\frac{2(UB_i - \Theta_i)}{(UB_i - LB_i)(UB_i - \hat{\Theta}_i)'} \qquad \hat{\Theta}_i \le \Theta_i \le UB_i$$

$$0 \qquad \Theta_i > UB_i \qquad (14)$$

where $f(.)$ represents the probability density of obtaining level Θ_i. We did not model b_1 as random, since the relative shape of the utility function in the net losses dimension would depend on a_1, which was modeled as a random variable. Also, the overall scaling constant K is determined from the values of the individual scaling constants k_1, k_2, k_3 (*4,6*) and did not need to be modeled explicitly. The parameter values of the distributions are given in Table 1.

TABLE 1 Parameter Value of Probability Distributions for Utility Function Parameters

Utility Function Parameter	Description	Lower Bound LB	Upper Bound UB	Mode $\hat{\theta}$
a_1	Unidimensional Utility of 0$	0.70	0.95	0.83
b_1	Exponent of Unidimensional Utility of Revenue Losses	Assumed	Deterministic	1.60
b_2	Exponent of Unidimensional Utility of Publicity Attribute	0.10	1.00	0.60
k_1	Scaling Parameter of Financial Attribute	0.43	0.81	0.68
k_2	Scaling Parameter of Publicity Attribute	0.04	0.43	0.20
k_3	Scaling Parameter of Medical Attribute	0.00	k_2	0.14
k_a	Scaling Parameter of Lives Saved Sub-Attribute	0.50	1.00	0.70
K	Overall Scaling Parameter for Multiplicative Function	Deterministic Function of k_1, k_2, k_3		-0.08

Probability Distribution

The probability mass functions $P(\mathbf{Y}/X)$ were obtained from Monte Carlo simulation and a series of stochastic models. Specifically, we encoded the director's subjective probability distributions (*15*) for the number of requests his program would receive in the upcoming year, conditional on the number of helicopters in the fleet. We then used Monte Carlo simulations to combine this distribution with a simulation model that we developed (*16*) to model the number of requests that could be serviced with given program configurations. This produced the density function for the number of patients carried—that is, Y_2—if Skymed were to operate X helicopters in the upcoming year.

We then simulated observations from this Y_2-distribution and input each observation into a stochastic model predicting extra net revenue (Y_1) and another stochastic model predicting the number of extra lives saved (Y_{3a}) and extra limbs saved (Y_{3b}). In determining the extra contributions, both models considered that some of the Y_2 (patients carried) would have been carried by Lifeflight had Skymed shut down. The revenue model also had to consider the likelihood that some of the patients that Lifeflight would have carried would have gone to university hospitals for the special treatments offered and some would be discretionary patients, going to the sponsoring hospital of the aeromedical program providing the transport. Once the appropriate number of extra patients was determined, revenues were determined from a stochastic model on the basis of past revenues and costs generated, and the number of lives and limbs saved was determined on the basis of probabilities given by Urdaneta et al. (*8*).

The specific numbers output from the revenue and medical models were then coupled with the specific Y_2-value input to them to form an "observed" attribute vector $[y_1, y_2, (y_{3a}, y_{3b})]$. We repeated this process 1,000 times, forming 1,000 $\mathbf{Y}(j) = [y_1(j), y_2(j), y_{3a}(j), y_{3b}(j)]$ vectors, $j = 1, 2, \ldots,$ 1,000. Assigning probabilities of .001 to each of these vectors formed $P(\mathbf{Y}/X)$ for X helicopters. The attribute levels in the joint distribution are probabilistically dependent, since the values of Y_1, Y_{3a}, and Y_{3b} depend on the value of Y_2 input, but the marginal distributions can be formed. We present the 0.25, 0.50, and 0.75 percentile values of the cumulative marginal distributions for $X = 1.0$ in Table 2.

If the program were to shut down, the attribute vector would be $\mathbf{Y} = [0,0,(0,0)]$ with certainty, by definition of the attributes. This vector, therefore, characterized the probability distribution of offering no service.

TABLE 2 Quartile Values of Attribute Marginal Cumulative Density Functions

Attribute Y_i	Units	$F_{Y_i}^{-1}(0.25)$	$F_{Y_i}^{-1}(0.50)$	$F_{Y_i}^{-1}(0.75)$
Y_1: Extra Net Revenue	10^6	-0.20	+0.20	+0.40
Y_2: Patients Carried	people	850	1050	1100
Y_{3a}: Extra Lives Saved	people	55	75	95
Y_{3b}: Extra Limbs Saved	people	30	50	70

ATTRIBUTE IMPORTANCE

The model reviewed can be used to compare the supplier's value of the multicommodity (i.e., net revenue, publicity, medical) bundle offered by operating one helicopter in the upcoming year to his value of the multicommodity bundle offered if the program were to shut down. By operating one helicopter ($X = 1$), Skymed would produce more publicity and medical benefits than if it were to shut down ($X = 0$), but it would incur fixed and operating costs that could be avoided. Moreover, operating the program with one helicopter would generate important revenues (as long as insurance companies are willing to reimburse charges for a large enough percentage of the patients transported), but it is not obvious that these revenues would offset the costs.

To determine the value of operating one helicopter in the upcoming year, we use Equation 1 with the 1,000 $\mathbf{Y}(j) = [y_1(j), y_2(j), y_{3a}(j), y_{3b}(j)]$ vectors (each occurring with probability .001) and calculate the expected utility, which we call $EU(\mathbf{Y})$:

$$EU(\mathbf{Y}) = \sum_{j=1}^{1,000} 0.001 * U[y_1(j), y_2(j), y_{3a}(j), y_{3b}(j)] \qquad (15)$$

The utility function U, given by Equations 3–7, 10, 12, and 13, depends on the vector of parameters $\boldsymbol{\Theta}$. (We do not explicitly write the dependence on $\boldsymbol{\Theta}$ to simplify the notation). Using the modes of the parameter distributions (i.e., $\hat{\Theta}_i$) from Table 1 as our best-guess estimates, which can be thought of as point estimates, we obtain $EU(\mathbf{Y}) = 0.65$.

The expected utility associated with shutting down (providing no service) is just $U[0,0,(0,0)]$, since $\mathbf{Y} = [0,0,(0,0)]$ would occur with certainty. Using the utility equations 3–7, 10, 12, and 13, we see that this expected utility is k_1a_1. Different $\boldsymbol{\Theta}$s give different k_1s and a_1s and, therefore, different expected utilities of the shut-down alternative. Using the best-guess $\hat{\Theta}_i$s gives $0.68 * 0.83 = 0.56$. This is less than value of $EU[\mathbf{Y}]$ determined above, meaning that when the best-guess estimates were used for the utility parameters, the multicommodity bundle associated with operating one helicopter is better than that associated with shutting down. That is, the program sees positive value in providing service.

To acknowledge the difficulties associated with determining the utility parameters $\boldsymbol{\Theta}$, we generated 1,000 observations from the distributions of the Θ_i parameters given in Equation 14, and Table 1, as described before (*3*). For each generated value we calculated the difference between the expected utility associated with operating one helicopter $EU[\mathbf{Y}]$ and that associated with operating no helicopters $EU[\mathbf{0}](= k_1a_1)$. We then had one $EU[\mathbf{Y}] - EU[\mathbf{0}]$ value for each of the 1,000 generated Θ-vectors. We present the distribution of these values for the first time in Figure 1. Note that all of the observations are positive, indicating that for all of the utility parameter vectors $\boldsymbol{\Theta}$, $EU[\mathbf{Y}] > EU[\mathbf{0}]$, and it would be better to operate one helicopter than to discontinue service.

Single-Attribute Analysis

Although the analysis shows that offering service is preferred to shutting down when considering the revenue, publicity, and medical benefits together, we are interested here in determining whether any single attribute could justify providing service when the costs of providing this service are included. The fixed costs associated with operating one helicopter (*3*) are $\$1.75(10^6)$. The variable costs depend on the number of hours flown, and from the performance characteristics, we found that, conditional on flying y_2 patients, they could be approximated as a deterministic $\$735y_2$ (*3*). Therefore, the cost $[c(j)]$ of providing the service associated with attribute vector $\mathbf{Y}(j)$ in the probability mass function is

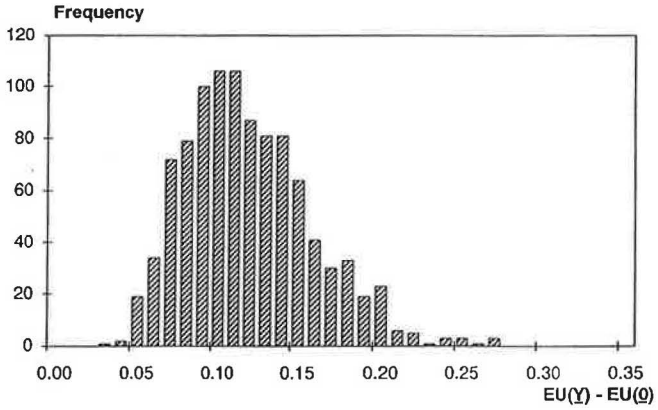

FIGURE 1 Distribution of $[\mathbf{Y}] - EU[\mathbf{0}]$ values across 1,000 Θ-vectors.

$$c(j) = 1.75(10^6) + 735y_2(j) \qquad (16)$$

To determine the value of the individual service benefits provided by operating one helicopter, we first define the following:

$$EU[Y_1] = \sum_{j=i}^{1,000} 0.001 * U[y_1(j), 0, (0,0)] \qquad (17)$$

$$EU[Y_2] = \sum_{j=i}^{1,000} 0.001 * U[-c(j), y_2(j), (0,0)] \qquad (18)$$

$$EU[Y_3] = \sum_{j=i}^{1,000} 0.001 * U\{-c(j), 0, [y_{3a}(j), y_{3b}(j)]\} \qquad (19)$$

$EU[Y_1]$ gives the expected utility of the revenues, combined with the costs (to form the net revenues y_1), generated from operating one helicopter, when there are no publicity or medical benefits (i.e., they are zeroed out). $EU[Y_2]$ gives the expected utility of the publicity, combined with the costs required to generate this publicity, when the revenues and medical benefits are zeroed out. $EU[Y_3]$ gives the expected utility of the medical benefits, combined with the costs required to generate these benefits, when the revenues and publicity benefits are zeroed out. That is, we consider the value of the individual attributes—revenues generated, publicity, medical benefits—along with the cost to produce them—by setting the other attributes to their alternative shut down levels. Given a vector of utility parameters $\boldsymbol{\Theta}$, it is straightforward to use our utility specifications to calculate these values.

To determine whether the attribute i is alone sufficient to justify service, we substituted $EU[\mathbf{0}]$ from $EU[Y_i]$ (both being

calculated with the same vector of utility parameters Θ_i). Because the utility function is internally scaled (*4,6,11*), the magnitude of the difference $EU[Y_i] - EU[\mathbf{0}]$ would be fixed up to the scaling imposed by $U[-6(10^6), 0, (0,0)] = 0$ and $U[+9(10^6), 3{,}500, 500, 300] = 1$ (*3*). These "dummy" vectors are not very meaningful, however (*17*). We therefore scaled the difference by $EU[Y_i] - EU[\mathbf{0}]$ to form RM_i:

$$RM_i = \frac{EU[Y_i] - EU[\mathbf{0}]}{EU[Y] - EU[\mathbf{0}]} \tag{20}$$

Because $EU[\mathbf{Y}] - EU[\mathbf{0}]$ is always positive (see Figure 1), the sign of RM_i would indicate whether the distribution of obtaining attribute i taken by itself (but considering the cost of obtaining the distribution) when providing service is better ($RM > 0$) or worse ($RM < 0$) than providing no service. The magnitudes of positive RM_is indicate how much (or little) of the increased value (expected utility) associated with the full set of attributes is obtained when only attribute i is considered, where the increased value is considered to be that above the value that would be obtained by providing no service. The magnitudes of negative RM_is give a feel for how far the value associated with attribute i falls short of the shut-down alternative, where the scale is again the increased value associated with operating one helicopter.

The distributions across the 1,000 Θs (Figure 2 and Table 3) show that RM_1, RM_2, and RM_3 overwhelmingly tend to be negative. That is, if considering any of the individual attributes in isolation, the preference would be to provide no service. This means that direct revenues are not sufficient ($RM_1 < 0$) to provide service. Moreover, the publicity or the medical benefits, when taken alone, do not offset the costs (RM_2, $RM_3 < 0$, respectively) of generating these benefits.

Importance of Revenue and Publicity

Service is provided to produce medical benefits. These results show, however, that the medical benefits alone do not outweigh the costs of providing service. To see if coupling the medical benefits (Y_3) with either the generated revenues (Y_1) or the associated publicity (Y_2) would be sufficient to justify service, we zero out the attribute not coupled. Specifically, we form

$$EU[Y_1, Y_3] = \sum_{j=i}^{1{,}000} 0.001 * U\{y_1(j), 0, [y_{3a}(j), y_{3b}(j)]\} \tag{21}$$

$$EU[Y_2, Y_3] = \sum_{j=i}^{1{,}000} 0.001 * U\{-c(j), y_2(j), [y_{3a}(j), y_{3b}(j)]\} \tag{22}$$

where U is the utility function used before and $c(j)$ is again obtained from Equation 16. As before, we subtract $EU[\mathbf{0}]$ and scale by $EU[\mathbf{Y}] - EU[\mathbf{0}]$ to form RM_{ij}:

$$RM_{ij} = \frac{EU[Y_i, Y_j] - EU[\mathbf{0}]}{EU[\mathbf{Y}] - EU[\mathbf{0}]} \tag{23}$$

and the interpretation is as before.

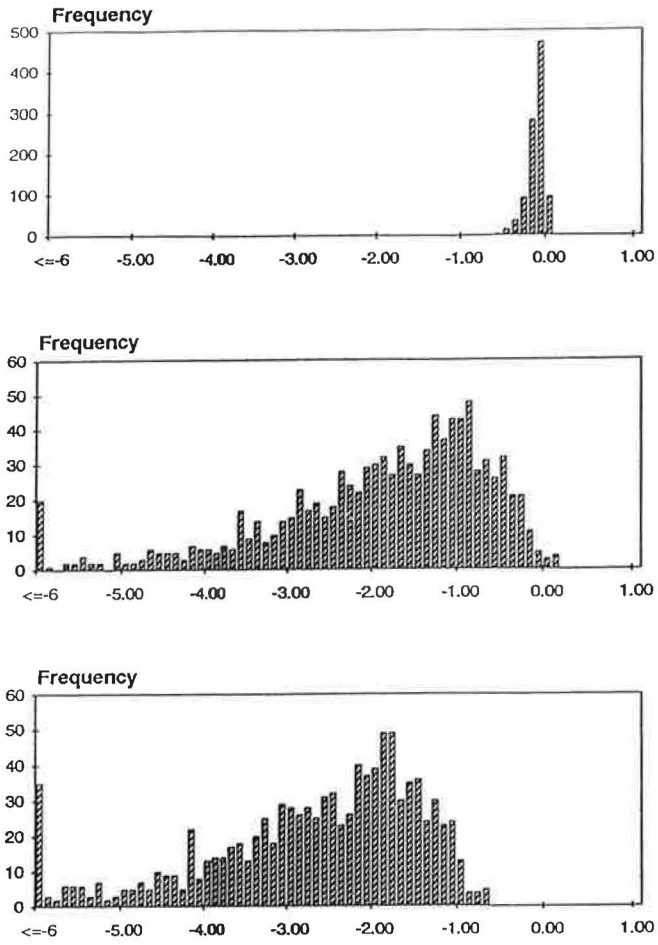

FIGURE 2 Distribution of relative value measures for single attributes RM_i across 1,000 Θ-vectors: *top*, RM_1 distribution; *middle*, RM_2 distribution; *bottom*, RM_3 distribution.

All of the terms in Equation 23 again depend on Θ-values, but they are easy to calculate once these values are given. In Figure 3 we present the RM_{ij} distributions across the same 1,000 Θ-values. The means and standard deviations of these distributions are presented in Table 3. There are only a few positive observations in the RM_{23} distribution, indicating that the combination of publicity and medical benefits would not be sufficient to justify providing service: Revenues must be generated to offset the cost of service. As seen by the RM_{13} distribution, the combination of revenues and medical ben-

TABLE 3 Means and Standard Deviations of *RM* Distribution Figures 2 and 3

Relative Value Measure	Distribution Mean	Distribution Standard Deviation
RM_1	-0.21	0.10
RM_2	-2.04	1.41
RM_3	-2.85	1.44
RM_{13}	-0.02	0.09
RM_{23}	-1.86	1.36

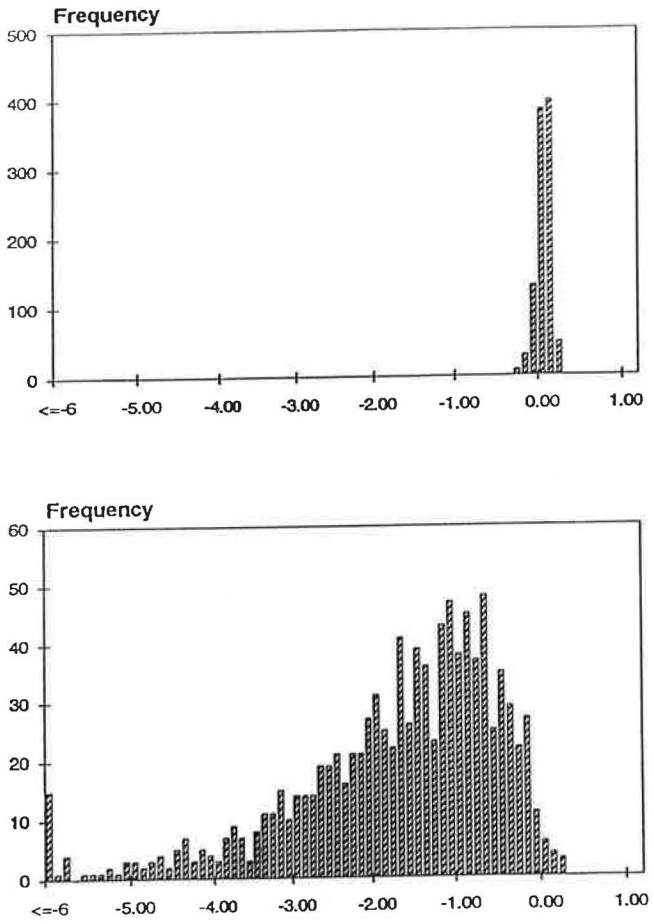

FIGURE 3 Distribution of (*top*) revenue-medical (RM_{13}) and (*bottom*) medical-publicity (RM_{23}) across 1,000 Θ-vectors.

efits might be enough to justify providing service, but the values are negative for many of the Θ-values, indicating that publicity might be needed to make service attractive in this setting.

DISCUSSION OF RESULTS

We emphasize again that these results must be considered preliminary, because the preference information was elicited from the program director, who put himself in the place of the real decision makers, for assistance in a slightly different problem (*3*). The director has worked with the decision-making board for several years, however, and has a good feel for its relative preferences in this area. Moreover, our experience in preference modeling and familiarity with the emergency medical industry has led us to believe that this preference information is representative of many service suppliers. The general results appear to be consistent across the great range of utility parameters allowed. The results may also be considered preliminary, since the performance side—that is, the probability distributions of the attribute vectors associated with providing service—is specific to one aeromedical program. We hope to investigate more programs in the future, but we have been impressed by the Skymed program and believe that our results would be representative of an efficient helicopter-based operation.

The results indicate that aeromedical service is truly multiobjective in nature. On the basis of revenue, publicity, and medical benefits offered, it was desirable to provide service for each of the 1,000 sets of preference values used. None of these service dimensions taken by itself was sufficient to justify operations, however, when the costs of supplying the service were considered. From an academic view, it is interesting to have identified a truly multiobjective service in this way. There are also more practical implications to this multiobjective nature.

First of all, the revenues generated are seen to be critical to the supplier. They make the greatest contribution to the positive value offered by the service. Without the revenues, Figure 3 (*bottom*) shows that the combined effect of publicity and medical benefits falls short of justifying supply of the service for all but a few (roughly 1 percent) of the vectors of utility parameters. Although the revenues make such an important contribution to the provision of service, the service cannot be considered a profit generator for the sponsoring hospital. Figure 2 (*top*) shows that, taken by itself, the net revenue is not enough to justify service. The median of the marginal cumulative net revenue function (Table 2) shows that there is a 50 percent chance of at least breaking even by providing service. Nevertheless, this is not considered sufficient because of the downside financial risks associated with, for example, low demand, bad weather, or lack of adequate insurance for an important percentage of the patients transported, and the aversion to losses shown in Equations 5 through 9. In the current system, the revenues are seen as a means to offset the cost of operations, not as profit.

To improve the revenue component of the service, the program could raise flight charges. Industry personnel believe that demand for transport is relatively price-inelastic. Raising charges would probably not be politically desirable, however, because of public concern with increasing health care costs. Moreover, it might have a negative net effect in the longer term if insurance companies decided that charges were too high and decided to stop reimbursing aeromedical transport completely. How much reimbursement could be cut before providing service becomes undesirable is an aspect we are now investigating.

If it would be difficult to increase revenues to the program, the financial desirability could be increased by decreasing costs. Fixed costs could be decreased by merging the two competing aeromedical systems in the area, an idea that definitely did not originate here. Countering this idea is, of course, the argument that competition increases long-term efficiency and could ultimately decrease costs. The results presented here do not contribute to this issue, either for or against. Our results show, however, that the publicity to the sponsoring hospital provided by flying emergency missions is perceived as real and important. Without it, Figure 3 (*top*) shows that the service would probably not be considered desirable. Merging programs would reduce, if not eliminate, the publicity dimension.

Finally, aeromedical service does provide medical benefits, and these contribute to the value of the service. When paired with the revenue contributions, there is roughly a 50 percent chance of making the service desirable [Figure 3 (*top*)]. If

costs could be recovered with certainty, the positive value associated with the medical benefits would, of course, be sufficient to justify service. Our model of the supply of medical benefits was based on the only relevant study (*8*) we could find, and there was a need to interpret the study for our needs. We are embarking on a sensitivity analysis of the results presented there, and we have noted a desire on the part of the industry to have more studies on the medical contributions of aeromedical transport. From Figure 2 (*bottom*), it appears, however, that medical benefits alone are far from sufficient to keep programs in the air, and if aeromedical programs are to continue to provide this life-saving service, suppliers will need to be ensured of improved financial security or continue to believe that the service is providing important publicity to the sponsoring hospitals.

ACKNOWLEDGMENTS

The authors would like to thank Skymed and Dave Kerins for their assistance. The views and interpretations presented are the authors', however. This work was supported by a National Science Foundation grant and a donation by Transportation Research Center, Inc., to Mark McCord's Presidential Young Investigator endowment fund.

REFERENCES

1. D. F. Kerins, *A Systems Approach to the Identification and Characterization of Civilian Air Ambulance Services*. M.S. thesis. Department of System Science, University of Louisville, Ky., 1989.
2. Top Ten Transport Cities. *Journal of Air Medical Transport*, Vol. 10, No. 5, 1991.
3. M. R. McCord, O. Franzese, and X. Sun. Multicriteria Analysis of Aeromedical Fleet Expansion. *Journal of Applied Mathematics and Computation* (in preparation).
4. R. de Neufville. *Applied System Analysis*. McGraw-Hill, New York, N.Y., 1990.
5. A. Goicoechea, D. R. Hansen, and L. Duckstein. *Multiobjective Decision Analysis with Engineering and Business Applications*. John Wiley & Sons, New York, N.Y., 1982.
6. R. L. Keeney and H. Raiffa. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. John Wiley & Sons, New York, N.Y., 1976.
7. *Readings in Decision Analysis*. Decision Analysis Group, Stanford Research Institute, Menlo Park, Calif., 1977.
8. L. F. Urdaneta, M. K. Sandberg, A. E. Cram, T. Vargish, P. R. Jochimsen, D. H. Scott, and T. J. Blommers. Evaluation of an Emergency Air Transport Service as a Component of a Rural EMS System. *The American Surgeon*, Vol. 50, 1984, pp. 183–188.
9. M. R. McCord and C. Leotsarakos. Investigating Utility and Value Functions with an Assessment Cube. In *Risk, Decision, and Rationality* (B. R. Munier, ed.). D. Reidel, Dordrecht, the Netherlands, 1988, pp. 59–75.
10. P. Delquie. *Contingent Weighting of the Response Dimension in Preference Matching*. Ph.D. dissertation. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, July 1989.
11. M. R. McCord and R. de Neufville. Lottery Equivalents: Reduction of the Certainty Effect Problem in Utility Assessment. *Management Science*, Vol. 32, No. 1, 1986, pp. 56–60.
12. M. R. McCord and R. de Neufville. Assessment Response Surface: Investigating Utility Dependence on Probability. *Theory and Decision*, Vol. 18, 1985, pp. 263–285.
13. J. C. Hershey, H. C. Kunreuther, and P. J. H. Schoemaker. Sources of Bias in Assessment Procedures for Utility Functions. *Management Science*, Vol. 28, 1982, pp. 936–954.
14. R. Krzysztofowicz and L. Duckstein. Assessment Errors in Multiattribute Utility Functions. *Organizational Behavior and Human Performance*, Vol. 26, No. 3, 1980, pp. 326–348.
15. C. S. Spetzler and C. S. Stael von Holstein. Probability Encoding in Decision Analysis. In *Readings in Decision Analysis*. Stanford Research Institute, Menlo Park, Calif., 1977, pp. 403–427.
16. M. R. McCord, X. Sun, D. Kerins, and O. Franzese. Simulating Aeromedical Helicopter Capacities. *Proc., 22nd Annual Pittsburgh Conference on Modeling and Simulation*, Vol. 22, Part 1, 1991, pp. 145–151.
17. B. Roy. *Methodologie Multicritere d'Aide à la Decision* (in French). Economica, Paris, France, 1985.

Publication of this paper sponsored by Committee on Transportation Supply Analysis.