TRANSPORTATION RESEARCH
RECORD

No. 1452

Planning and Administration

Travel Forecasting and Supply Models

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1994

Transportation Research Record 1452
ISSN 0361-1981
ISBN 0-309-06060-5
Price: $22.00

Subscriber Category
IA planning and administration

Printed in the United States of America

Sponsorship of Transportation Research Record 1452

GROUP 1—TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Chairman: Thomas F. Humphrey, Massachusetts Institute of Technology

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

Committee on Passenger Travel Demand Forecasting
Chairman: Eric Ivan Pas, Duke University
Bernard Alpern, Moshe E. Ben-Akiva, Jeffrey M. Bruggeman, William A. Davidson, Christopher R. Fleet, David A. Hensher, Alan Joel Horowitz, Joel L. Horowitz, Ron Jensen-Fisher, Peter M. Jones, Frank S. Koppelman, David L. Kurth, T. Keith Lawton, David M. Levinsohn, Fred L. Mannering, Eric J. Miller, Michael R. Morris, Joseph N. Prashker, Charles L. Purvis, Martin G. Richards, Earl R. Ruiter, G. Scott Rutherford, Galal M. Said, Gordon W. Schultz, Peter R. Stopher, Antti Talvitie, A. Van Der Hoorn

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, Randolph W. Hall, Rudi Hamerslag, Bruce N. Janson, Haris N. Koutsopoulos, Chryssi Malandraki, Eric J. Miller, Anna Nagurney, Earl R. Ruiter, K. Nabil A. Safwat, Mark A. Turnquist

Task Force on Transportation Modeling Research Needs
Chairman: Eric Ivan Pas, Duke University
Paul E. Benson, David E. Boyce, Elizabeth Deakin, Frederick W. Ducca, Peter W. Glynn, Greig W. Harvey, Ryuichi Kitamura, George T. Lathrop, T. Keith Lawton, Hani S. Mahmassani, Carroll J. Messer, Richard H. Pratt, Stephen H. Putman, Earl R. Ruiter, Gordon W. Schultz, Gordon A. Shunk, Peter R. Stopher, Mary Lynn Tischer, Edward Weiner

Transportation Systems Planning Section
Chairman: Bruce D. McDowell, US ACIR

Committee on Statewide Multimodal Transportation Planning
Chairman: Michael D. Meyer, Georgia Institute of Technology
David Preston Albright, Eng Reda Ba-Faqueeh, Linda Bohlinger, James L. Covil, Marta V. Fernandez, John W. Fuller, John S. Hassell, Jr., Ronald G. Hoffman, Thomas F. Humphrey, Gloria J. Jeff, William O. Knox, David M. Levinsohn, Melvin L. Mitchell, Roland A. Nesslinger, Lance A. Neumann, Neil J. Pedersen, Judy A. Perkins, Catherine L. Ross, Roger L. Schrantz, David G. Snider, James P. Toohey, Richard A. Torbik, Jeffrey W. Trombly, Joanne Walsh

Public Transportation Section
Chairman: Subhash R. Mundle, Mundle & Associates, Inc.

Committee on Public Transportation Planning and Development
Chairman: Patricia V. McLaughlin, Metropolitan Transit Authority
Secretary: David R. Miller, Parsons Brinckerhoff et al.
Paul J. Ballard, Edward A. Beimborn, Chester E. Colby, Sally Hill Cooper, Stephen P. Gordon, George Edward Gray, S. Olof Gunnarsson, Brendon Hemily, Jay A. Higle, Nathan L. Jaschik, Hermann Knoflacher, Perry J. Maull, Marianne A. Payne, Patti Post, James P. Redeker, Jacques P. Roulet, G. Scott Rutherford, Dennis H. Ryan, George M. Smerk, Katherine F. Turnbull, Samuel L. Zimmerman

Transportation Research Board Staff
Robert E. Spicher, Director, Technical Activities
James A. Scott, Transportation Planner
Peter L. Shaw, Public Transportation Specialist
Nancy A. Ackerman, Director, Reports and Editorial Services
Norman Solomon, Editor

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

Transportation Research Record 1452

Contents

Foreword

The papers contained in this volume focus on the applications of modeling techniques to various aspects of transportation.

One paper discusses the applications of a travel forecasting model to transit system applications. Other papers discuss modeling of travel demand to analyze the effects of new transportation policies on travel patterns and modeling of travel behavior for various aspects of transportation control measures. Also included are a simulation for evaluating performance of emerging response fleets, a model for vehicle sizing for bus transit networks, a model for forecasting transit ridership based on the Seattle experience, and a control model addressing the dynamic freeway control process.

The final paper presents a discussion of the potential of implementing multimodal planning and programming.

Assessing User Benefits of Transit System Improvements with Spatially Varying Demands

ALAN J. HOROWITZ

Transit planners recognize that spatially varying demands affect the assessment of transit system alternatives. However, they do not yet possess the tools necessary to properly determine the effects of the variation on estimates of user benefits. An extended measure of user benefits that is consistent with net consumer surplus from classical economic theory is presented. Also presented is the structure of a travel forecasting model that can show the effects of activity allocation, trip distribution, and route choice on net consumer surplus. Individual components of the model have already been extensively tested in practice and are described in the academic literature, but the transit ridership properties of the model, as a whole, have not been established. The model is capable of finding a joint equilibrium solution between activity allocation, mode split, trip distribution, and traffic assignment. Tests of the model on real networks indicate that spatial redistribution of activities resulting from a transit service improvement can be large enough to determine whether the improvement should be implemented.

The measurement of user benefits and its role in transit decision making has recently become a hotly debated subject (*1*). Some of the debate relates to issues of integrity and competence, which cannot be resolved by better measurement methods. However, many remaining issues could be resolved by adopting the best available forecasting techniques, applying them properly, and developing good indicators of user benefits.

Two closely related arguments have attracted considerable attention from planners and researchers. First is the contention that improvements in transportation systems, such as increases in capacity, may not always result in improved user benefits. Some economists argue that demand can become so elastic that an improvement can attract too many users in the long term, causing the whole system to operate less efficiently than before (*2*). That argument is counterintuitive and unlikely to apply to transit system improvements, but it focuses renewed attention on the nature of travel demand and how it affects user benefits. For example, correct assumptions regarding demand elasticity may give lower estimates of benefits than would result if current ways of thinking were applied.

Second, some communities have requested funds for transit system expansion despite poor ridership forecasts. Leaders of the communities have argued that there is an intrinsic relationship between transit supply and the long-term distribution of activities in their region. It is further argued that we do not yet possess the methodology to measure that relationship, so user benefits must be greater than indicated. Although that argument seems plausible, it lacks convincing verification.

Center for Urban Transportation Studies, University of Wisconsin–Milwaukee, EMS Building, Room E 387, P.O. Box 784, Milwaukee, Wis. 53201.

Both arguments could be laid to rest, at least partially, by forecasting models that properly reflect the amount of elasticity found in an actual transportation system. At least such a model must be able to achieve a joint equilibrium solution between mode split, trip distribution, activity allocation, and highway traffic assignment—all with sufficiently realistic relationships. Although the individual components of such a model have been identified for almost two decades, it is only recently that equilibrium solutions satisfying Wardrop's first principle could be obtained.

This paper describes the overall structure of such a model and explains how it might be operated and validated. Tests are then performed to identify the likely "winners" of the two arguments.

NET CONSUMER SURPLUS OF SERVICE CHANGES

User benefits are those that result from increased accessibility when a transit system improves. Benefits accrue to a transit patron because a trip can be made with less cost or greater convenience. Benefits also can accrue to an automobile driver or a passenger, because there might be less congestion on some streets. Furthermore, benefits could accrue to a traveler who chooses to make an additional trip by either mode or to switch modes.

Many benefit studies in the past determined that the greater user benefit resulting from a transportation system improvement is travel time savings. Additional user benefits include user savings from lower costs of fuel, tolls, fares, and vehicle maintenance. In addition, intangible user benefits could include travel comfort and the ability either to make entirely new trips or to satisfy old trip purposes by traveling to a better, but more distant, destination.

In our largest cities, there is increasing interest in transit's impact on traffic congestion. There are two aspects of this impact: (*a*) degradation of traffic flow associated with buses sharing roads with automobiles and (*b*) improvements in traffic flow that might occur if some drivers could be persuaded to take transit. Both of these effects, which are components of user benefits, can be measured with the proper methodology.

Economists tell us that user benefits of any public project can be ascertained by calculating net consumer surplus. Consumer surplus is the difference between the amount an individual is willing to pay for a good and the amount the individual actually pays. Net consumer surplus is the change in consumer surplus caused by the public project.

When dealing exclusively with highway travel, it is sometimes possible to estimate user benefits by adding individual components. However, transit benefits are far more complicated, so it is easiest

to estimate directly the net consumer surplus of the system change from a travel forecasting model. If calculated correctly, net consumer surplus will include all of the previously cited benefits, both tangible and intangible.

Classical economic theory deals mainly with changes in price. Clearly, benefits still can accrue to transit users even if fare is constant, such as with improved headways, elimination of transfers, faster speeds, or line extensions. Some service improvements can reduce the duration of trips; other service changes improve the convenience of trips. It is important to include these nonmonetary changes in any estimate of net consumer surplus.

For any given transit trip, it is possible to calculate a comprehensive measure of its costs and inconveniences, that is, the trip's disutility. Disutility is most easily interpreted when it is expressed in units of automobile riding time. A typical disutility function would look like this:

$$\begin{aligned}\text{Disutility} = {} & \text{automobile riding time} \\ & + (\text{transit riding time})(\text{transit riding weight}) \\ & + (\text{walking time})(\text{walking weight}) \\ & + (\text{waiting time})(\text{waiting weight}) \\ & + (\text{transfer time})(\text{transfer weight}) \\ & + \text{initial wait penalty} + \text{first transfer penalty} \\ & + \text{second transfer penalty} \\ & + [\text{fare}/(\text{value of time})] + [(\text{tolls} + \text{parking costs} \\ & + \text{vehicle operating costs})/(\text{value of time})] \\ & + [(\text{vehicle ownership costs})/(\text{value of time})] \end{aligned} \quad (1)$$

In this equation, the value of time is the rate at which travelers would be willing to trade money for time. Typical values of weights and penalties are given in Table 1. The weights originally were derived through a psychological scaling experiment (*3,4*), but they are consistent with weights observed from the calibration of mode split models and have been adopted widely for travel forecasting.

Equation 1 deals exclusively with cost and convenience issues. Additional terms could be provided for other significant elements of comfort, such as protection from inclement weather and privacy.

The only vehicle ownership costs that should be included in Equation 1 are those that can be attributed to a single trip. Because it has been found that travelers do not correctly perceive the full value of their vehicle ownership costs while making mode choice decisions, this term is often omitted.

Travelers have a willingness-to-pay in units of travel time (*5*). They will choose to ride only if the disutility of travel (in time units) is less than their willingness-to-pay (in time units). Consequently, any traveler must possess a consumer surplus of disutility. That disutility may be expressed mathematically as time savings, or it can be converted to monetary units by multiplying time savings by the value of time. For this research, consumer surplus is left in time units to avoid complications associated with time valuation.

A disutility measure of consumer surplus is shown in Figure 1 for a single trip. A demand curve represents the relationship between numbers of trips and trip disutility, expressed in time units. Point 1 represents the original disutility and number of riders taking the trip. Point 2 indicates a new disutility and the number of riders after a service change, such as shortening the headway. Because of the service improvement, more people have chosen to take the trip. Some new riders switched from the automobile, some new riders have changed their choice of destination, and some new riders are making an entirely new trip. T_1 is the original disutility, and T_2 is the new disutility. All of the old riders receive windfall consumer surplus of $T_1 - T_2$. The windfall is shown as the shaded area A. New riders have a net consumer surplus represented by the shaded area B. The new riders' net consumer surplus is almost a triangular area. Consequently, the total consumer surplus could be found from the roughly trapezoidal, combined area:

$$\text{Net consumer surplus} \approx (T_1 - T_2)(Q_1 + Q_2)/2 \quad (2)$$

Net consumer surplus may be found by subdividing the shaded area into several flat and wide trapezoids and adding their areas, as shown in Figure 2. The process of finding the area of several smaller trapezoids can be expressed mathematically as:

$$\text{Net consumer surplus} = -\int_{T_1}^{T_2} Q(T)\,dT \quad (3)$$

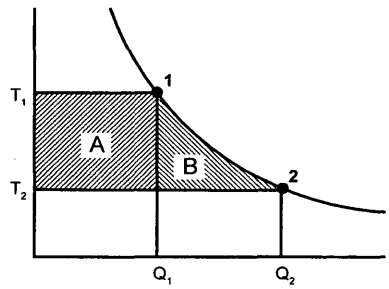


FIGURE 1 Net consumer surplus of trips by a single mode and origin-destination pair.

TABLE 1 Typical Weights and Penalties for Travel Disutility

Weight or Penalty	Value
Transit Riding Weight	1 + 2.0 x (fraction of person-time standing)
Walking Weight (good weather)	1.3
Waiting Weight	1.9
Transfer Weight	1.6
Initial Weight Penalty	8.4 minutes
Transfer Penalty (first or second)	23 minutes
Value of Time	0.167 to 0.333 of the average wage of choice riders

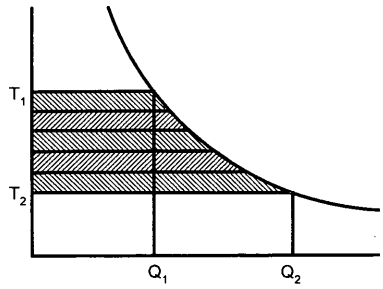


FIGURE 2 Integrating net consumer surplus with trapezoids.

where $Q(T)$ is ridership as a function of disutility (*6*). Because of the integral sign, Equation 3 looks more complicated than it really is. Integral calculus is never actually used to perform such a computation. Instead, one would simply divide the service change into several small increments and compute the net consumer surplus with Equation 2 as each increment is applied.

In a multimodal transportation system it is necessary to sum the net consumer surplus over all possible modes. For example, highway traffic could decline slightly as the result of the service improvement illustrated in Figure 1. Total net consumer surplus for the whole system can be found from the relationship

$$\text{Net consumer surplus} = -\sum_m \sum_i \sum_j \int_{T_{1mij}}^{T_{2mij}} Q_{mij}(T)\, dT \qquad (4)$$

for all modes *m*, all origins *i*, and all destinations *j*. As before, the integral is performed by summing the areas of flat, wide trapezoids (*7*). Unlike the example in Figure 1, Equation 4 also applies to modes that result in losses for users. For example, a highway system can contribute positively to consumer surplus by congestion relief while still giving some of its users to the transit system. Those highway users that remain will realize a windfall consumer surplus equal to the travel time reduction for the trip. Drivers choosing to switch paths are also counted properly by Equation 4.

It is sometimes useful to break net consumer surplus into components to determine the primary sources of the benefits. For example, highway consumer surplus may be differentiated from transit consumer surplus.

FORECASTING TRANSIT RIDERSHIP WITH SPATIALLY VARYING DEMANDS

A travel forecast that can properly measure net consumer surplus is only slightly more difficult than a conventional forecast, provided care is taken to compute the necessary values of disutility and demand for all modes. The model described in the following paragraphs should be considered a reasonable way of doing travel forecasting, but it is certainly not the only way, and it may not even be the best way. The model is an assemblage of tried-and-true components from current planning practice and the academic literature. Techniques that are now considered experimental may later prove to be more accurate. However, it is important for the purposes of this research to adopt only components that have achieved a consensus as to their validity.

The travel forecasting model was custom built for this research, because there are not any commercial forecasting packages that contain all the necessary elements. Nonetheless, each separate component should be considered off-the-shelf technology, even if the entire package seems unusually complex. The activity allocation step was adapted from source code of the Highway Land Use Forecasting Model II, and the traffic and transit forecasting components were taken from the source code of the Quick Response System II. The various pieces were compiled into a single executable module. The pieces are discussed briefly in the following paragraphs.

Activity Allocation Issues

The allocation of activities throughout a region must be sensitive to the quality of transportation services. The most widely researched way of achieving that sensitivity is the Lowry land use model (*8,9*). The Lowry-Garin model, specifically implemented for this paper, is shown in Figure 3. The underlying mathematical relationships have been described elsewhere (*10,11*). Other land use formulations exist that are based on similar principles.

A Lowry-Garin model allocates residences proximately to workplaces and allocates services proximately to their markets. Within a Lowry-type model, services are defined as those employers who derive their income from within the region and who are sensitive to the locations of their customers. Services are further subdivided into two classes: (*a*) those that serve people and tend to locate proximately to concentrations of population and (*b*) those that serve businesses and tend to locate proximately to concentrations of employees.

Population is allocated to zones with a residential location model on the basis of the residential attractiveness of the zone (typically, net developable area) and on the basis of the disutility of travel between the zone of residence and all zones of employment. Services are allocated to zones in much the same way as residences are allocated to zones, considering both service attractiveness and the disutility of travel.

A Lowry-type model cannot allocate "basic" industries (businesses that derive their income from outside the region), so their locations must be provided as input.

Lowry-type models become computationally messy because services themselves must be served and because services have employees needing residences. Consequently, Lowry-type models simultaneously solve for the number of people and the number of service employees in every zone. Such a solution requires a large amount of computation, especially if the model must also resolve conflicts over land, satisfy hard constraints on population or on service employment, or introduce agglomeration effects.

Once population and services have been allocated, it is possible to perform a traffic forecast in the usual way. The traffic forecast may reveal unanticipated congestion effects, so the activity allocation step may have to be repeated.

Equilibrium Assignment Issues

When computing consumer surplus, it is important that automobile disutility be consistent with the amount of traffic along the path from origin to destination. In addition, the amount of traffic should be sensitive to possible variations in activity allocation, mode split, and the distribution of trips, all of which depend on automobile disutility. This consistency is sometimes referred to as an elastic demand-equilibrium assignment.

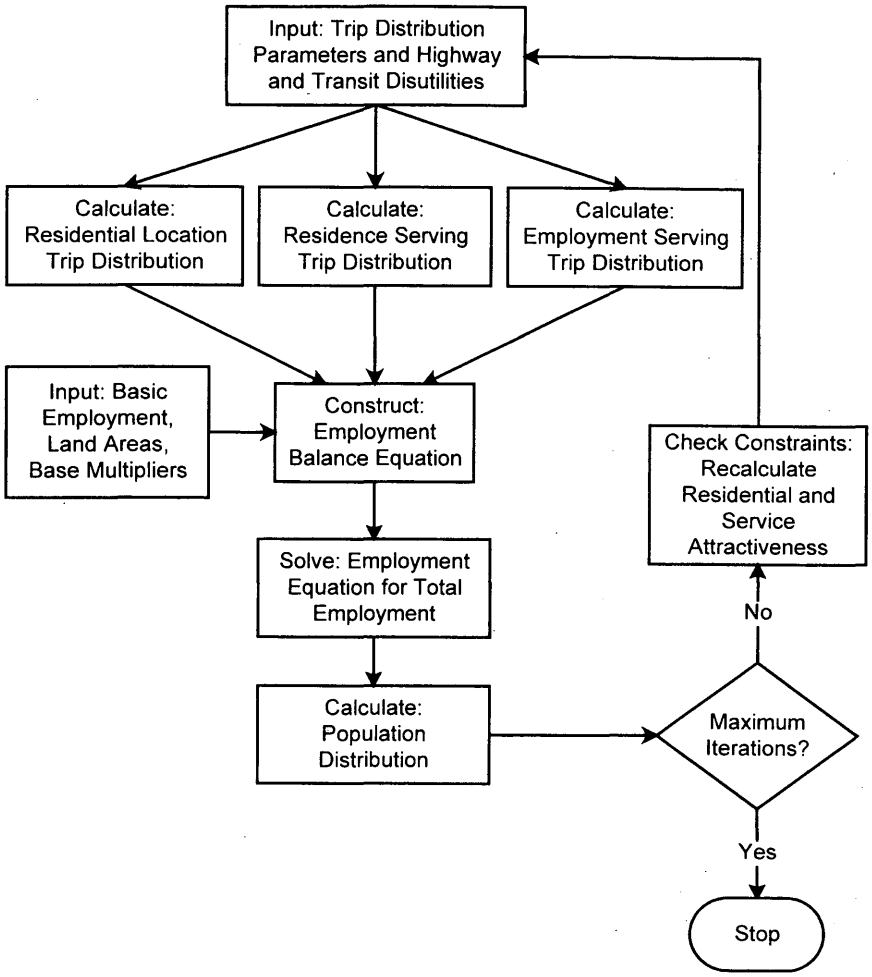


FIGURE 3 Activity allocation step of the travel forecast.

The chosen method of obtaining an equilibrium assignment is shown in Figure 4 and contains many of the same steps as a traditional travel forecast. However, the model shown in Figure 4 differs from traditional travel forecasting by routing the feedback loop so that the trip distribution, mode split, and activity allocation steps can be based on the highway disutilities that are appropriate for the amount of traffic congestion. Critical to the feedback loop is an averaging step (*12*). At this step, traffic volumes from all previous all-or-nothing traffic assignments are averaged together. Then new disutilities on each link are obtained. It has been previously shown that equilibrium solutions can be consistently obtained in this manner. An unweighted average works consistently well, although convergence is slow (*13,14*).

Assignment Convergence Issues

Given the complexity of the model and the method chosen for obtaining an equilibrium solution, we lack standard means to determine when convergence has been reached. The accepted method of monitoring an objective function of a nonlinear program is not available in this case. The surest method of determining whether an equilibrium solution has been reached is to compare the final total assigned travel time with the travel time obtained by loading the final trip table on to the network with all-or-nothing assignment. The two total travel times must be equal for the network to be in equilibrium. Less precise, but almost as effective, is to monitor assigned volumes on successive iterations. In either case, the uniqueness of the equilibrium solution cannot be established.

Location Models within Activity Allocation Step

All three location models are singly constrained, entropy-maximizing gravity models.

Composite Disutilities

Most travel forecasts find the distribution of trips throughout the community with a model step that excludes information about the quality of transit service. Consequently, such a forecast will not be properly sensitive to changes in transit service. Forecasters have

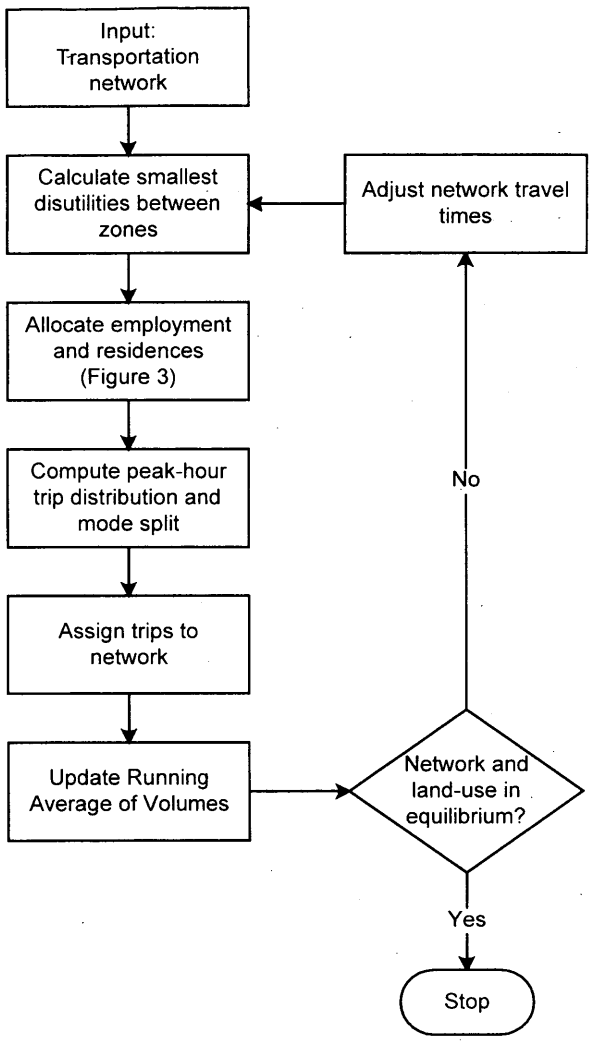


FIGURE 4 Combined-steps method of travel forecasting.

sometimes included transit service in the trip distribution step by computing composite disutilities between origins and destinations that account for both highway and transit service. The following composite disutility function has been found to provide the correct degree of sensitivity:

$$T^c_{ij} = \ln[\exp(-\alpha\, T^b_{ij}) + \exp(-\alpha\, T^a_{ij})] / (-\alpha) \quad (5)$$

where

T^c_{ij} = composite disutility from origin i to destination j,
T^b_{ij} = disutility by transit,
T^a_{ij} = disutility by automobile, and
α = coefficient for an unweighted minute of travel time in a mode split model (*15*).

The composite disutility is always smaller than the smallest value of its components. Composite disutilities should not be used for trips that are captive to any particular mode.

Any forecast can be performed either with or without Equation 5. The differences in the two forecasts can be interpreted as transit's impact on the spatial distribution of activities.

Application of Composite Disutility Function

The composite disutility function was used for the distribution of all trip purposes, except for employment-serving trips in the activity allocation step. Employment-serving trips are rarely made by transit and are assumed to be captive to the automobile.

Other Distribution Issues

The full trip distributions from the activity allocation step are discarded before the trip distribution step of the traffic and transit forecast. Somewhat inefficiently, the procedure retains only the estimates of activity levels in each zone. The trip distribution step in Figure 4 uses a doubly constrained gravity model, which satisfies both attraction and production end constraints. Consequently, trip distribution is less sensitive to variations in disutility than the location models of the activity allocation step (Figure 3).

Mode Split

Mode split was handled with a binary logit model (automobile, generalize transit) with two market segments (captive, choice). Transit trips were loaded with a stochastic, multipath assignment algorithm.

ISSUES OF MEASURING CONSUMER SURPLUS

Approximating Net Consumer Surplus Integral with Trapezoids

Transit service changes can be either discrete or continuous. An example of a discrete service change would be the addition of a new rail station. An example of a relatively continuous service change would be an improvement in headways. It would make sense to compute the net consumer surplus of only part of a headway improvement, but it would make little sense to compute the net consumer surplus of only part of a new station. For discrete service changes, there can be only two possible valid forecasts—with and without the change. Consequently, net consumer surplus must be computed by Equation 2, recognizing that an overestimate in benefits is possible.

For continuous service changes, the calculation of net consumer surplus can be more precise. The service change can be divided arbitrarily into several increments and the net consumer surplus computed for each increment. The sum of the net consumer surpluses for each increment is the total net consumer surplus. The major drawback to subdividing service changes in this manner is the added computation time necessary to evaluate each amount of intermediate service.

Double Counting

When benefits are calculated for a project, it is important to avoid double counting. Because consumer surplus as defined here is such a broad measure, it encompasses effects, such as land value changes, that can be measured separately. Environmentally related benefits, such as land preservation, are not included in consumer surplus.

Need for Realistic Null Alternative

Net consumer surplus is always calculated between a before case and an after case. The most relevant before case is the null alternative, that is, the most likely state of the community without the service change. The null alternative is not necessarily the current state of affairs. The null alternative should include growth or decline, redistribution of activities, or natural changes in the character of the community. Good null alternatives are difficult to construct, but they are essential to a valid calculation of consumer surplus.

CASE STUDY NETWORKS AND SCENARIOS

The case study region selected for this study was Wausau, Wisconsin, because its networks have been extensively used for testing both travel forecasting models and land use models. Its networks are known to behave similarly to those from much larger cities. The Wausau network has the advantage of computational speed, because it has only 36 highway zones and 9 external stations. Calibration runs were made to ensure that the forecasting model produced reasonable highway volumes and transit loads.

Separate networks were created for the highway and transit systems. The transit system had only five bus routes, operating on 30-min headways. Highway zones were subdivided into 60 transit zones for the purposes of transit trip assignment. All highway links were given a capacity (Level of Service C, design capacity), and delay was calculated exclusively with the BPR speed and volume function.

Wausau does not have much traffic. To determine how activity levels influence the net consumer surplus of a service change, two different states of the city were created:

- Scenario 1: current activity levels, basic employment at existing conditions; and
- Scenario 2: twice current activity, basic employment doubled in each zone.

Doubling the basic employment in the Lowry-Garin model has the effect of doubling both population and service employment in the region as a whole. Scenario 2 is quite congested, and drivers have ample incentive to choose transit.

It is possible that a forecasting model can take on a much different character when evaluating large service changes instead of small ones. To determine whether the magnitude of the service change had interesting effects, two different service changes were created:

- Service Change A: a reduction of headways from 30 to 15 min (mild), and
- Service Change B: a reduction of headways from 30 to 5 min and elimination of the $0.50 fare (aggressive).

CASE STUDY MEASUREMENT OF NET CONSUMER SURPLUS

Except as noted, calculations were conducted with sufficient precision for routine regionwide travel forecasting. Of special concern were errors associated with insufficient equilibrium iterations (Figure 4) and the method of integration. A good equilibrium solution is essential to accurate estimates of consumer surplus.

Convergence Error

An earlier study indicated that approximately 20 iterations of the equilibrium loop of Figure 4 would usually be sufficient for travel forecasting purposes (*11*). However, the number of iterations necessary for precise calculation of net consumer surplus was not determined. Because net consumer surplus is found by comparing two different forecasts, convergence errors in each forecast can combine unpredictably.

Table 2 gives the values of net consumer surplus for 10, 20, 40, and 100 equilibrium iterations for Scenario 2 and Service Change B. It appears that net consumer surplus stabilizes at about 20 iterations for this network. Assuming that a 100-iteration forecast contains essentially no convergence error, the error at 20 iterations is an acceptable 0.13 percent.

All remaining simulations described in this paper were still nicely converging to an equilibrium solution at 20 iterations.

Integration Slices

Simulations are time consuming, so it is tempting to use a single slice when integrating the net consumer surplus. A single slice will cause an overestimate, which becomes worse as the difference between the alternatives becomes larger. Table 3 compares the net consumer surplus with a single slice to that with five slices for Scenario 2 and Service Change B. Because the combination of scenario and service change produces a large net consumer surplus, the integration error is also large. The overall integration error for a single slice is at least 20.4 percent.

TABLE 2 Variation in Net Consumer Surplus with Equilibrium Iterations*

Number of Iterations	Net Consumer Surplus		
	Highway	Transit	Total
10	19957	349679	369636
20	23703	348872	372575
40	22817	349531	372348
100	23200	348901	372101

*Twice current activity levels (Scenario 2), with composite disutilities. Before: 50 cent fare; 30 minute headways. After: 0 fare; 5 minute headways (Service Change B). Units are minutes of riding time.

TABLE 3 Comparison Between a One-Slice and a Five-Slice Integration*

Number of Slices	Net Consumer Surplus		
	Highway	Transit	Total
5	26002	283370	309372
1	23703	348872	372575
Percent Difference	-8.8%	23.1%	20.4%

*Twice current activity levels (Scenario 2), 20 iterations, with composite disutilities. Before: 50 cent fare; 30 minute headways. After: 0 fare; 5 minute headways (Service Change B). Units are minutes of riding time.

Because the conclusions are unaffected by this type of systematic error, the remaining analysis in this paper uses a single integration slice. However, it would be a mistake to rely exclusively on single slices in practical applications. User benefits can be seriously overestimated with big system changes, making them seem more cost-effective than they really are.

Computational Considerations for Lowry-Garin Model

The solution of the employment balance equations within the Lowry-Garin model is always exact, but the need to resolve conflicts over available land required repeated solutions of these equations. The number of needed land use iterations (the loop of Figure 3) during any given equilibrium iteration (Figure 4) was reduced by capturing the attractiveness values from the previous equilibrium iteration. Thus, it was possible to reduce the number of land use iterations to just three, achieving a significant reduction in computation time. Because of capturing, the resolution of land conflicts becomes increasingly better with each equilibrium iteration.

RESULTS

Transit Ridership

Not all of Wausau is served by transit. When transit service is improved, the model tends to concentrate activities in zones served by transit. Some zones pick up additional services, whereas other zones pick up population. In either case, the concentration has a positive effect on ridership. Table 4 compares gains in forecast transit ridership with and without the composite disutility function. Without the composite disutility function, transit can influence the distribution of activities only by relieving highway congestion, a relatively minor effect.

Table 4 indicates that redistribution of activities causes a ridership increase of 3 to 14 percent. Although such increases are not dramatic, they would lead one to favor the more aggressive alternatives for transit service improvement. A particularly good transit system can induce some of its own demand.

Table 4 also indicates, as expected, that transit ridership increases substantially faster than the activity level. The composite disutility function does not appear to interact with the level of activity in the region.

Consumer Surplus

Not surprisingly, the effect on net consumer surplus of activity redistribution is similar to the effect on ridership gains. Table 5 presents net consumer surplus for both highway and transit users. Net consumer surplus increases from 3 to 12 percent with the composite disutilities.

The values of net consumer surplus indicated in Table 5 are uncorrected for integration errors, which would be especially pronounced for all cases of Service Change B.

TABLE 4 Forecast Transit Ridership Gains due to Service Changes*

With Composite Disutilities:

Scenario	Service Change	
	Mild (A)	Aggressive (B)
Current Activity (1)	2122	9185
Twice Current (2)	5212	20877

Without Composite Disutilities:

Scenario	Service Change	
	Mild (A)	Aggressive (B)
Current Activity (1)	2021	8069
Twice Current (2)	5045	19499

Percent Difference:

Scenario	Service Change	
	Mild (A)	Aggressive (B)
Current Activity (1)	5.0%	13.8%
Twice Current (2)	3.3%	7.1%

*Units are system riders.

TABLE 5 Effect of Composite Disutilities on Net Consumer Surplus*

	Service Change	
Scenario	Mild (A)	Aggressive (B)
Current Activity (1)	45450	154275
Twice Current (2)	112138	372575

Without Composite Disutilities:

	Service Change	
Scenario	Mild (A)	Aggressive (B)
Current Activity (1)	44219	138214
Twice Current (2)	104458	339537

Percent Differences:

	Service Change	
Scenario	Mild (A)	Aggressive (B)
Current Activity (1)	2.8%	11.6%
Twice Current (2)	7.4%	9.7%

*Units are minutes of riding time.

Highway-Related Consumer Surplus

Referring to Table 3 (Scenario 2, Service Change B), one observes that the net consumer surplus from highway users is positive and is about 10 percent of the net consumer surplus from transit users. This component of net consumer surplus is almost entirely the result of congestion relief. The highway net consumer surpluses for both service changes for Scenario 1, with little congestion, are negligible.

Discussion of Findings

The tested service changes applied to the whole transit system, so shifts in activity levels occurred, essentially, between nonserved and served areas. A major transit alternative that upgrades service in a single corridor could prompt a relatively larger redistribution. The effects of the redistribution on net consumer surplus could be substantial.

The model is sensitive only to service changes that affect the disutility function (Equation 1). For activity redistribution to occur, there must be measurable advantages for a traveler. For example, the model will not redistribute activity when a bus line is replaced by a light rail line that operates at the same speeds and headways.

Wausau, the test city, is small but exhibits many of the same characteristics as larger cities. It would not be possible to extrapolate specific numbers to larger cities; however, the general trends discussed would still hold.

CONCLUSIONS

It is possible to build a travel forecasting model that finds a joint equilibrium solution between activity allocation, mode split, trip distribution, and traffic assignment. It is possible for such a model to include all major aspects of spatial variations in travel and still find an internally consistent solution.

When measuring net consumer surplus of transit alternatives, it is important to observe computational requirements. There should be sufficient equilibrium iterations to eliminate biases from convergence error. Furthermore, for big service changes, there should be sufficient slices in the integration of net consumer surplus to avoid a substantial overestimate of net consumer surplus.

Given the assumptions of the forecasting model, which represent a consensus of the transportation planning literature, activity redistribution increases net consumer surplus of transit service changes. The increase can be large enough to affect decisions regarding aggressive service improvements, especially those in well-defined corridors.

There is no evidence, either for highway users or transit users, that net consumer surplus would be negative (or even significantly retarded) for any reasonable service change.

ACKNOWLEDGMENTS

Portions of the research reported in this paper were funded by the Federal Transit Administration, University Research Program. This paper incorporates valuable suggestions from Edward Beimborn of the University of Wisconsin–Milwaukee and from four anonymous referees.

REFERENCES

1. Beimborn, E., A. Horowitz, J. Schuetz, and G. Zejun. *Measurement of Transit Benefits.* Report WI-11-0013, FTA, June 1993.
2. Mogridge, M. J. H., D. J. Holden, J. Bird, and G. C. Terzis. The Downs-Thomson Paradox and the Transport Planning Process. *International Journal of Transport Economics,* Vol. 14, 1987, pp. 283–311.
3. Horowitz, A. J. Subjective Value of Time in Bus Transit Travel. *Transportation,* Vol. 10, 1981, pp. 149–164.
4. Horowitz, A. J. The Subjective Value of Time Spent in Travel. *Transportation Research,* Vol. 12 1979, pp. 385 –393.
5. Horowitz, A. J. Assessing Transportation User Benefits with Maximum Trip Lengths. *Transportation Planning and Technology,* Vol. 6, 1980, pp. 175–182.
6. Hotelling, H. The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates. *Ecomometrica,* Vol. 6, 1938, pp. 242–269.
7. Williams, H. C. W. L., et al. Transport Policy Appraisal with Equilibrium Models I: Generated Traffic and Highway Investment Benefits. *Transportation Research B,* Vol. 25B, 1991, pp. 253–279.
8. Lowry, I. S. *A Model of Metropolis.* RM-4035-RC, The Rand Corporation, 1964.

9. Garin, R. A. A Matrix Formulation of the Lowry Model for Intrametropolitan Activity Allocation. *AIP Journal,* Vol. 32, Nov. 1966, pp. 361–365.
10. Horowitz, A. J. Subarea Focusing with Combined Models of Spatial Interaction and Equilibrium Assignment. In *Transportation Research Record 1285,* TRB, National Research Council, Washington, D.C., 1990, pp. 1–8.
11. Horowitz, A. J. Convergence of Certain Traffic/Land-Use Equilibrium Assignment Models. *Environment and Planning A,* Vol. 23, 1991, pp. 371–383.
12. Evans S. P. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research,* Vol. 10, 1976, pp. 37–57.
13. Horowitz, A. J. Convergence Properties of Some Iterative Traffic Assignment Algorithms. In *Transportation Research Record 1220,* TRB, National Research Council, Washington, D.C., 1989, pp. 21–27.
14. Powell, W. B., and Y. Sheffi. The Convergence of Equilibrium Algorithms and Predetermined Step Sizes. *Transportation Science,* Vol. 16, 1982, pp. 45–55.
15. Williams, H. C. W. L. On the Formation of Travel Demand Models and Economic Evaluation Measures of Users Benefit. *Environment and Planning A,* Vol. 9, 1977, pp. 284–344.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Transportation Policy Analysis Using a Combined Model of Travel Choice

MAYA R. TATINENI, MARY R. LUPA, DEAN B. ENGLUND, AND DAVID E. BOYCE

A combined model of travel demand is introduced to analyze the effects of new transportation policies on travel patterns. A brief discussion of the need for new forecasting procedures is followed by a description of the combined model's formulation. The procedure used to solve the combined model (i.e., to forecast travel demand) is then compared with a sequential modeling process currently used by planning agencies. Three scenarios representing possible policy results are analyzed in terms of their effects on different travel and network-related variables. The scenarios studied represent changes in transit fares, fuel costs, and land use changes that consider dispersed employment locations. Finally, research possibilities that could enhance the applicability of the combined model for various policy analyses are examined.

As a result of the 1990 Clean Air Act Amendments, greater attention is being focused on transportation policies for mitigating congestion and reducing total vehicle kilometers (miles) of travel. To design and evaluate alternative policies that will influence travel choices in the desired direction, it is necessary to model the effects of these policies in an accurate and consistent manner.

Transportation policy analysis traditionally has been performed using a sequential travel forecasting procedure, which involves the application of models for trip generation, trip distribution, mode split, and trip assignment. Increasingly, this procedure has proven deficient in several respects:

- The values of the variables used in the different models are inconsistent; in particular, travel times and costs used in the trip distribution and mode split models are not equal to those times and costs determined in the solution of the trip assignment model.
- The basis for forecasting travel choices, as defined in terms of variables and parameters, is inconsistent across the several models; for example, trip assignment is often based on travel times only, whereas mode split is based on a weighted combination of travel times and operating costs and fares.
- Evaluation of alternative policies using the sequential modeling process is complex and time consuming in that it is often not possible to produce results easily in the time frame decision makers desire. Moreover, the effects of these policies on travel patterns may not be clearly visible because of inconsistencies in the models.

The need for a new generation of forecasting procedures that take into account these deficiencies and provide a better basis for estimating travel choice has been expressed by various practitioners in the field (*1*). The purpose of this paper is to illustrate the application of a forecasting procedure that avoids these difficulties by combining the trip distribution, mode split, and trip assignment models into a single model and solution procedure. In this "combined" model, the same travel choice principles and relationships are incorporated as are used in the sequential procedure. By solving the problem as one model, however, the inconsistencies of the sequential procedure are eliminated, and model results are much easier to obtain.

In the paper, the formulation of a combined model and the procedure used to solve it are described. The results of solving the model for three scenarios representative of possible policy changes are then analyzed in terms of their impact on travel patterns and network-related variables. Finally, extensions of the model needed to facilitate its application in practice are discussed.

M.R. Tatineni and D.E. Boyce, Urban Transportation Center, University of Illinois at Chicago, 1033 West Van Buren Street, Suite 700 South, Chicago, Ill. 60607-9940. M.R. Lupa and D.B. Englund, Chicago Area Transportation Study, 300 West Adams Street, Chicago, Ill. 60606.

DESCRIPTION OF MODEL AND SOLUTION PROCEDURE

Combined models are based on the assumption that travel choices should result in equilibrium or user optimal conditions of travel. Wardrop (*2*) defined equilibrium route choice conditions such that "the journey times on all the routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route."

The concept of equilibrium has been widely accepted by practitioners with regard to traffic assignment or route choice. That is, given the number of trips between different origin-destination pairs, the resulting traffic is assigned iteratively to multiple sets of links or routes until all used routes between each origin-destination pair have equal travel times.

Most equilibrium or iterative traffic assignment models used in practice are applied to a given trip table that may be the result of earlier modeling steps involving trip distribution and mode choice. However, trip distribution (or destination choice) models and mode choice models also incorporate the interzonal travel costs as an important variable in determining travel choices. Because these models are applied before the traffic assignment model, the travel costs assumed for trip distribution and mode choice may in fact be quite different from the travel costs that result from the traffic assignment step. To be consistent, the travel choices that result from the route choice model must be equal to the travel costs used in the trip distribution and mode choice models.

To model true equilibrium conditions of travel, it is necessary, therefore, to feed back the travel costs that are determined as a result of the traffic assignment model to the trip distribution and mode choice models. The models must then be solved iteratively until the costs that are used to model trip distribution and mode choice are equal to the costs that result from the traffic assignment (or route choice) model. Such iterative procedures using the sequential models are rarely done in practice. The process itself can be time consuming, and the use of different variables to express travel costs in

the different models leads to inconsistencies that make it harder to reach a solution with the desired properties.

In the combined models, the problems associated with the sequential models are overcome by

- Considering a common cost function to model the various travel choices, and
- Using an iterative solution procedure so that the final equilibrium travel conditions are a result of the destination, mode, and route choices, instead of just route choice.

In a combined model of travel choice, a trip maker's choices regarding destination, mode, and route are considered as part of a single decision-making process with user cost minimization the main criterion. The model is formulated as a minimization problem, with the objective function representing a generalized cost that is a weighted sum of the travel times and monetary costs associated with a trip. The objective function is subject to constraints with regard to nonnegativity of flow and flow conservation, in terms of trip origins, destinations, and route flows. Furthermore, in order to consider dispersion of choices from a strictly cost minimizing behavior, which might occur either because of imperfect knowledge on the part of the user or because of consideration of other factors, such as convenience and comfort, that are not accounted for in the model, a choice dispersion constraint is introduced. The dispersion constraint is derived from an entropy function that originally was defined as a measure of dispersion in information theory (*3*).

The equivalent optimization problem is to minimize the following expression:

$$\frac{R}{N}\gamma_1 \sum_a \int_0^{v_a} c_a(x)dx + \frac{1}{N}\gamma_2 \sum_a \int_0^{v_a} k_a(x)dx + \gamma_3 \sum_i \sum_j P_{ijh} w_{ijh} + \gamma_1 \sum_i \sum_j P_{ijt} c_{ijt} + \gamma_2 \sum_i \sum_j P_{ijt} k_{ijt} + \gamma_3 \sum_i \sum_j P_{ijt} w_{ijt} + \gamma_4 \sum_i \sum_j P_{ijt} \quad (1)$$

The constraints are as follows:

$$\sum_{r \in R_{ij}} f_r = P_{ijh} N/R + T_{ij} \quad (2)$$

$$\sum_j \sum_m P_{ijm} = \overline{P}_i \quad (3)$$

$$\sum_i \sum_m P_{ijm} = \overline{P}_j \quad (4)$$

$$-\sum_i \sum_j \sum_m P_{ijm} \ln \frac{P_{ijm}}{\overline{P}_i \overline{P}_j} \geq S \quad (5)$$

$$f_r \geq 0 \quad (6)$$

where

$$v_a = \sum_i \sum_j \sum_{r \in R_{ij}} f_r \delta_r^a \quad (7)$$

The terms used in the preceding expressions are defined as follows:

N = total number of trips/hr,
v_a = total flow of vehicles on Link a,
c_a = in-vehicle travel time function that increases with link flow,
k_a = automobile operating cost function that increases with link flow,
R = automobile occupancy factor (persons/vehicle),
R_{ij} = set of highway routes between Zones i and j,
T_{ij} = number of truck trips/hr in automobile equivalent units,
f_r = total vehicle flow on Route r (vehicles/hr),
δ_r^a = 1 if Link a belongs to Route r and 0 otherwise,
$\overline{P}_i$ = fixed proportion of trips originating from Zone i,
$\overline{P}_j$ = fixed proportion of trips terminating at Zone j,
P_{ijm} = proportion of person trips by Mode m between Zones i and j,
P_{ijh} = proportion of automobile person trips between Zones i and j,
P_{ijt} = proportion of transit person trips between Zones i and j,
w_{ijh} = out-of-vehicle travel time for automobile trips between Zones i and j,
w_{ijt} = out-of-vehicle travel time for transit trips between Zones i and j,
c_{ijt} = in-vehicle transit travel time between Zones i and j,
k_{ijt} = transit fare between Zones i and j, and
S = observed dispersion of choices.

γ_1, γ_2, and γ_3 are the weights for the three cost components considered: in-vehicle travel time, operating cost or fare, and out-of-vehicle travel time; γ_4 is the estimated transit bias. This last term in the objective function is considered part of the transit cost. These weights are found by calibrating the model to represent the relative importance of the associated travel costs in determining travel choices for a particular region. Operating costs are expressed in terms of the average cost per vehicle, whereas times are stated per person.

The solution to the minimization problem defined above results in traffic flow conditions that are equivalent to the equilibrium flow conditions as defined by Wardrop. The equation to determine the interzonal automobile costs, u_{ij}, is derived from the optimality conditions of the model as a weighted sum of the link flow dependent travel times and operating costs and may be written as

$$u_{ij} = \gamma_1 \sum_a c_a(v_a)\delta_r^a + \frac{\gamma_2}{R} \sum_a k_a(v_a)\delta_r^a \quad (8)$$

For routes that are not used for a given zone pair, the cost on that route will be not less than u_{ij}, in accordance with Wardrop's definition. This result may also be derived directly from the optimality conditions of the model.

The generalized cost of travel by automobile between a zone pair, *i-j*, is a weighted sum of the travel time, operating cost, and parking cost associated with the trip and is given by the following equation:

$$c_{ijh} = u_{ij} + \frac{\gamma_2}{R} p_j + \frac{\gamma_3}{R} w_{ijh} \quad (9)$$

The generalized cost of travel by transit is given by a similar equation as

$$c_{ijt} = \gamma_1 c_{ijt} + \gamma_2 k_{ijt} + \gamma_3 w_{ijt} + \gamma_4 \quad (10)$$

Travel times and costs for transit are taken from a fixed matrix of travel times and fares on the basis of the minimum paths between each zone pair. The equation to determine the proportion of trips between Zones i and j by Mode m is also derived from the optimality conditions as a function of these generalized costs:

$$P_{ijm} = A_i \bar{P}_i B_j \bar{P}_j \exp(-\mu c_{ijm}) \quad (11)$$

where c_{ijm} is the generalized cost of travel between Zones i and j by Mode m. We may interpret A_i and B_j as the balancing factors for trip productions and attractions to satisfy the constraints on flow conservation defined on $\bar{P}_i$, the fixed proportion of trips originating from Zone i, and $\bar{P}_j$, the fixed proportion of trips terminating at Zone j. Thus, the origin-destination and modal choices are specified as direct functions of the proportion of trips leaving origin zones and entering destination zones and inverse functions of the interzonal generalized mode costs. The equations for the interzonal trip costs and proportions, as derived from the optimality conditions of the model, are used to find a solution to the model in an iterative procedure discussed below.

The Evans algorithm (*4*), based on the partial linearization technique, is used to solve the model. The steps involved in using the Evans algorithm to find a solution to the combined model may be summarized as follows:

Step 0 (initialization). Find initial trip proportions P_{ijm} and link flows v_a using an all-or-nothing assignment based on zero flow link costs.

Step 1. Update link costs on the basis of the new flows, v_a.

Step 2. Find new minimal-cost routes on the basis of costs from Step 1 and compute new generalized costs on the basis of these routes.

Step 3. Find the feasible descent direction as follows. First, compute new travel demands or trip proportions, Q_{ijm}, using the new generalized cost values:

$$Q_{ijm} = A_i \bar{P}_i B_j \bar{P}_j \exp(-\mu C_{ijm}) \quad (12)$$

and solve for A_i and B_j. Second, compute link flows, w_a, by assigning these new travel demands to new minimal-cost routes.

Step 4. Conduct line search; find an optimal step size λ such that if x represents the current solution (P_{ijm} and v_a) and y represents the subproblem or new solution (Q_{ijm} and w_a), then $x' = x + \lambda(y - x)$ minimizes the objective function value.

Step 5. Update the trip proportions, P_{ijm}, and link flows, v_a, using the step size λ such that $P'_{ijm} = (1 - \lambda)P_{ijm} + \lambda Q_{ijm}$ and $v'_a = (1 - \lambda)v_a + \lambda w_a$. The costs based on the updated link flows are then used to find a new subproblem solution at Step 2. Steps 2 to 5 are repeated until a stipulated convergence criterion is satisfied. About 20 iterations of the algorithm were sufficient to find a solution within 0.5 percent of the true optimal solution corresponding to the desired equilibrium of origin-destination, mode, and route choice.

In the equilibrium traffic assignment model used in the sequential modeling process, the link flows are assigned iteratively until travel costs for all used routes between a given zone pair are approximately equal. However, no attempt is made to update the trip matrices for the new travel costs. In the combined model, the trip matrices are updated for every iteration of the link flow assignment by calculating new trip proportions, Q_{ijm}, corresponding to updated link flows, w_a. Thus, the combined model solves a larger problem than the equilibrium traffic assignment model. Whereas the equilibrium traffic assignment model solves for the equilibrium travel conditions by reassigning only the link flows, the combined model solves for the same equilibrium conditions by both reassigning the link flows and revising the corresponding trip matrices. The solution procedure used in the combined model is compared with the sequential solution procedure in Figure 1.

The first formulation of a combined model was made in 1956 by Beckmann et al. (*5*), about the same time that the sequential procedure was conceived. This kind of formulation was specialized for the trip distribution model that was used by Evans in the sequential procedure in 1973. Evans proposed an algorithm for solving the model as well as proving that the solution converges to the desired conditions outlined above. The combined model, including trip distribution, mode, and route choice, was first implemented in 1982 on a network of realistic size for the Chicago region by Boyce et al. (*6*). Development and implementation of similar models for the northeastern Illinois region, based on a sketch planning network and zone system, have been the subject of ongoing research efforts involving the staff and faculty of the University of Illinois at Chicago and the Chicago Area Transportation Study. This paper is based in part on a report by Boyce et al. (*7*) [see also Boyce et al. (*8*)].

The data used for the analysis reported here are for 1980 from the Chicago region. A sketch planning or aggregated zone system and

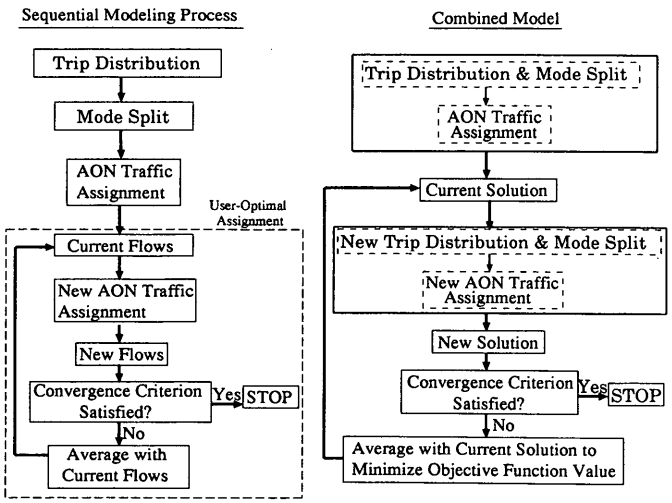


FIGURE 1 Comparison of solution procedures.

network was used in the analysis. The zone system, which includes 317 zones, is shown in Figure 2. The highway network has 2,902 links. The transit network is represented by a fixed matrix of travel times and fares.

DESCRIPTION OF SCENARIOS

Three policy changes represented by varying transit fares, fuel costs, and land use densities are considered. The monetary cost of travel is an important factor influencing travel choices. Both transit fares and automobile fuel costs may vary as a result of policy changes. To analyze the effects of those costs on travel patterns, both transit fares and gasoline prices are varied by multiplying the base year values by factors ranging from 0.25 to 3.0. The change in gasoline prices will affect only the cost of trips by automobile and does not affect transit trip costs.

Changes in land use variables are restricted to the consideration of changes in employment location. Work trip destinations are used to represent the availability of employment in different areas. The scenario examined here is the relocation of employment from the central business district (CBD) to the suburbs. Thus, work trip destinations in the CBD zone in the base data are reallocated to 11 suburban zones. The number of trips redistributed from the CBD varies from 0 to 44,000 out of the total 139,000 CBD destinations in the peak hour.

Six measures are selected to evaluate each scenario: mode choice, average trip length, total and congested vehicle kilometers (miles) of travel, average travel time, and average generalized cost of travel. Congested vehicle kilometers (miles) are the total vehicle kilometers (miles) on all links with flow exceeding capacity. Each of these scenarios is analyzed in detail in the following sections.

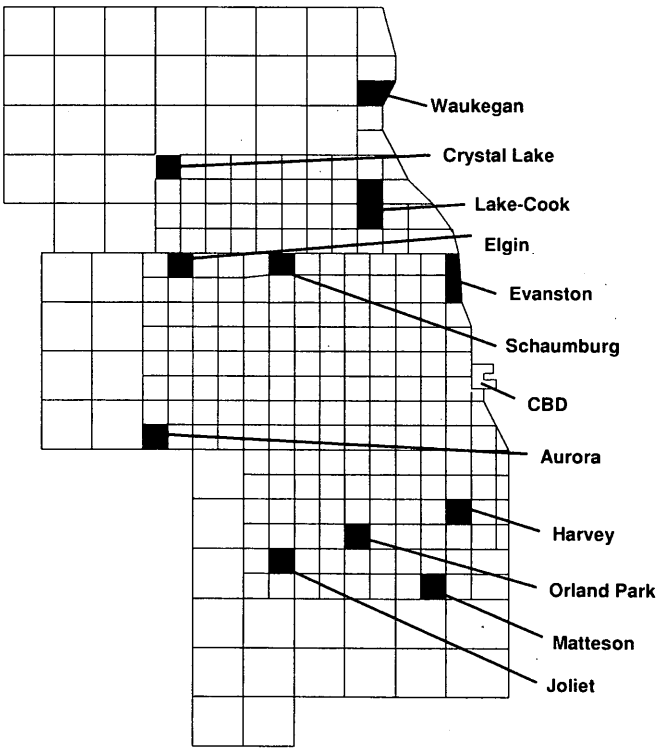


FIGURE 2 Location of the Chicago region's regional work centers.

EFFECTS OF VARYING TRANSIT FARES

The effects on each of the output variables considered of varying transit fares are shown in Figures 3 and 4 and may be summarized as follows.

Mode Choice

As transit fares increase, the relative attractiveness of transit compared with the automobile decreases. Thus, we find a decrease in the proportion of trips by transit and a corresponding increase in the proportion of trips by automobile. The effect is nonlinear, however, because the increase in automobile trips also results in increased automobile travel times, thereby reducing the attractiveness of the automobile.

Trip Length

As shown in Figure 3, an increase in transit fares is marked by a decrease in the average trip length for both modes, transit and automobile. Therefore, as transit fares are increased from low to higher values, there is a tendency for shorter trips to shift to automobile, resulting in a decrease in the average trip length for automobile. Increased costs associated with transit trips lead to a decrease in the average trip length for this mode as well. In addition, as the highway network becomes congested, the average length of transit trips reaches a stable minimum of about 14.5 km (8.7 mi).

Travel Time

The decrease in the average trip length for transit is accompanied by a decrease in the average travel time. On the other hand, as the number of trips by automobile increases as a result of the shifting of some trips from transit, there is an increase in the number of automobiles on the network to accommodate these trips, resulting in a reduction in the average speed on the network. Thus, slower speeds on the network result in an increase in the average travel time for automobile trips, despite a decrease in the average trip length.

Congested Vehicle Kilometers (Miles)

As explained before, the increase in the number of automobile trips results in an increased number of automobiles on the network, leading to an increase in congested vehicle kilometers (miles) of travel.

Generalized Costs

The average generalized cost increases for both transit and automobile. In the case of the automobile, the increase may be attributed to longer travel times. For transit, although there is a decrease in the average travel time, the increase in transit fares results in an increase in the average generalized cost.

EFFECTS OF VARYING FUEL PRICES

The changes in trip characteristics due to an increase in fuel prices are shown in Figures 5 and 6. These effects may be summarized as follows.

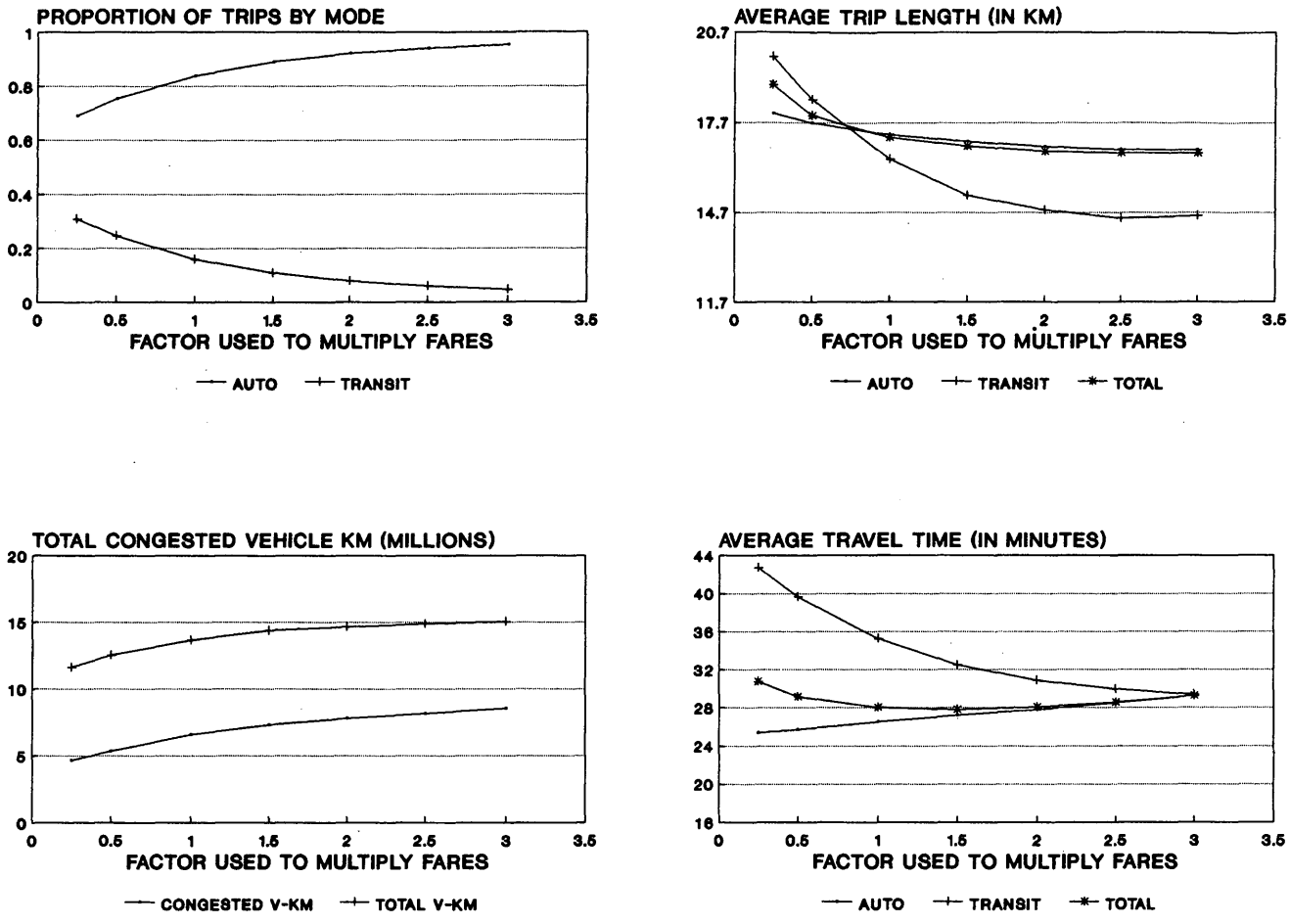


FIGURE 3 Effect of varying transit fares.

Mode Choice

As in the case of transit fares, an increase in fuel prices reduces the tendency to make trips by this mode and, accordingly, a decrease in the proportion of automobile trips may be observed. A corresponding increase in the proportion of transit trips may also be noted.

Trip Length

An increase in fuel prices should suppress longer trips, and this result is reflected in the decrease in the average trip length for the automobile. However, there is a corresponding increase in the average trip length for transit, indicating that increasing fuel prices shifts longer trips to transit.

Travel Time

The increase in the average trip length for transit and the decrease in average trip length for automobile are associated with longer average travel times for transit and shorter ones for the automobile. Whereas the increase in transit travel times may be attributed to the increase in transit trip lengths only, in the case of the automobile, shorter travel times also result from an increase in speeds on the network as congestion decreases because of fewer automobile trips.

Congested Vehicle Kilometers (Miles)

As fuel prices increase, there is a corresponding decrease in automobile trips, which results in fewer automobiles on the network, thus decreasing congested vehicle kilometers (miles) of travel.

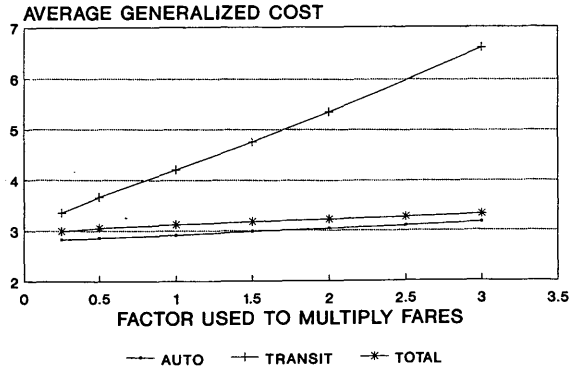


FIGURE 4 Effect of transit fares on generalized cost.

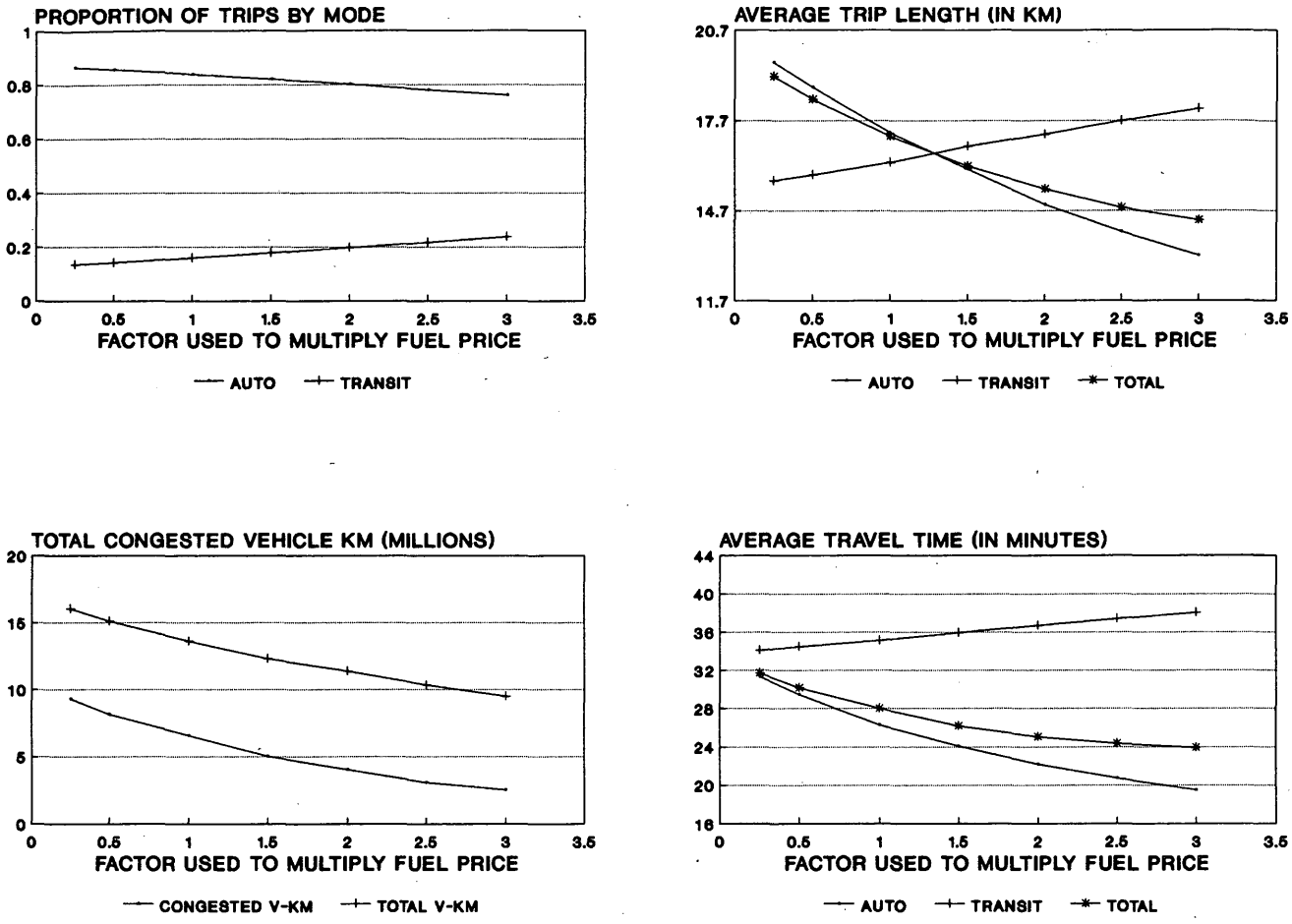


FIGURE 5 Effect of varying fuel prices.

Generalized Cost

An increase in fuel prices results in an increase in the average generalized cost for both modes considered. In the case of the automobile, the increase in fuel prices contributes to the increase in the average generalized cost, whereas the increase in the generalized cost associated with transit results from the increase in the average travel time for transit trips.

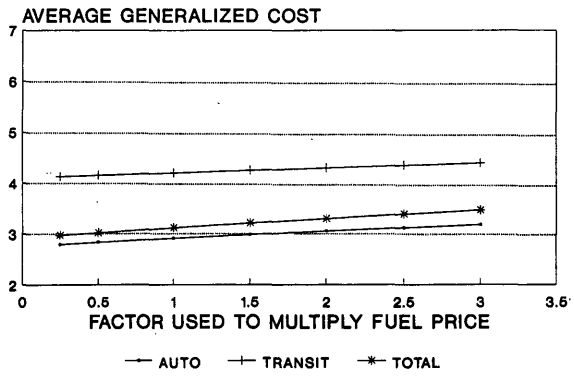


FIGURE 6 Effect of varying fuel prices on generalized cost.

EFFECTS OF VARYING EMPLOYMENT LOCATION

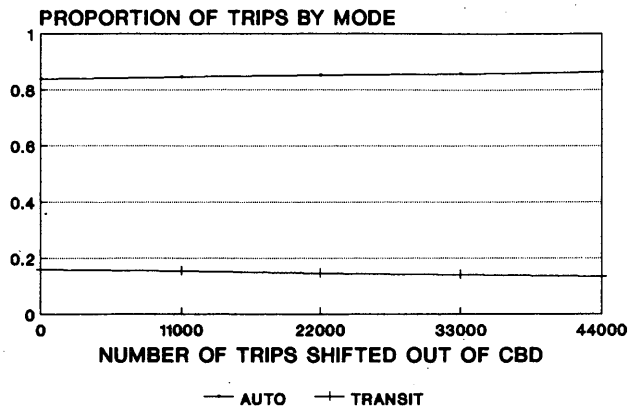
Locations of the regional centers to which trip destinations are shifted from the CBD are shown in Figure 2. The effects on travel choices of moving jobs to the suburbs are shown in Figures 7 and 8 and are discussed below.

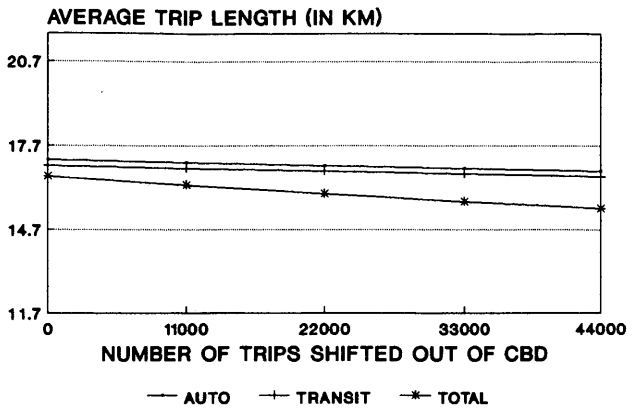
Mode Choice

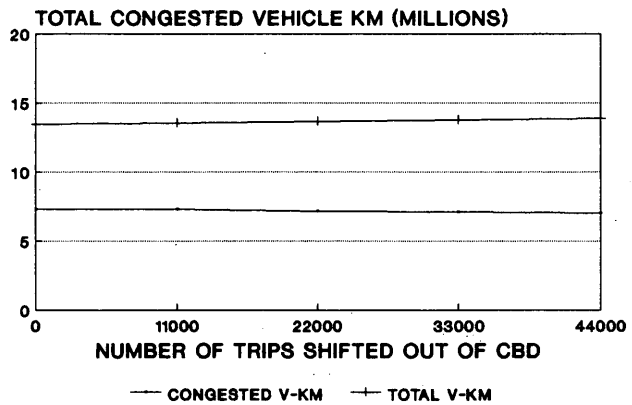
The decrease in transit trips may be attributed to the fact that workers employed outside the CBD are unable to use transit for their trip to work because the transit network for the Chicago region is designed to serve suburb-to-CBD trips rather than suburb-to-suburb trips. There is a corresponding increase in automobile trips.

Trip Length

Shifting employment from the CBD to suburban regional centers is marked by a decrease in the average trip length for both modes. In the suburban scenario, workers drive to work sites closer to their homes, thus decreasing the average trip length for both automobile and transit. Some of the long transit trips from the suburbs to the







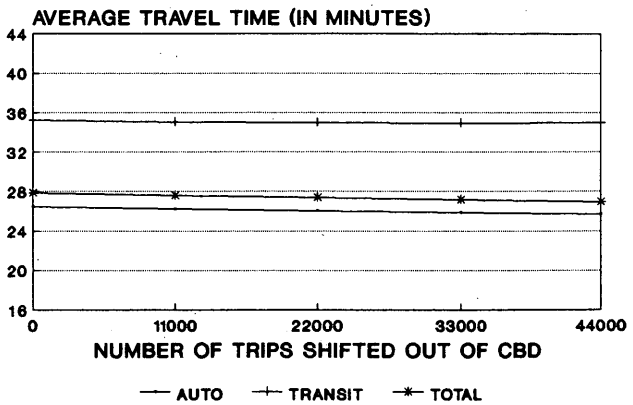


FIGURE 7 Effect of varying employment location.

CBD shift to automobile, resulting in a substantial decrease in transit trip length.

Travel Time

The increase in automobiles on the network is offset somewhat by the shorter distances those automobiles travel. Therefore, the average travel time for both automobile and transit decreases. The decrease in average travel time for transit is also a result of fewer workers bound for the CBD.

Congested Vehicle Kilometers (Miles)

Total vehicle kilometers (miles) traveled increase because of the larger percentage of automobiles on the network overall. However, congested vehicle kilometers (miles) decrease.

Generalized Cost

There is a decrease in the average generalized cost for both automobile and transit, which may be attributed in both cases to shorter work trips (i.e., decreased values of both travel times and monetary costs).

CONCLUSIONS AND SUGGESTIONS FOR FUTURE RESEARCH

The internal consistency with which travel costs and choices are determined in the combined model provides a better basis to analyze the effects and cross effects of varying costs on travel patterns. Moreover, the ease of applicability of the model enabled this analysis to be completed in a relatively short time. Much of the work in

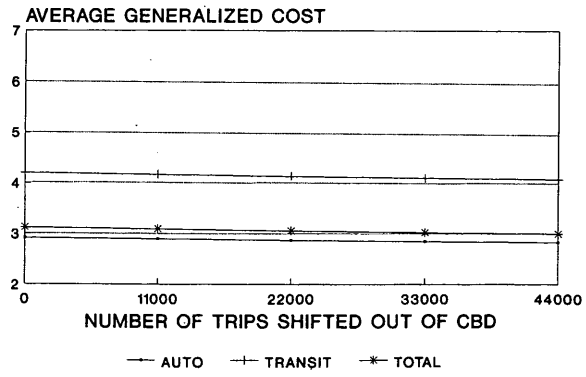


FIGURE 8 Effect of employment location on generalized cost.

developing the model, and subsequent solutions for calibration and analysis work with the model, was done using a CRAY supercomputer. In recent months, however, the model has been solved on a Sun SPARCstation 10 with 64 megabytes of memory at the Chicago Area Transportation Study (CATS). Solving the model at CATS for 20 iterations takes 45 min (i.e., 2.25 min per iteration). There is no doubt that rapid advances in desktop computing technology will make it much easier to use such models in the future.

There are still many possible extensions to the model in its present form that would widen its applicability. For instance, the model could be revised to enable one to predict variations in the overall trip rate and average automobile occupancy. Dispersion of travel choices from strictly cost-minimizing behavior occurs either because of differences in the perceived values of cost/time or consideration of factors not accounted for in the modeling process. Although the model in its present form accounts for choice dispersion in location and mode choices, further research could enable consideration of a similar dispersion measure for route choice.

At the present time, the model is solved by directly executing a source code written in FORTRAN. Further coding work, however, could make this model much more user friendly. Indeed, that should be one of the first steps taken to enhance the model's applicability in practical planning analyses. Another approach would be to solve the model using transportation planning software, such as the EMME/2 system.

It is clear that the planning profession needs to take a close look at present modeling methods and revise them to be more consistent. Obviously, the travel choice process does not consist of separate decisions with regard to destination, mode, and route. The interdependency of these choices, and the common costs considered, must be reflected in the modeling process. Planning agencies must incorporate relevant research findings into their modeling process to allow the models to represent more closely traveler response to real-life policy changes.

ACKNOWLEDGMENTS

Computational support for a portion of this research was provided through a grant from the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. The support of FHWA, through the Illinois Department of Transportation, is gratefully acknowledged. Support for much of the research, in terms of staff time and computer support, was provided by CATS.

REFERENCES

1. Weiner, E. Upgrading U.S. Travel Demand Forecasting Capabilities. *The Urban Transportation Monitor,* July 9, 1993.
2. Wardrop, J. G. Some Theoretical Aspects of Road Traffic Research. *Proc., Institute of Civil Engineering,* Part II, Vol. 1, 1952, pp. 325–378.
3. Shannon, C. E., and W. Weaver. *The Mathematical Theory of Communication.* University of Illinois Press, Urbana-Champaign, 1949.
4. Evans, S. P. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research,* Vol. 10, 1976, pp. 37–57.
5. Beckmann, M., C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation.* Yale University Press, New Haven, Conn., 1956.
6. Boyce, D. E., K. S. Chon, and R. W . Eash. Development of a Family of Sketch Planning Models. *CATS Research News,* Chicago, Ill. 1982.
7. Boyce, D. E., M. Tatineni, and Y. Zhang. *Scenario Analyses for the Chicago Region with a Sketch Planning Model of Origin-Destination, Mode and Route Choice.* University of Illinois, Chicago, 1992.
8. Boyce, D. E., K. S. Chon, M. E. Ferris, Y. J. Lee, K-T. Lin, and R. W. Eash. *Implementation and Evaluation of Combined Models of Urban Travel and Location on a Sketch Planning Network.* University of Illinois at Urbana-Champaign and Chicago Area Transportation Study, 1985.

Publication of this paper sponsored by Committee on Passenger Travel Demand Forecasting.

Critique of Metropolitan Planning Organizations' Capabilities for Modeling Transportation Control Measures in California

ROBERT A. JOHNSTON AND CAROLINE J. RODIER

For each class of transportation control measures (TCMs), the relevant travel behaviors expected to change are identified and techniques for simulating these changes are listed. Then, the latest round of analysis of TCMs in each of the four largest urban regions in California is studied carefully to see whether the relevant behaviors were modeled in a credible fashion, on the basis of local data. In modeling TCMs that change travel time and costs or expand transit options, models were found to lack automobile ownership steps and accessibility variables in some steps. Intersection capacity and delay should be entered into the road networks, and the networks need to be more detailed. In addition, more cost data are needed. Household income should be retained in final trip tables to allow for equity evaluations of changes in travel patterns. In simulating policies that change land uses, walk and bicycle modes should be explicit, and better land use data are needed. For analysis of clean vehicle incentive programs, vehicle types should be linked to trip purposes. Most agencies did a poor job evaluating TCMs; in some cases, they did not even use their travel demand models but instead used spreadsheets with generalized default values. Many improvements are being made to these models, and practice will be improved.

The regional travel demand models of metropolitan planning organizations (MPOs) have been used in the past primarily for the undemanding task of projecting relative levels of traffic congestion or transit demand in urban corridors. The new federal Clean Air Act, however, now requires models that can project travel (and on-road mobile emissions) with absolute accuracy. Air quality plans in nonattainment regions must include transportation control measures (TCMs) and, for example, must reduce emissions of volatile organic compounds according to certain schedules (by 15 percent in 6 years or 9 percent in 3 years). Furthermore, the TCMs to be evaluated include pricing and land use measures, policies not traditionally modeled by most MPOs.

Because of the uneven quality of MPO models across the United States, and because of the incomplete and fragmented modeling regulations that have come from the Environmental Protection Agency (EPA) to date, MPOs have developed their own national guidelines for good modeling practice (*1*). Whereas that report and papers commenting on its drafts (*2*) consider regional models in general, examination of specific MPO models in the major California urban regions (the San Francisco Bay Area, Sacramento, Southern California, and San Diego) will help further understanding of how models need to be improved to simulate accurately the effects of TCMs on travel and emissions.

Institute of Transportation Studies, University of California at Davis, Davis, Calif. 95616.

We set forth an exhaustive list of desired modeling capabilities but believe that at least one MPO in the United States has realized them. We would not expect an MPO to develop all of these capabilities within the next few years, because of data limitations. However, we would expect MPOs to accelerate data collection for the next round of model development and at least attempt most of the recommended improvements. Under the Intermodal Surface Transportation Efficiency Act, many categories of funds for planning and model development are available including some types of project funds; thus, funding should not be a limitation in the future.

Apparently, environmental groups are poised to sue some of the large MPOs, at least partly on the basis of their modeling methods. Perhaps MPOs should work toward making their models close to the state of the art instead of merely acceptable according to EPA guidelines. Lawsuits can easily cost more than a major program of model development. Each MPO will have to weigh these matters considering its current models and data sets, and develop a model improvement work plan that suits its local needs. Our critique of existing models does not represent legal requirements; those are unique to each region.

Our modeling reviews were drafted in more detail than appears here and were reviewed by the agencies. We attempted to respond to staff members' concerns in every case but were often hampered because their written and oral reviews differed from the agency documents or those of other staff members. Turnover of staff also made it difficult to ensure the accuracy of some details, as did lack of documentation for some modeling exercises. In some cases, MPO reviews were antagonistic, because of past or threatened lawsuits. In all cases, we found agency staffs overworked and had to ask repeatedly for their assistance in reviewing the drafts. We have tried very hard to represent accurately the modeling practices of the MPOs.

We begin by categorizing TCMs into eight different classes and identifying the TCMs' likely behavioral effects and the model components needed to capture those effects. The categories and criteria are based on a selective review of the literature on modeling theory and practice (*3*) and on work by Harvey (*4*). Next, we discuss issues related to the criteria set out, including the magnitude of TCMs' effects, forecasting variables, the feasibility of the proposed improvements, and synergistic effects of TCMs. We then examine the MPOs' analyses of TCMs in their most recent round of transportation and air quality plans and compare their TCM analyses with our criteria to identify shortcomings. We also discuss improvements under way on their models and recommend additional improvements needed for better TCM modeling.

CATEGORIES OF TCMs AND CRITERIA FOR ACCURATE MODELING

Categories described in this section, their behavioral effects, as well as many modeling criteria, were informed by Harvey's report (*4*).

Change Travel Times

TCMs that alter travel times include high-occupancy vehicle (HOV) lanes, arterial operation improvements, preferential parking, and reduced transit wait times. These TCMs are designed to decrease travel times for high-occupancy modes or increase travel times for low-occupancy modes. The primary behavioral effect of these TCMs should be mode shifts. But they may also result in reduced automobile ownership, fewer and shorter trips by automobile, and closer proximity of residential and work locations.

To capture the mode shift effects of these TCMs, a reliable mode choice model is needed, one that accurately represents congested and free-flow travel times, transit and automobile access times (e.g., walk and wait), and signal and intersection delays for all trip purposes.

Currently, many models represent access to transit only crudely (*5*). The representation of transit access times can be improved by incorporating variables such as proximity of work and housing to transit, bicycle and pedestrian conditions, and the location of park-and-ride lots (*5*). The representation of automobile access times can also be improved with increased sensitivity to parking capacity constraints (*5*).

Highway and transit travel times and costs or composite impedances should be represented in the trip distribution, trip generation, and automobile ownership steps (in addition to the mode choice step, as described above) to simulate changes in trip lengths, the number of trips made, and the number of cars owned by households. It is important that an endogenous automobile ownership step be included in the travel demand model because of the significant effect of automobile ownership on trip generation. The representation of accessibility in the automobile ownership step should include parking availability (*5*). All model steps should be fully iterated on impedances from assignment (i.e., congested impedances for peak models and uncongested impedances for nonpeak models).

If these TCMs result in large changes in accessibilities, even just for some subareas, a land allocation model that is fully iterated with the travel demand model can be used to simulate changes in the location of new employment and residential development.

Change Travel Costs

TCMs that alter travel costs include increased fuel taxes, pay-as-you-drive insurance, highway peak-period congestion fees, increased bridge tolls, parking fees, subsidized transit, ridesharing incentives, and vehicle purchase fees. These TCMs are designed to increase the monetary cost of traveling in single-occupancy vehicles and to decrease the cost of traveling in high occupancy modes. The primary result of these TCMs should be a shift in mode from single-occupancy vehicles to HOVs. In addition, these TCMs may result in fewer, shorter discretionary trips and time-of-travel shifts, particularly in the case of peak-period congestion pricing. However, reduced travel times resulting from mode shifts, reduced trip making, and time-of-day shifts may induce some single-occupant vehicles back onto facilities and thus offset some portion of the mode shift. Further, if pricing measures are imposed only on certain roadways, route shifts instead of mode shifts may occur. A secondary effect of large changes in travel pricing may be changes in employment and residential locations for existing and new land uses. Finally, issues of equity also need to be considered when evaluating travel-pricing TCMs.

The mode shift effects of these TCMs can be simulated with the use of a reliable mode choice model that accurately reflects changes in travel costs in composite impedances, as discussed above. Again, an endogenous automobile ownership step that is sensitive to travel costs (including parking costs) is needed to capture these TCMs' effects on automobile ownership levels and thus on trip generation. Generalized travel costs should also be included in the trip distribution and trip generation steps to better simulate changes in the number and length of trips made as a result of these TCMs. A departure time choice model that is sensitive to direct travel costs as well as time costs is needed to represent time-of-day shifts due to TCMs that impose additional monetary costs on peak-period travel, such as congestion pricing. To simulate these TCMs' effects on route choice, Harvey suggests that travel time components be "supplemented by a network assignment model capable of capturing the 'equilibrium' between price and time effects" (*4*). All model steps should be fully iterated on composite impedances from assignment.

Detailed pricing data in the base year data set must be available to properly specify the model's travel cost variables. Replogle suggests that the data should include information about "the share of employees getting free parking at individual sites or within compact zones, the cost of short and long term parking at individual sites or within compact zones, the cost of short and long term commercial parking, HOV pricing incentives and other commuter subsidies, as well as transit cost on an origin-destination basis (if appropriate by mode)" (*5*).

A data base or model that links vehicle types to trip categories is needed to project the emission effects of TCMs that increase costs for high-emitting vehicles (*4*).

For equity evaluation of TCMs that alter travel costs, household income classes should be retained in the final trip tables. This makes information related to the number of people by income class affected by a particular pricing policy readily available.

Again, if these TCMs result in large changes in accessibilities, a land allocation model that is fully iterated with the travel demand model can be used to simulate changes in the location of new employment and residential development.

Expand Transit Options

TCMs that expand travel options include, for example, improved access to bus and rail transit. These TCMs are designed to expand travel options by serving areas with new modes. The primary behavioral effect of these TCMs should be mode shifts; however, large changes in transit service may affect automobile ownership levels, trip lengths, and trip generation. Heavy rail (subway or commuter rail) may also alter new land development patterns.

To accurately capture the travel demand for new modes, mode choice models ideally should incorporate unobserved attributes, such as comfort and reliability, as explicit variables in mode choice,

in addition to travel time and cost. However, Harvey suggests that such variables can be difficult to quantify, and thus "conventional studies get around the problem by using observed shares to create a one-time set of adjustments"(*4*).

Because these TCMs also affect transit travel times and costs, composite impedances should be included in the mode choice, trip distribution, trip generation, and automobile ownership steps to represent changes in these behaviors. In addition, if TCMs result in large changes in accessibilities, a land allocation model should be used.

Change Land Uses

Some TCMs encompass a range of land development policies aimed at encouraging a more compact pattern of urban development coordinated with transit services and with improvements to walking and bicycling facilities. These TCMs may result in mode shifts, shorter trips, fewer automobile trips, and reduced automobile ownership.

Generally, walk, bicycle, and transit accessibility variables (i.e., measures of the walk and bicycle environment and transit travel time and cost) are needed in the mode choice, trip distribution, trip generation, and automobile ownership steps to simulate the effects of these TCMs (*5*). More specifically, the mode choice step should include explicit walk and bicycle modes as well as indices of zonal or discrete household-based bicycle and pedestrian "friendliness" to simulate mode shifts due to these TCMs (*5*). Further, to represent the diversion of short automobile trips to nonmotorized modes, a person-trip-based trip generation step should be used in which central business district (CBD) and other locational variables have been replaced with variables that represent nonmotorized access to retail and pedestrian and bicycle friendliness (*5*). Detailed networks and smaller zones can be used to improve representation of walk, bicycle, and transit accessibility (*5*). Model steps should be fully iterated on zone-to-zone travel impedances from assignment.

To properly specify walk, bicycle, and transit accessibility variables, Replogle suggests collecting "inventories of transportation supply, with information on road widths, number of lanes, presence of medians, intersection configurations, transit services including transit stop locations and service frequency, parking inventories including park-and-ride lots, location and character of sidewalks and bicycle paths and lanes, availability of secure bicycle parking spaces at transit stops, and other factors" (*5*).

If TCMs result in large changes in accessibilities, a land allocation model should be used.

Clean Vehicle Technology

These TCMs include vehicle technologies designed to reduce emissions, for example, technologies that change the internal combustion engine, electric vehicles, vehicle inspection and maintenance, new car standards, clean fuels, or retirement of high-polluting vehicles. Such TCMs are designed to alter the vehicle rather than travel behavior.

For TCMs that affect the entire fleet in a uniform manner, Harvey suggests that "emissions improvements can be calculated simply by substituting a revised set of composite emission factors" (*4*). Harvey has pointed out the difficulties in evaluating emission reductions for TCMs that affect only a portion of the fleet (*4*):

> There is a danger that the altered portion of the fleet will be used in a way that is not representative of the overall vehicle use pattern. Two clear examples come to mind: (1) conversion of a dedicated fleet to alcohol or electric propulsion might have a disproportionately small effect on emissions because so much of fleet VMT occurs in the hot stabilized operating regime; and (2) subsidized or mandated retirement of the oldest personal vehicles might have a disproportionately large effect on emissions because so much of the VMT of the old vehicles occurs in the cold and hot start modes. Simple adjustment of the fleet composite emissions factors would not accurately represent either of these changes.

Partial fleet changes should be evaluated with a model or data base that links vehicle types to trip categories in addition to having revised emission factors (*4*). If vehicle or fuel costs rise uniformly, these changes can be simulated, as discussed earlier with respect to TCMs that change travel costs.

Ease Activity Constraints

These TCMs attempt to reduce the place and time restrictions of work travel that force travelers to use limited transportation services. Examples of TCMs that ease activity constraints are flextime and telecommuting. The behavioral effects of these TCMs are highly complex; however, they should affect mode choice, departure time choice, trip making, and possibly automobile ownership.

Models of human activity scheduling behavior can capture the effects of flextime and telecommuting, but as yet these models exist only in experimental form (*4*). Without such models, the behavioral effects of these TCMs must be assessed by extrapolating from carefully interpreted case study data and manually adjusting mode choice projections (*4*), trip generation rates, and possibly automobile ownership rates. However, a time-of-day choice step that is included in the travel demand model can help simulate changes in travelers' choice of departure time resulting from flextime policies.

Promote Alternative Modes

TCMs that promote alternative modes are designed to educate travelers about their travel options and thus help them make more rational travel decisions. Such promotion can be very effective where modal choices are substitutable. These TCMs should result primarily in mode shifts.

Currently, it is very difficult for travel demand models to simulate the effects of promotional TCMs. Case studies, if carefully interpreted, can be used to manually adjust the mode choice projections (*4*).

Limit Travel Options

These TCMs are intended to reduce modal options (i.e., use of automobiles) either temporarily or in the long term and include, for example, fuel rationing and exclusion of single-occupant automobiles from key facilities. In the short term, these TCMs may result in large mode shifts, reduced trip making, and shorter trips. If enacted frequently or in the long term, these TCMs may result in changes in automobile ownership, and changes in new and existing residential and employment location might occur.

To reflect reduced modal options on key facilities, Harvey cites the need for a detailed network of freeways, arterials, and roads as well as a "mode choice model with a 'choice set' (i.e., range of alternatives) that can be adjusted to reflect limited availability" (*4*).

Because TCMs that limit travel options will increase the time and cost of automobile travel, composite impedances that reflect these increases should be represented in the mode choice, trip distribution, trip generation, and automobile ownership steps. If TCMs result in large changes in accessibilities, a land allocation model should be used.

TCM EFFECTS WARRANT MODEL IMPROVEMENTS

Wachs, in a comprehensive review of recent behavioral research in transportation demand management, found that there is clear evidence that travel time, out-of-pocket travel costs, and the comfort and reliability of travel modes have a significant effect on trip generation, mode choice, departure time choice, and route choice (*6*). Stopher summarizes the literature on the effects of capacity constraints (e.g., congestion, which increases the time costs of travel) on travel behavior and concludes that such constraints result in changes in new development, automobile ownership, trip making, the length of trips, mode choice, departure time choice, and route choice (*7*).

Bae, however, in an examination of transportation and land use measures included in Southern California's Air Quality Management Plan (particularly, alternative work schedules, mode shift strategies, and growth management), found that these measures were projected to have a relatively modest impact on reducing air pollution (*8*). It should be noted that Bae's examination made use of some weak sources. Bae suggests that clean vehicle technology and pricing TCMs are more effective alternatives. Cameron's study of pricing policies in Southern California found that pricing policies would have a significant effect on trip generation, VMT, and mode choice. The Transportation Incentive Planning System (TRIPS) travel demand model, which includes an endogenous automobile ownership step and composite travel costs throughout the model hierarchy, was used for this study (*9*).

In the end, however, transportation planners must use their own judgment as to whether the effects of particular TCMs in a particular region will be large enough to warrant the model improvements suggested here, particularly inclusion of composite impedances in the trip generation and automobile ownership steps and feedback to those steps and to a land allocation model.

FORECASTING TCM VARIABLES

Most of the variables at issue in this paper, (i.e., accessibility and demographic variables) are currently forecast in most regional travel demand models. Life-cycle stage variables, which have been shown to be significant in predicting travel demand, are less commonly forecast in regional travel demand models. However, the Portland, Oregon, and Montgomery County, Maryland, models have incorporated life-cycle variables (e.g., ages of household members). Forecasts of these variables are likely to be reasonable within a 20-year time frame. Because land use forecasts are subject to local political pressures, we advocate simulation of land use variables through land allocation models (i.e., development location choice models) to avoid political bias and improve accuracy.

FEASIBILITY OF PROPOSED CHANGES

Travel time and travel cost variables can be included throughout the chain of travel demand models. The original Metropolitan Transportation Commission models (1978) are an example of a set of regional travel demand models that have successfully incorporated composite impedances in the mode choice, trip distribution, trip generation, and automobile ownership steps. Land allocation models that are sensitive to transportation supply are available (e.g., the DRAM/EMPAL model); however, their sensitivity is limited (*7*).

Separate walk and bicycle modes can be added to mode choice models fairly easily. The difficulty arises in developing the travel times for these modes. Greatly increased network detail is needed to estimate travel times for short walk and bicycle trips (*7*). In the short term, however, rough approximations of walk and bicycle travel times can be derived from the roadway network. The integration of geographic information systems into travel demand models will assist in the development of the network detail needed for improved specification of walk, bicycle, and transit accessibilities. In the short term, however, zonal and discrete household-based walk, bicycle, and transit accessibility indexes have been incorporated effectively into some regional travel demand models, for example, that of Montgomery County, Maryland.

Currently, time-of-day choice modes are not generally included in regional travel demand models. Stopher states that "some form of time-of-day modeling can be developed to work within travel-forecasting procedures" (*7*). Portland, Oregon, and Sacramento, California, are incorporating explicit time choice components in their updated travel demand models (*1*).

Comfort and reliability variables are difficult to incorporate in regional travel demand models. However, academic models have successfully incorporated such variables (*10*). Wachs suggests the use of market segmentation to help clarify the relationship between attitudes and travel behavior (*6*).

Finally, the additions and extensions suggested in this paper require data that are not generally included in conventional data bases used to estimate and calibrate models. Conventional data bases should be expanded to obtain needed travel behavior data. Such estimation and calibration of model steps and of overall system models is more time-consuming than past practices. However, Portland has calibrated its socioeconomic/demographic models (i.e., worker, children, and automobile ownership models) and travel demand models (i.e., trip generation, destination choice, pre-mode choice, and mode choice models) to survey data and has calibrated its automobile assignment and transit assignment models to count data.

SYNERGISM

TCMs tend to be modeled separately instead of together as a package. However, some combinations of TCMs can increase or decrease the effectiveness of individual TCMs (*11*). The findings regarding potential synergistic effects were summarized as follows (*11*):

> In general, it was found that improvements in driving conditions work counter to efforts to shift commuters from their own cars onto public transit or to participate in ridesharing programs. Penalties associated with driving, on the other hand, support these efforts, as well as attempts to reduce overall travel by changing land uses and substituting communications for work trips. All transit improvement and incentive techniques are mutually supportive to a high degree. Carpooling, which in itself appears to be a moderately effective and inexpensive approach, does not blend well with many other approaches; efforts to reduce travel demand by changing land use, to spread peak commuting time, to provide transit alternatives, or to improve traffic flow through improvements to roadways all reduce the motivation for participating in prearranged ridesharing.

Thus, TCMs should be modeled in various packages, not separately, to capture synergistic effects and thereby avoid overestimating or underestimating the effects of TCMs.

PAST TCM MODELING PRACTICES IN FOUR REGIONS

This section and the next are based on a study performed for the California Energy Commission that reviews the MPOs' regional travel demand models and their modeling of TCMs (*3*). Agency documents and interviews were used to prepare these reports, and the reports were reviewed by the agencies for accuracy.

San Francisco Bay Area

The TRIPS model was the primary travel demand model used in the San Francisco Bay Area to evaluate TCMs. TRIPS was used to evaluate most TCMs within the travel cost category and some TCMs within the travel time category. Local data, empirical studies reported in the literature, and interviews with experts were used for categories of TCMs involving expanded travel options, travel time, land use changes, activity constraints, and promotion of alternative modes (*12,13*). TCMs involving walk and bicycle improvements were modeled with "a regional mode choice model developed by Deakin in the mid 1980's with bicycle and walk as explicit modes" (*12,13*). Traffic operations models, such as TRANSYT and NETSIM, were also used in the analysis (*13*).

The TRIPS model was derived from models originally developed for the Metropolitan Transportation Commission in the mid- to late 1970s; it incorporates transit and highway travel times and costs in all of its model steps, includes an automobile ownership step, and is fully iterated (*14*). TRIPS uses a sample of households from the most recent Bay Area travel survey, which preserves the variation in the distribution of population characteristics and thus produces more accurate travel demand predictions (*15*). Household totals are expanded to represent the larger population and summed in regional categories (*15*). TRIPS lacks a detailed network representation and traffic assignment component. Instead, as an approximation, a simple routing for estimating changes in level of service has been incorporated in the model. Thus, TRIPS achieves great detail in representing demand at the expense of detailed network representation (*14*).

Sacramento Region

As part of the regional mobility plan, the Sacramento region used its regional transportation demand model to evaluate parking pricing and new HOV lanes (*16,17*). Cumulative estimates of VMT and emission reductions due to the other TCMs included in the plan were derived from the results of TCMs modeled by other regions in California, particularly the Bay Area (*17*). Analyses of TCM effectiveness in the Bay Area and other areas cannot be transferred credibly to the Sacramento region, however, because of large differences in urban structure and transportation infrastructure, particularly modal options.

Southern California Region

For its 1992 Air Quality Management Plan, the Southern California region used its regional travel demand model to evaluate TCMs involving alternative work weeks, flextime, telecommuting, employer rideshare and transit incentives, parking management, vanpool purchase incentives, merchant transportation incentives, automobile use restrictions, new HOV facilities, and transit improvements, as part of its regional mobility plan (*18*). These strategies were modeled primarily through manual adjustments to the trip generation tables and mode choice model. In other words, each TCM was assumed to reduce single-occupant vehicle trips by a certain percentage, and trip generation rates and mode choice projections were adjusted to reflect that reduction (*18*). That method of modeling is not adequate because it begs the question of whether the TCMs will have their anticipated behavioral effects. Some strategies that involve pricing incentives were modeled correctly with sensitivity runs, which resulted in changes in mode choice (*3*).

The Southern California region modeled TCMs related to goods movement, traffic flow improvements, nonrecurrent congestion relief, airport ground access, and rail consolidation with elasticities obtained from the regional travel demand model and from elasticities reported in the literature (*3,18*). Elasticities that are used to adjust VMT or trips without running these changes through the model set will not represent the complete effects of the TCMs, however. Also, point elasticities obtained from the literature are valid only if they are used for the same ranges and starting points on the basis of which the elasticities were calculated.

San Diego Region

The San Diego Region evaluated its TCMs with the use of TCM Tools (*19*), a spreadsheet that aggregates the effects of TCMs at the regional level and uses input data obtained from expert judgment. The spreadsheet has default values for most outputs or uses point elasticities to produce outputs. The default values can be overridden with area-specific data obtained from a regional transportation model. The spreadsheet does not represent the effects of changes in congestion on travel. Most of the effects of land use changes and traffic flow improvements must be estimated apart from the spreadsheet. In general, the spreadsheet is primarily a screening tool and generally predicts the best, instead of the most likely, outcomes of TCMs (*15,20*).

The TCM Tools spreadsheet is acceptable as an accounting system for measuring the effects of TCMs only if it is used in conjunction with a fully run set of regional travel demand models and its default values are overridden with area-specific values obtained from the regional travel demand model.

Most default values were not overridden in the modeling of the San Diego region's TCMs. Small adjustments were made for some default values for the HOV and park-and-ride TCMs. Because the elasticities were so small and it was thought that area-specific values would not be much different from the default values, no area-specific adjustments were deemed necessary. In general, the San Diego region lacks data with which to develop area-specific values (*3*).

MPOs' POTENTIAL ABILITIES TO ANALYZE TCMs

Current Models

As described above, not all TCMs in the regions that should have been were modeled with regional travel demand models. However,

accurate evaluation of most TCMs requires that analyses be performed by fully run travel demand models. Therefore, regional travel demand models' current abilities, if they were used to evaluate the categories of TCMs, were examined to identify needed model improvements.

Categories of TCMs related to changes in travel time, changes in travel cost, and expanded transit options can only be evaluated adequately with TRIPS, particularly if it is used in conjunction with a network (assignment) model. That is primarily because TRIPS incorporates highway and transit travel time and cost in its mode choice, trip distribution, trip generation, and automobile ownership steps and the model is fully iterated. The Sacramento, Southern California, and San Diego regions incorporate highway and transit travel time directly only in the mode choice and trip distribution steps and incorporate travel cost directly only in the mode choice step. However, travel cost is included indirectly in the Southern California and San Diego regions' trip distribution steps through feedback. None of these three regions has an automobile ownership model that is endogenous and is affected by accessibility or by other variables that can be altered with policy. The Sacramento region does not recycle assigned travel impedances back to trip distribution.

To improve the accuracy of travel times in the models, all MPOs should improve their representation of access to transit and automobile in the mode choice step. Further, only San Diego's model represents signal and intersection delay separately from link capacity and delay. None of the models include explicit comfort and reliability variables to capture demands for expanded travel options accurately. To simulate the effects of TCMs that increase the monetary cost of peak-period travel (e.g., congestion pricing), all of the MPOs need to develop time-of-day choice models. In addition, only TRIPS retains income in all the trip tables, which allows analysis of the equity implications of pricing measures.

All of the MPOs have pricing data related to automobile operating costs, tolls, transit fares and discounts, and some parking cost data. The Bay Area region has daily and monthly parking cost data. The San Diego region's parking data are adequate except that more data are needed regarding the share of employees with free parking. Sacramento has parking cost data (monthly zonal averages) only for the downtown area and none for suburban or special generator areas.

None of the regions use travel demand models that can evaluate adequately TCMs that improve walk, bicycle, and transit environments. None of the regions represent walk and bicycle modes separately in the mode choice step except for the San Diego region, and its walk and bicycle modes are not policy sensitive (they are exogenous). In general, walk, bicycle, and transit accessibility variables (for example, proximity of employment and housing to transit and services, and walk and bicycle characteristics of zones) are lacking in the mode choice, trip distribution, trip generation, and automobile ownership steps. The Bay Area and Sacramento regions are able to represent, to some degree, the homogeneity and heterogeneity of land uses, because they include a variable for employment in the zone of residence. All MPOs should replace CBD and other locational variables with variables that represent regional accessibility and improve the detail of their networks to represent the proximity of employment and housing to transit and services. The regions all use reasonably small zones in areas of dense land use.

All the regions lack sufficient transportation and land use supply data, particularly related to zonal walk and bicycle characteristics. All have transportation supply data related to the transit and automobile travel times, roadway lanes, park-and-ride lots, and transit stops. Only the San Diego region has data on intersection configurations, parking inventories, and walk and bicycle distance. All need data related to the character of sidewalks, bicycle paths and lanes, availability of secure bicycle parking spaces at transit stops, and roadway medians.

For TCMs related to clean vehicle technology, all regions can calculate emission improvements from TCMs that affect the entire fleet in a uniform manner by substituting a revised set of composite factors. None has the capacity yet to evaluate partial fleet changes with regional models. However, some data are available on vehicles from the California Department of Motor Vehicles, which could be used in conjunction with the TRIPS model (and perhaps other models) to evaluate the effects of this TCM (*4*).

For categories of TCMs related to promotion of alternative modes and to the easing of activity constraints (e.g., flextime and telecommuting), all regions can use carefully interpreted case studies to manually adjust their mode choice projections. All MPOs use case study data, but available documentation suggests that none, except the Southern California region, used them to manually adjust mode choice projections.

Only the Southern California and Sacramento regions included TCMs intended to limit travel options. However, to model these TCMs, all the regions would need to improve the detail of their networks (i.e., obtain a more detailed depiction of roadways in restricted areas) and use an adjustable choice set in their mode choice models.

To assess secondary effects of changes in new residential and employment locations due to TCMs, only the Southern California region iterated the travel demand projections with land allocation model projections. The Bay Area and San Diego regions could do this, but they did not do so for their TCM analyses. The Sacramento region currently does not have a land allocation model. None of the regions used alternative land use projections as a TCM, although all of them have done such studies in the past.

Planned Model Improvements

The San Francisco Bay Area plans to pursue the following travel demand model improvements, which should improve their ability to analyze TCMs: (*a*) incorporate walk and bike accessibility (land use) variables in the trip generation step; (*b*) develop a mode of access to rail, investigate land use density effects on transit ridership, compare generic with mode-specific time and cost parameters, and examine HOV time saving coefficients in the mode choice step; (*c*) improve the forecasting method for projecting vehicle occupancy rates, especially for nonwork trips; and (*d*) develop time-of-day choice models (*21*). These changes should be incorporated into the TRIPS model if it is used for future TCM evaluations.

The Sacramento region is currently undertaking a major update of its travel demand models and plans to incorporate the following: (*a*) an automobile ownership step that is sensitive to walk and bicycle accessibility (land use) variables and to transit access; (*b*) a trip generation step that is also sensitive to land use variables; (*c*) a mode choice step in which walk and bicycle modes are represented and zonal indexes of pedestrian and bicycle friendliness are incorporated; (*d*) travel cost variables in all model steps; (*e*) a time-of-day choice model (*22*); and (*f*) more data related to the pedestrian and bicycle environment of zones (*23*). The Sacramento region is also considering a land allocation model and is gathering the needed land use data (*23*).

The Southern California Association of Governments currently is preparing its strategic plan for improving its model, and thus the

TABLE 1 Improvements Needed in Regional Travel Demand Models To Evaluate Transportation Control Measures

	CHANGE TRAVEL TIME, CHANGE TRAVEL COSTS, AND EXPAND TRANSIT OPTIONS
BAY AREA	1. improve access to transit and auto 2. explicit comfort and reliability variables in mode choice 3. fully iterate with a land allocation model 4. signal and intersection capacity and delay separate from link 5. time-of-day choice
SACRAMENTO	1. an auto ownership step 2. travel time and travel cost in all steps 3. recycle congested impedances back to auto ownership 4. explicit comfort and reliability variables 5. retain income in final trip tables 6. signal and intersection delay separate from link 7. time-of-day choice 8. more detailed pricing data 9. full iteration with a land allocation model
SOUTHERN CALIFORNIA	1. an auto ownership step 2. travel time and travel cost in all steps 3. recycle congested impedances back to auto ownership 4. explicit comfort and reliability variables 5. retain income in final trip tables 6. signal and intersection capacity and delay separate from link 7. time-of-day choice
SAN DIEGO	1. an auto ownership step 2. travel time and travel cost in all steps 3. recycle congested impedances back to auto ownership 4. explicit comfort and reliability variables 5. retain income in final trip tables 6. time-of-day choice 7. more detailed pricing data

(*continued on next page*)

association was able to provide us only with information about its proposed mode choice model improvements. It is considering incorporating the following into its mode choice step: (*a*) expanded subdivisions of modes, whereby transit may be subdivided into bus, commuter rail, and rail transit, for example (however, explicit walk and bicycle modes are not being considered because variables that influence their use are difficult to quantify); (*b*) improved representation of highway terminal times, automobile and walk access to transit, automobile parking cost, and automobile operating costs; and (*c*) increased market segmentation, which may include expanded trip purposes, household income or automobile ownership, other household characteristics (e.g., household size, number of workers, and number of children), parking pricing, and travel time of day (*24*).

The San Diego region planned, by the end of 1992, to (*a*) consider incorporating trip-chaining and accessibility in the trip generation step; (*b*) include direct travel costs in impedance measures in the trip distribution step; (*c*) improve feedback mechanisms, where possible; (*d*) consider adding a light rail mode; (*e*) double modeled roadway mileage and code separate HOV facilities in the network; and (*f*) add simultaneous HOV trip table assignment (*25*).

CONCLUSIONS REGARDING MPOs' NEAR-TERM CAPABILITIES

The Bay Area, using the TRIPS model and incorporating the model changes under way, will be the best equipped to capture the effects of TCMs involving changes in travel times and costs and expanded transit options, primarily because TRIPS incorporates travel time and travel cost in all model steps and recycles assigned impedances back through automobile ownership and subsequent steps. Generally, the other regions incorporate travel time and travel cost only in their mode choice and trip distribution steps, and assigned impedances are recycled, at best, only back through trip distribution. However, Sacramento and San Diego plan to expand their inclusion of travel time and cost in more model steps, as described above, which will improve their analyses of these TCMs. Both the Sacramento and Bay Area regions are taking steps to improve their models' depictions of peak spreading. Sacramento is also planning to develop an automobile ownership model. Items in the "Change Travel Time, Change Travel Costs, and Expand Transit Options" box in Table 1 that the regions still need to add to their programs for model improvements are as follows: Bay Area,

TABLE 1 *(continued)*

	CHANGE LAND USES
BAY AREA SACRAMENTO SOUTHERN CALIFORNIA SAN DIEGO	1. fully represent walk and bike modes 2. walk, bike, and transit accessibility variables in all model steps 3. more transportation and land use supply data 4. regional accessibility variables, not CBD 5. improve network detail

	CLEAN VEHICLE TECHNOLOGIES
BAY AREA SACRAMENTO SOUTHERN CALIFORNIA SAN DIEGO	1. a model or data base that links vehicle types to trip categories

	EASE ACTIVITY CONSTRAINTS
SOUTHERN CALIFORNIA	1. when available, use model of human activity scheduling
BAY AREA SACRAMENTO SAN DIEGO	1. careful interpretation of case studies to manually adjust mode choice projections 2. when available, use model of human activity scheduling

	PROMOTION OF ALTERNATIVE MODES
BAY AREA SACRAMENTO SAN DIEGO	1. careful interpretation of case studies to manually adjust mode choice projections

	LIMIT TRAVEL OPTIONS
BAY AREA SACRAMENTO SOUTHERN CALIFORNIA SAN DIEGO	1. improve network detail 2. use an adjustable choice set

2, 4; Sacramento, 4–6, 9; Southern California, all; and San Diego, 1, 3, 4–7.

Currently, all the regions have a limited ability to evaluate TCMs related to changes in land uses (i.e., improved walk, bicycle, and transit accessibility). In general, the MPOs can all improve their models' abilities to evaluate these TCMs by representing walk and bicycle modes in the mode choice step; including walk, bicycle, and transit accessibility variables in all model steps; and obtaining more detailed land use and transportation supply data. Sacramento plans to incorporate expanded land use variables in the automobile ownership, trip generation, and mode choice steps, and the Bay Area is considering incorporating more land use variables in the trip generation and mode choice steps. Sacramento is adding walk and bicycle modes. The other regions should attempt this. Sacramento should acquire a land allocation model.

To evaluate TCMs related to clean vehicle technology, all the MPOs will have to develop a model or use a data base that can link vehicle types to trip categories. Available documentation suggests that none of the MPOs has plans to develop that capability.

Carefully interpreted case studies can be used to manually adjust mode choice projections in evaluating TCMs that promote the use of alternative modes or that impose activity constraints. However, models of human activity scheduling should be used as they become available. Finally, all the MPOs should be able to model TCMs that limit travel options by increasing their network detail and using an adjustable choice set in the mode choice model.

ACKNOWLEDGMENTS

We thank the California Energy Commission for sponsoring this work and the University of California, Universitywide Energy Research Group, for its support. Helpful comments were received from Greig Harvey and Leigh Stamets.

REFERENCES

1. Harvey, G., and E. Deakin. *Toward Improved Regional Transportation Modeling Practice.* National Association of Regional Councils, Washington, D. C., 1993.
2. Replogle, M. *Best Practices in Transportation Modeling for Air Quality Planning.* Silver Spring, Md., 1991.
3. Johnston, R. A., and C. J. Rodier. *Critique of Regional Travel Demand Models in California.* California Energy Commission, Sacramento, 1993.
4. Harvey, G. *Screening Transportation Control Measures for the San Francisco Bay Area.* Metropolitan Transportation Commission, Oakland, Calif., 1989.
5. Replogle, M. Improving Transportation Modeling for Air Quality Planning. Presented at 72nd Annual Meeting of the Transportation Research Board, Washington D. C., 1993.
6. Wachs, W. Policy Implications of Recent Behavioral Research in Transportation Demand Management. *Journal of Planning Literature,* Vol. 5, No. 4, May 1991, pp. 333–341.
7. Stopher, P. R. Deficiencies of Travel-Forecasting Methods Relative to Mobile Emissions. *Journal of Transportation Engineering,* Vol. 119, No. 5, Sept. 1993, pp. 723–741.
8. Bae, C. Air Quality and Travel Behavior, Untying the Knot. *Journal of the American Planning Association,* Vol. 59, No. 1, Winter 1993, pp. 65–75.
9. Cameron, M. *Transportation Efficiency: Tackling Southern California's Air Pollution and Congestion.* Environmental Defense Fund and Regional Institute of Southern California, March 1991.
10. Hartgen, D. Attitudinal and Situational Variables Influencing Urban Mode Choice: Some Empirical Findings. *Transportation,* Vol. 3, No. 4, 1974, pp. 377–392.
11. Eisinger, D. S., E. A. Deakin, L. A. Mahoney, R. E. Morris, and R. G. Ireson. *Transportation Control Measures: State Implementation Plan Guidance.* Report SYSAPP-90/084. U.S. Environmental Protection Agency, 1990.
12. Applied Development Economics and Deakin, Harvey, Skabardonis Inc. *Socioeconomic Analysis of Proposed Regulation 13: Rule 1 Trip Reduction Requirements for Large Employers.* Bay Area Air Quality Management District, San Francisco, Calif., 1992.
13. Harvey, G., and E. Deakin. *Transportation Control Measures for the San Francisco Bay Area: Analyses of Effectiveness and Costs.* Bay Area Air Quality Management District, San Francisco, Calif., 1991.
14. *Air Quality, Congestion, Energy, and Equity Impacts of Market-Based Transportation Control Measures, Part 2: Technical Proposal.* Deakin, Harvey, Skabardonis Inc., Berkeley, Calif., Oct. 1992.
15. JHK & Associates. *Draft Review of TCM Impact Assessment Tools and Regional Travel Demand Modeling Capabilities.* Southern California Association of Governments, Los Angeles, 1992.
16. *Regional Transportation Plan.* Sacramento Area Council of Governments, Sacramento, Calif., 1992.
17. *Regional Transportation Plan Technical Appendix.* Sacramento Area Council of Governments, Sacramento, Calif., 1992.
18. *Air Quality Management Plan.* South Coast Air Quality Management District, Los Angeles, Calif., 1991.
19. *1992 Air Quality Strategy.* San Diego Air Pollution Control District, San Diego, Calif., 1992.
20. Sierra Research, Inc. *User Manuals for Software Developed To Quantify the Emission Reductions of Transportation Control Measures.* San Diego Association of Governments, San Diego, Calif., 1991.
21. *Research Design and Strategic Plan 1992–1995 Bay Area Regional Transportation Database/Travel Demand Models and Travel Forecasting.* Metropolitan Transportation Commission, Oakland, Calif., 1992.
22. DKS Associates. *Technical Work Program.* Sacramento Area Council of Governments, Sacramento, Calif., 1992.
23. *Workshop on the Sacramento Area Travel Demand Model and Land Use Model Issues.* Sacramento Area Council of Governments, Sacramento, Calif., 1993.
24. Cambridge Systematics, Inc. *Southern California Association of Governments Regional Mode Choice Model.* Southern California Association of Governments, Los Angeles, 1993.
25. *Transportation Model Revisions for Series 8 Forecasts.* San Diego Association of Governments, San Diego, Calif. (undated).

The views presented in this paper are solely those of the authors.

Publication of this paper sponsored by Task Force on Transportation Modeling Research Needs.

Simulation Model for Evaluating the Performance of Emergency Response Fleets

K. G. ZOGRAFOS, C. DOULIGERIS, AND L. CHAOXI

A simulation model for evaluating the performance of an emergency response fleet of an electric utility company is presented. The proposed model considers the spatial, temporal, and severity distribution of calls and has the capability to simulate alternative configurations of service districts and dispatching policies of the emergency response fleet. A nonstationary Poisson process is used to simulate the temporal distribution of service calls, whereas discrete simulation is employed for the spatial and severity distribution of the service calls. A mixed planar and network model is used to calculate the shortest travel time between the service calls and the location of emergency response vehicles. The model is validated on the basis of historical data. It is used to evaluate the relationship between fleet size and total incident service time and to compare alternative configurations of service districts for the same fleet size and dispatching policy.

A central problem in managing spatially distributed emergency response operations, such as police, ambulance, fire, emergency repair, and roadway assistance, is to determine the number of mobile units (fleet size) that should be available to respond to emergency calls, the service territories, and the dispatching strategies of the Emergency Response Units (ERUs) (*1,2*). The primary measure of effectiveness of an emergency response system is the minimization of the average response time to emergency calls (*1–3*). Average response time depends on the spatial, temporal, and severity distribution of the service calls and the size and deployment strategy of the emergency response fleet (*1,2*).

Thus, efficient deployment of an emergency response fleet requires examination of trade-offs between the cost of the emergency response fleet operations, that is, the number of ERUs available for deployment, and resulting performance, or average response time. Examination of this trade-off requires development of analytical tools that will be able to evaluate the performance of an emergency response mechanism for various levels of expected work load and manpower availability.

The objective of this paper is to present a simulation tool, developed for evaluating the performance of alternative districting patterns and fleet size of an emergency response system. The paper focuses on the development of the simulation model and its application as an evaluation tool. The work presented is motivated by the emergency repair operations of a large utility company. The proposed simulation tool is part of an integrated decision support system (*4*) that was developed to help the service restoration managers improve the effectiveness of the utility's service restoration fleet.

K.G. Zografos, Department of Transportation Planning and Engineering, National Technical University of Athens, Athens, Greece. 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.

EMERGENCY RESPONSE FLEET OPERATIONS

The need for emergency response services arises when incidents requiring prompt response and attention occur randomly in space and time. Depending on the type of incident, whether a fire, medical emergency, police emergency, emergency repair, or roadway incident, a mobile ERU should be dispatched to the scene of the incident to offer the necessary services. A crucial parameter involved in the design and evaluation of emergency response operations is the total incident service time (TIST). TIST is defined as the time elapsed between the occurrence of an incident and the completion of requested services (*3,4*).

TIST consists of four time intervals. The first interval, T_1, the incident detection and identification time, is determined by the time elapsed between the occurrence of an incident and the arrival of a call at a dispatching center announcing the incident and requesting services. The duration of T_1 depends on the technology used to detect the incident (*5*). For certain types of incidents, there are opportunities for automatic incident detection, depending on the capacity of the switchboard receiving the incident calls and the technology used to associate calls with the location of the incident (*6,7*).

The second interval, T_2, or the dispatch delay, is determined by the time elapsed between the detection and identification of the incident and its assignment to the first available ERU. The magnitude of a dispatch delay depends on the ERUs' degree of use. In the case of a congested system, when the utilization rate of the servers exceeds a threshold value, the dispatch time is the major determinant of response time (*5–8*).

The third component, T_3, is the time interval required by an ERU to travel from its current location to the scene of the incident. The fourth interval, T_4, involves the time that an ERU spends in providing the requested services. The duration of the incident service time depends mainly on the severity of the incident (*1*).

TIST, and consequently the performance of the emergency response system, can be enhanced substantially by reducing dispatch delay (T_2) and travel time (T_3). Thus, any modeling effort regarding the improvement of the deployment of an emergency response fleet should take into account the interactions between the various components of the TIST and their effect on T_2 and T_3.

PREVIOUS RELATED WORK

Computer simulation methods have been used extensively to study the performance of emergency response systems and are described in the literature. Simulation models offer the capability to study the trade-off between the number of servers and TIST for complex, large-scale systems that are not amenable to exact queuing theory formulations. In addition, simulation models can be used to evalu-

ate the performance of alternative districting patterns generated by districting models.

Savas (*9*) used simulation as a tool to perform a cost-effectiveness analysis of New York's emergency ambulance services, thereby linking operations research models, decision making, and computer techniques.

Larson (*10*) used the hypercube queuing model, which incorporates theoretical queuing theory results and simulation as a tool in dispatching of police patrol cars. Brandeau and Larson (*11*) extended the use of the hypercube model to the deployment of emergency ambulances.

Ignall et al. (*12*) used simulation to suggest approximate analytical models to be used for police patrol and fire operations in New York City. The link between simulation and analytical models has been analyzed further and evaluated in a paper by Shantikumar and Sargent (*13*), in which several uses of these hybrid models are suggested.

Green and Kolesar (*14*) have used simulation as a tool to evaluate a multiple car dispatching model for police patrol. Their application involved the police patrol fleet of New York City.

Zografos et al. (*5*) developed a simulation model for studying the trade-off between freeway incident delay and the size of a freeway emergency response fleet, and for studying the effect of alternative dispatching strategies on the performance of the freeway emergency response fleet.

Goldberg et al. (*15*) developed a simulation model for evaluating alternative base locations for an emergency (paramedic) response fleet in Tucson, Arizona.

Although emergency response systems share common characteristics, operations cannot be simulated in a generic sense. Evaluation of the performance of different emergency response systems requires the development of simulation models that can capture the particular administrative, organizational, technical, technological, and operational characteristics of the system under consideration. In the emergency repair operations of electric utility companies, for example, there is no patrolling as in police and freeway incident management operations; there is no multiple vehicle dispatching as in fire, police, and freeway incident management operations; and the service unit does not necessarily return to its home base as in ambulance and fire protection operations.

RELATIONSHIP BETWEEN DISTRICTING AND SIMULATION

The proposed simulation model is part of an integrated decision support system that was developed to help managers of emergency repair operations in the electric utility industry to rationalize the deployment of an emergency repair fleet. The integrated decision support system for emergency repair operations consists of two interrelated modules: a districting module and a simulation module.

For a service area consisting of a number of elementary spatial units (atoms) with a given level of emergency repair activity, the objective of the districting module is to determine contiguous, nonoverlapping repair service areas, or "truck-areas," in such a way as to minimize the weighted distance between atoms belonging to the truck-area and the center of the truck-area.

Two more criteria can be incorporated into the districting model. The first criterion expresses the requirement of a balanced work load assignment between truck-areas, whereas the second criterion requires that the area of each truck-area be within a given percentage of the average.

The objective of the simulation module is to simulate service restoration operations within each truck-area generated by the districting module. The proposed simulation model uses historical data describing the spatial, temporal, and priority distribution of emergency repair calls and simulates alternative dispatching strategies for the emergency repair fleet. The output of the simulation module provides statistical information related to the performance of the service restoration units within each truck-area. The performance of the service restoration mechanism is evaluated in terms of the average dispatch, travel, and repair (DTR) time (i.e., the time interval between the identification of the service call and the completion of service). T_l is not included in the performance analysis because it is not affected by fleet size or the shape of the districts.

SPATIAL, TEMPORAL, AND SEVERITY CHARACTERISTICS OF SERVICE REPAIR CALLS

An essential step in developing the proposed simulation model was to understand the operations and the behavior of the service restoration mechanism. That was done through an analysis of historical data describing the demand for repair services and the performance of the service restoration mechanism. The data base used for the analysis of the service restoration operations was provided by a major utility company. The data base included data covering a period of 1 year and each record of the data base corresponded to a call for an emergency repair. Each record contained information regarding the time that the service call arrived at the dispatching center, the time that a work order (ticket) was generated for the call, the time that the ticket was assigned to a field repair unit, the time that the field unit arrived at the scene of the incident, and the time that the repair was completed. In addition, each record contained information describing the location of the call in terms of the X and Y coordinates of a major and a minor reference grid used by the company to identify its facilities and customers in the two-dimensional space. The size of the major grid was 1 mi^2, whereas the size of the minor grid was 2,500 ft^2. Finally, each record included information describing the severity of the service call in terms of the type of the failure and the type of the equipment that failed. A separate data base providing information on the number of field units available on a shift basis per day was also provided by the same utility.

Analysis of the data indicated statistically significant differences in terms of the spatial, temporal, and severity distribution of the service restoration calls. Important information regarding service restoration operations was obtained by analyzing the duration of the repair for work orders of different priorities. In this case, it was found that the duration of the repair time was related to the severity of the incident (*4*).

Information obtained from analysis of existing data was used to develop probability density functions describing (*a*) the temporal distribution of the service calls, (*b*) the spatial distribution of the service calls, (*c*) the priority distribution of service calls, and (*d*) the distribution of the average repair time for service calls having different priorities. In addition, the average travel time between the centroids of the major grid atoms was calibrated using travel time information provided by the data base.

STRUCTURE OF PROPOSED SIMULATION MODEL

The simulation module input consists of the geographic definition of the truck areas generated by the districting module and the travel

speed information for the links of the underlying transportation network. The input information is provided for all three shifts and for the entire area under consideration. The output of the simulation program provides the statistics describing the performance of the service restoration mechanism (i.e., the statistics of the total incident response time and its components). Figure 1 shows the relationship between the various components of the simulation module at a macroscopic level.

The proposed simulation program offers the opportunity to simulate a wide range of operational characteristics of the service restoration mechanism. The alternative simulations can be run by varying a set of control data in a control input file.

On a more detailed level, the simulation module proceeds as follows: from the given data base, the priority distribution of service calls per shift is calculated for the whole area under consideration. Furthermore, the priority and type of service call distributions are calculated for each shift. On the basis of the calculated distributions and information regarding geographic and operational characteristics of the subarea, which is provided in control parameter input files, the following are calculated:

1. Repair time distribution for a given subarea, shift, priority, and type of service call;
2. Spatial distribution of service calls for a given shift, priority, and type;
3. Temporal distribution of service calls; and
4. Travel time matrix for the three shifts.

Once all the necessary distributions have been defined and calculated, the program proceeds with the generation of the following:

1. Calls on a shift basis;
2. Calls by priority;
3. Calls by type;
4. Call locations; and
5. Exact service call coordinates within the corresponding atom, using a uniform distribution.

The simulation program continues with the assignment of a service call to a particular service unit according to the defined dispatching policy. The activity of the ERU is followed throughout its shift. Any unfinished work load is transferred to the next shift. Performance and utilization statistics are gathered throughout the process. Figure 2 shows the detailed structure of the proposed simulation model. A 10-min delay is assigned to each ticket to compensate for T_0. Tickets are served according to their priority. Unserved tickets are transferred to the next shift.

Temporal Generation of Calls

Analysis of historical data describing the arrival of the service calls to the switchboard of the emergency repair system revealed that arrival of service calls can be described by a nonstationary Poisson process.

Therefore, a nonstationary Poisson process with rate $\lambda(t)$ was used to fit the service call arrival rate data (*16*). The rate $\lambda(t)$ was expressed by a polynomial function of the form

$$\lambda(t) = \sum_{i=0}^{n} a_i t^i$$

where the constants a_i and the order n of the polynomial were calculated on the basis of the best fit to the historical data.

The generation of the temporal distribution of the service calls requires the solution of the following problem:

Given the instant t_i of the arrival of the ith call, find $t_{i+1} = t_i + \Delta t$ (i.e., find the time of the next service call arrival).

From the nonstationary Poisson process, the distribution of an interval (t_i, t_{i+1}) is given by

$$\xi = F(t, \Delta t) = 1 - \exp\left[- \int_t^{t+\Delta t} \lambda(\tau) d\tau \right]$$

From the time interval distribution we get

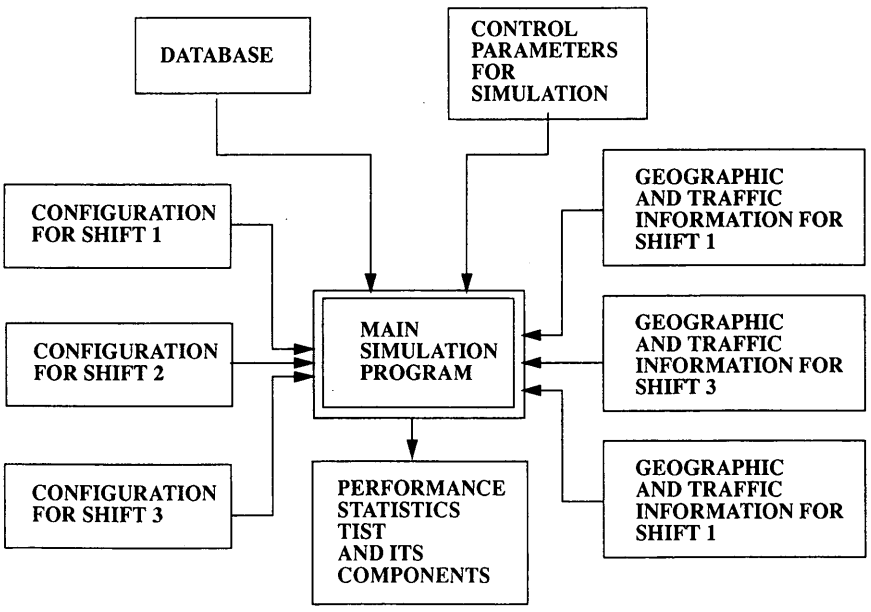


FIGURE 1 Interdependencies in the simulation mode.

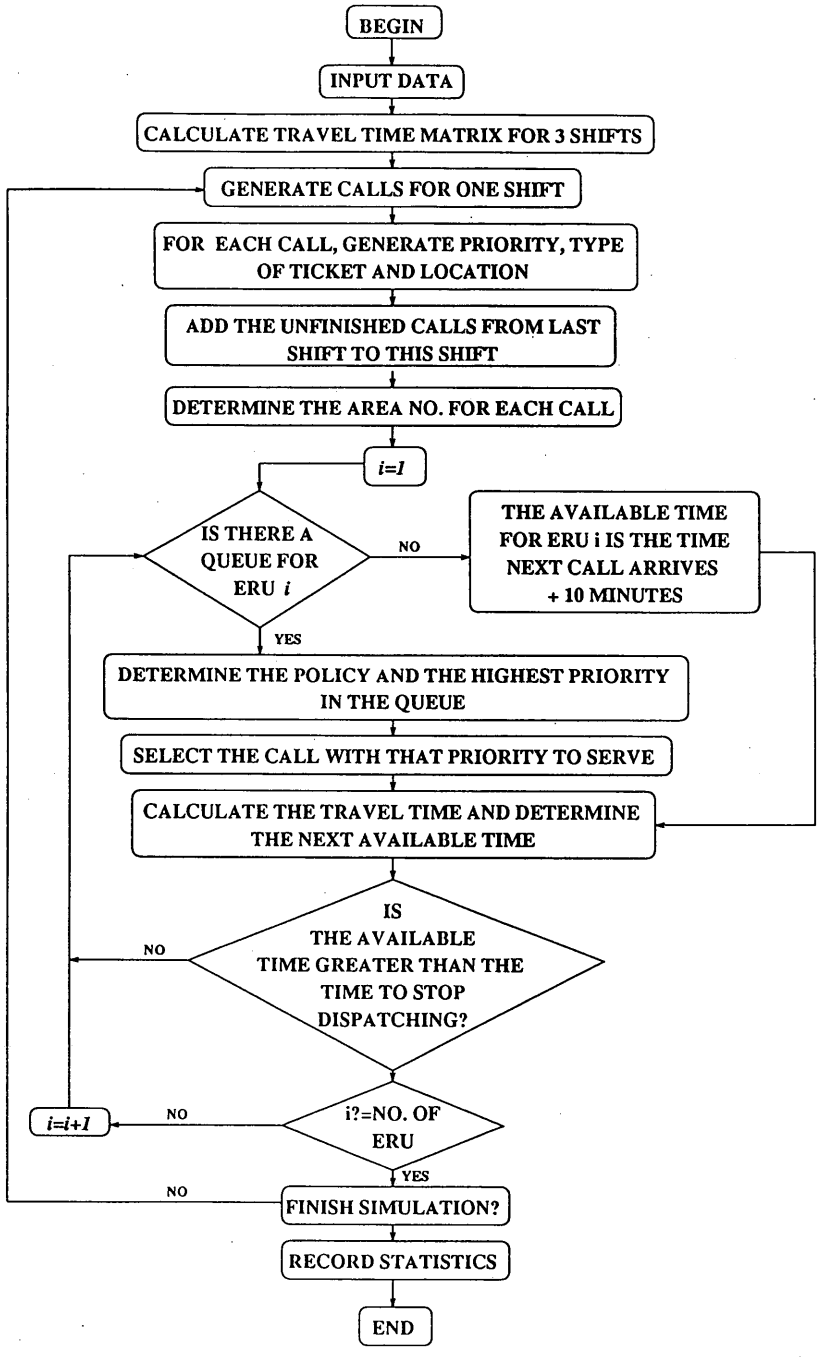


FIGURE 2 Flowchart for the simulation of 1 day of operations.

$$-\ln(1-\xi) = \int_{t_i}^{t_i+\Delta t} \lambda\tau\, d\tau$$

The following computational procedure was used to determine the interval Δt:

1. Generate a uniform random number ξ.
2. Calculate $z_1 = \ln(1-\xi)$, and let $z_2 = 0$, $t_{i0} = t_i$. These are the initialization conditions for dummy variables z_2 and t_{i0}, which are used subsequently for the determination of the stopping criterion and the updating procedure of the algorithm. Let Δt be a chosen small increment.
3. On the basis of the chosen Δt and Simpson's rule, calculate

$$\Delta z_2 = \frac{\Delta t}{3}[\lambda(t_{i0}+\Delta t) + 4\,\lambda(t_{i0}+2\Delta t)]$$

Let $z_2 = z_2 + \Delta z_2$ and $t_{i0} = t_{i0} + 2\Delta t$.

4. If $z_2 \geq z_1$, there is a need to get a smaller increase by interpolation, then

$$t_{i+1} = t_{i0} - (z_2 - z_1)\frac{2\Delta t}{\Delta z_2}$$

else go to Step 3.

5. End.

The procedure described above provides a stable solution to the temporal generation of calls.

Spatial Generation of Calls

A two-step procedure was used for the spatial generation of the service calls. First, the atom to which a call belongs was specified, and then the exact location within a specific atom was determined. To generate service calls in atoms, a discrete probability distribution was used.

The next step proceeds with the uniform generation of the service call within the specific atom of area A, as determined in the previous step. A common procedure to generate a point uniformly in an atom of area A requires first the generation of a point within a rectangle that covers the given region and then a check to determine whether the point is located within the specific atom (*17*). This two-step procedure is repeated until a candidate point is finally acceptable (i.e., located within the atom).

For this paper, a new procedure is used, which assigns two random numbers to a corresponding point in the given atom of area A without the need for an iteration and checking procedure.

Because most of the atoms met in practical problems can be described or closely approximated as polygons, the problem is formulated for a polygon, which is described by the N counterclockwise-ordered vertices $(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_N, y_N)$. We assume that a point (x_0, y_0) can be selected within the polygon so that the segment line $(x_0, y_0) - (x_i, y_i)$ is located within the polygon for $i = 1, 2, \ldots, N$ (Figure 3a). A ray from (x_0, y_0) intersects the polygon at only one point. Let us denote the following:

A_i = area defined by the triangle with vertices at (x_0, y_0), (x_i, y_i), and (x_{i+1}, y_{i+1}) (Figure 3b);
$A_t = \sum_{i=1}^{N} A_i$, the total area of the polygon;
$B_i = (\sum_{j=1}^{i} A_j)/A_t$, the fraction of the polygon area in the first i triangles; and
$B_0 = 0$.

The procedure is described as follows: First, generate a pair of uniformly distributed random numbers (R_1, R_2) in (0, 1). Second, find the location of R_1 among the B_i. If $B_{i-1} \leq R_1 \leq B_i$, calculate α, x', y' such that

$$\alpha = \frac{R_1 - B_{i-1}}{B_i - B_{i-1}}$$

$$x' = x_i + \alpha(x_{i+1} - x_i)$$

$$y' = y_i + \alpha(y_{i+1} - x_i)$$

The desired point (x, y) is given by

$$x = x_0 + \sqrt{R_2}\,(x' - x_0),\ y = y_0 + \sqrt{R_2}\,(y' - y_0)$$

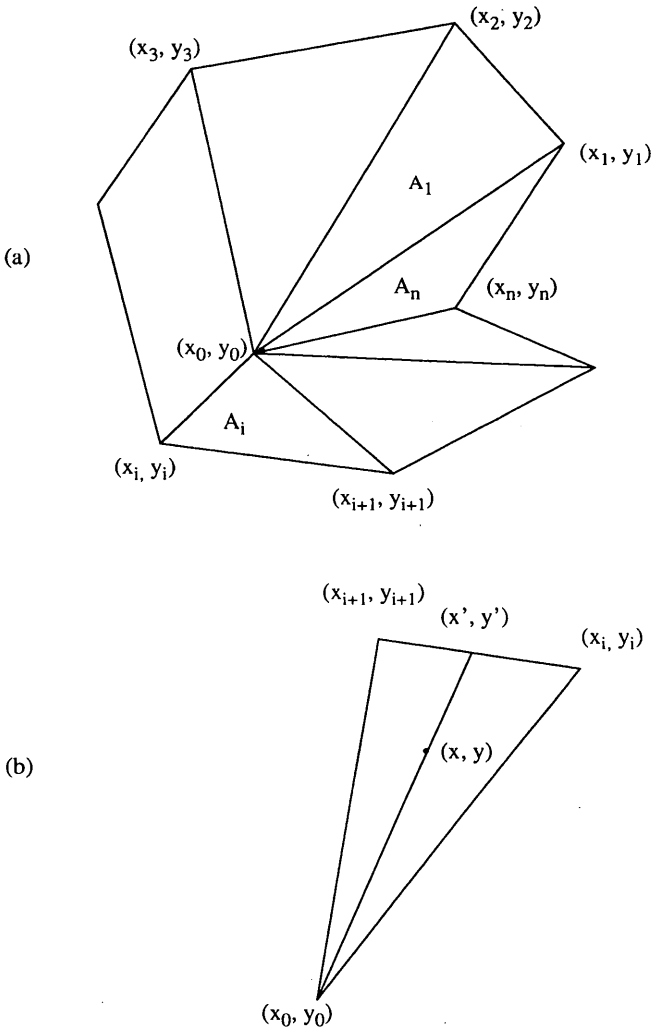


FIGURE 3 Uniform generation of calls within an atom.

This procedure directly generates points within the polygon without the need to check for acceptance or rejection of a certain point and the subsequently required iteration.

Priority, Type, and Repair Time Generation of Calls

A discrete distribution based on historical data is used to generate the priority of the service calls, the type of service calls of a certain priority, and the repair time of the service calls for a given shift, priority, and type.

Travel Time Estimation

The travel time estimation involves two steps. First, the travel time between the centroid of the atom where an ERU is located and the centroid of the atom where the service call originated is calculated. Second, the travel time from the actual location of the ERU and the service call to the corresponding centroid is calculated. The calculation of the travel time between the centroids of all atoms results in a travel time matrix whose elements are stored in a working file.

The matrix is calculated once for each shift. When a call is generated, the program identifies the atoms where the call and the ERUs are located and reads the corresponding travel time from the travel time matrix.

For the estimation of the shortest travel time matrix, all atoms are represented by the centroid (x^*, y^*) of the atom.

Because of the nature of the underlying transportation network (i.e., regular grid pattern), the Manhattan metric was selected to represent the distance for the surface street system of the network (i.e., local streets and main arteries). When the assumption of the regular grid pattern is violated because of travel on freeways, the existence of barriers to movement, a rural area's sparse road network, or for other reasons, special links are defined, and the network that was created by the special links is superimposed on the plane. Special links for freeways were defined by nodes representing entrance points to the freeway and the centroids of the neighboring atoms of the freeway. Special links around barriers were generated by a procedure described by Larson and Li (*18*).

Travel speed data for local streets, arterials, and freeways were collected for different hours of the day. For the surface street system, average travel speeds for each shift along the two major directions in the area (v_x, v_y) (the X axis corresponds to an east-west movement and the Y axis to a north-south movement) were obtained from the utility's personnel.

Arterial street speeds first were translated into equivalent local street speeds.

For freeway links, an average travel speed, v, for each shift was used; whereas for the special links, an average speed, v_a, for each shift was used. An algorithm is used to find the shortest travel time between all pairs of nodes and to determine the travel time matrix.

Given the atom center coordinate (x_i, y_i), the average travel speed (v_{x_i}, v_{y_i}) within the atom, the coordinates of freeway points, and the allowable travel directions on it (entrance, exit, or both), the average travel speeds between two freeway points, special link information, and the travel speeds on special links, the following algorithm determines the travel time matrix.

1. Initialize travel matrix: $T_{ij} = \infty$.
2. Calculate the travel time for each link.

For link i-j connecting the centroids of atoms i and j, we use the weighted travel speeds:

$$T_{ij} = \frac{|x_i - x_j|}{\alpha_i v_{x_i} + \alpha_j v_{x_j}} + \frac{|y_i - y_j|}{\alpha_i v_{y_i} + \alpha_j v_{y_j}}$$

where α_i is the percentage of the distance between the two atom centroids that belongs to atom i (*2*).

For atom i with centroid coordinates (x_i, y_i) to freeway point link j with coordinates (x_j, y_j) (note that freeway points are treated as centroids of atoms), use the average speed (v_{x_i}, v_{y_i}) within the atom:

$$T_{ij} = \frac{|x_i - x_j|}{v_{x_i}} + \frac{|y_i - y_j|}{v_{y_i}}$$

For a freeway point link connecting points (x_i, y_i) to (x_j, y_j), use given speed v:

$$T_{ij} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{v}$$

For special links connecting points (x_i, y_i) to (x_j, y_j), use given average speed v_a:

$$T_{ij} = \frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{v_a}$$

3. Find the shortest way between all pairs of nodes using Floyd's algorithm (a node is the centroid of an atom or a freeway point).

CASE STUDY

The simulation model described in the previous section was used to examine the trade-off between the number of available ERUs and the service restoration time and to evaluate the effectiveness of alternative dispatching and districting patterns produced by the solution of the districting problem. Initially, the simulation model was validated in terms of its capability to reproduce the inputs of the simulation process (i.e., temporal, spatial, and priority distribution of calls and distribution of the duration of the repair time). Statistical significance tests were performed to compare the means of the simulated and observed data for various shifts and districts of the study area. The results of the tests suggest that there is no statistically significant difference between the observed and simulated data.

The model is applicable to any type of area, transportation network, and population density. The calibration of the specific parameters is subject, however, to a case-by-case validation.

The validated model was used to study the relationship between the number of available ERUs (the fleet size) and the performance of the service restoration mechanism, as it is manifested by the duration of DTR time and its components T_2, T_3, and T_4. First-come-first-served (FCFS) and nearest neighbor (NN) policies were used for the evaluation. Two distinct districting procedures were evaluated: constrained districting and unconstrained districting. According to constrained districting, the whole districting was divided by the utility personnel in three subdistricts, which satisfied their previous organizational and managerial structure, and then the model was used independently in each subdistrict. In unconstrained districting, the whole district was divided optimally into truck areas according to our model.

Figures 4 through 7 summarize the results of the evaluation for a particular district, for various numbers of ERUs, and compare them with results for the current districting pattern. The results presented involve Shift 1 (7 a.m.–3 p.m.) and are run for $N = 11$–17. The current districting pattern has $N = 14$. A first result emerging from this analysis is that the average total DTR time decreases as the number of servers increases. The reduction is attributed mainly to the reduction of the dispatch delay (T_2) component. The equalization of work load achieved by the districting module has a very clear impact on T_2; reductions in travel time are not that significant. The differences in average time shown in these figures were found to be statistically significant at the $\alpha = 0.05$ level.

The results of the comparisons suggest that the new districting patterns lead to a more efficient performance of the service restoration mechanism, which is manifested through a statistically significant ($\alpha = 0.05$) reduction of the average DTR time for the same number of servers. In most cases, as many as two ERUs can be removed before any change in quality of service, as it is manifested by the value of DTR. Given the cost of operating each ERU, one can easily determine the cost savings from observed reductions in fleet size.

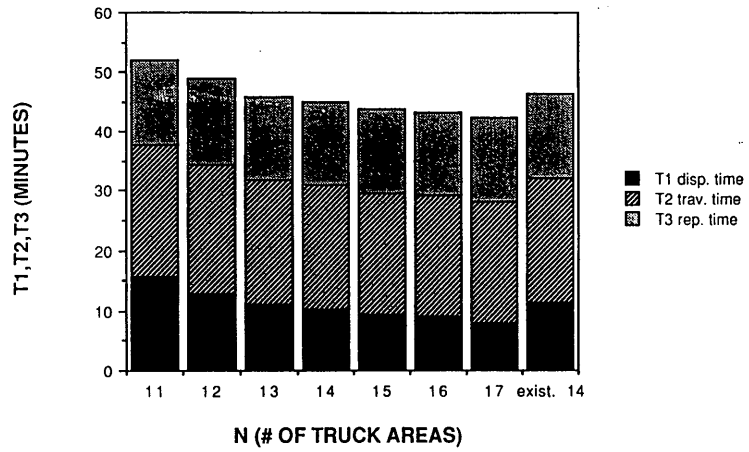


FIGURE 4 Relationship between DTR and number of ERUs (unconstrained districting, FCFS, Shift 1).

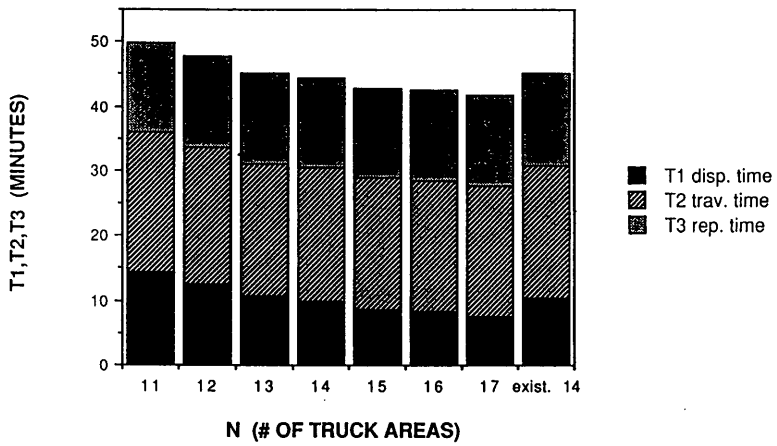


FIGURE 5 Relationship between DTR and number of ERUs (unconstrained districting, NN, Shift 1).

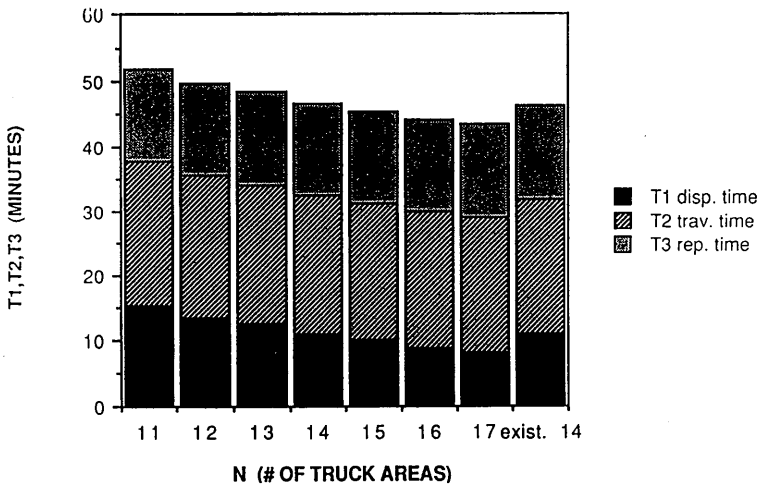


FIGURE 6 Relationship between DTR and number of ERUs (constrained districting, FCFS, Shift 1).

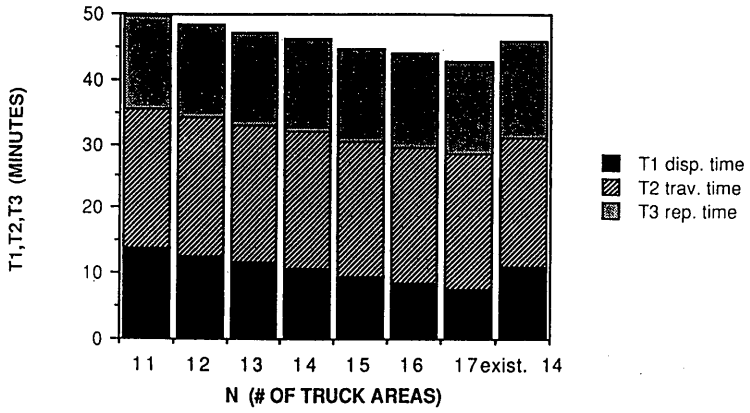


FIGURE 7 Relationship between DTR and number of ERUs (constrained districting, NN, Shift 1).

A comparison of the results of constrained and unconstrained districting indicates that unconstrained districting (Figures 4 and 5) provides better performance than constrained districting (Figures 6 and 7).

A comparison of the results of FCFS and NN policies indicates that NN gives consistently better results than FCFS, especially when the work load becomes very high (as with a small number of trucks).

CONCLUDING REMARKS

A simulation model for evaluating operations of the emergency repair fleet of an electric utility company was presented. The proposed simulation model has the capability to simulate alternative dispatching strategies (i.e., FCFS, NN, and their combinations). It is able to evaluate the performance of the emergency repair fleet in terms of the duration of DTR time. The model provides utility companies with an effective analysis and decision-making tool. Sample applications of the simulation model include an examination of the trade-offs between fleet size and duration of DTR, evaluation of alternative designs of service restoration districts, and examination of the effectiveness of alternative dispatching policies under various conditions of temporal, spatial, and severity distributions of service calls.

ACKNOWLEDGMENT

This work was supported by Florida Power & Light Company.

REFERENCES

1. Zografos, K. G., C. Douligeris, and L. Chaoxi. Model for Optimum Deployment of Emergency Repair Trucks: Application in Electric Utility Industry. In *Transportation Research Record 1358,* TRB, National Research Council, Washington, D. C., 1992, pp. 88–94.
2. Larson, R. C. *Urban Police Patrol Analysis.* MIT Press, Cambridge, Mass., 1972.
3. Jarvis, J. P. Models for the Location and Dispatch of Emergency Medical Vehicles. In *Emergency Medical Services* (T.R. Willemain and R. C. Larson, eds.), Lexington Books, Cambridge, Mass., 1977.
4. Zografos, K. G., C. Douligeris, L. Chaoxi, and P. Tsoumpas. Modeling Issues Related to the Service Restoration Problem. CEN 90-2. University of Miami, Coral Gables, Fla., 1990.
5. Zografos, K. G., T. Nathanail, and P. Michalopoulos. Analytical Framework for Minimizing Freeway Incident Response Time. *Transportation Engineering Journal, ASCE,* Vol. 119, No. 4, 1993, pp. 535–549.
6. Zografos, K. G., and C. Douligeris. Integrating Geographic Information System (GIS) and Automatic Vehicle Location (AVL) Technologies for Improving the Emergency Response Capabilities of Electric Utilities. Presented at VNIS, Dearborn, Mich., Oct. 20–23, 1991.
7. Zografos, K. G., C. Douligeris, J. Haupt, and J. Jordan. A Methodological Framework for Evaluating on Board Computer Technology in Emergency Dispatch Operations. Presented at VNIS, Dearborn Mich., Oct. 20–23, 1991.
8. Zografos, K. G., C. Douligeris, L. Chaoxi, and G. Develekos. Analysis and Optimization of Distribution System Reliability Through the Optimization of Emergency Response Operations. Presented at IEEE/NTUA Athens Power Tech, Athens, Greece, Sept. 5–8, 1993.
9. Savas, E. S. Simulation and Cost-Effectiveness Analysis of New York's Emergency Ambulance Service. *Management Science,* Vol. 15, No. 12, Aug. 1969, pp. B608–B617.
10. Larson, R. C. A Hypercube Model for Facility Location and Redistricting in Urban Emergency Services. *Computers and Operations Research,* Vol. 1, 1974, pp. 67–95.
11. Brandeau, M. L., and R. C. Larson. Extending and Applying the Hypercube Queueing Model To Deploy Ambulances in Boston. *TIMS Studies in the Management Sciences,* Vol. 22, 1986, pp. 121–153.
12. Ignall, E. D., P. Kolesar, and W. E. Walker. Using Simulation To Develop and Validate Analytic Models: Some Case Studies. *Operations Research,* Vol. 26, No. 2, March–April 1978, pp. 237–253.
13. Shantikumar, J. G., and R. G. Sargent. A Unifying View of Hybrid Simulation/Analytic Models and Modeling. *Operations Research,* Vol. 31, No. 6, Nov.–Dec. 1983, pp. 1030–1052.
14. Green, L., and P. Kolesar. Testing the Validity of a Queueing Model of Police Patrol. *Management Science,* Vol. 15, No. 2, Feb. 1989, pp. 127–147.
15. Goldberg, J., R. Dietrich, J. M. Chen, and M. Mitwasi. A Simulation Model for Evaluating a Set of Emergency Base Locations: Development, Validation, and Usage. *Socio-Economic Planning,* Vol. 24, No. 2, 1990. pp. 125–141.
16. MacLean. C. J. Estimation and Testing of an Exponential Polynomial Rate Function within the Nonstationary Poisson Process. *Biometrika,* Vol. 61, No. 1, 1974, pp. 81–84.
17. Larson, R., and A. Odoni. *Urban Operations Research.* Prentice Hall, Englewood Cliffs, N.J., 1981.
18. Larson, R. C., and V. O. K. Li. Finding Minimum Rectilinear Distance Paths in the Presence of Barriers. *Networks,* Vol. 11, 1981, pp. 285–298.

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

Vehicle Sizing Model for Bus Transit Networks

MAO-CHANG SHIH AND HANI S. MAHMASSANI

An iterative procedure to select vehicle sizes for the routes of a bus system with a given network configuration and origin-destination demand matrix is presented. The procedure starts by assigning a set of initial route service frequencies to compute route-level descriptors through a transit trip assignment model. The vehicle size for each route is computed analytically by a mathematical model that minimizes the total cost (operator cost and user cost) of each individual bus route. Revised frequencies are determined by applying a maximum allowed load factor consistent with the calculated vehicle sizes. The procedure terminates when frequencies of two consecutive iterations converge. The model is illustrated through a case application to the existing transit system in Austin, Texas. The result confirms the potential benefits of using variable vehicle sizes on different routes. However, the number of vehicle sizes in a system should be limited to avoid operational complexity and associated maintenance costs. In general, it appears that smaller buses could be operated on most of the bus routes in most North American cities to provide better service quality and lower operator cost.

Although both vehicle size and route frequency are important elements of bus service plans, most previous bus network design procedures treat vehicle size as a fixed value and compute route frequency either to achieve a minimum total generalized cost or to provide the capacity needed during peak-hour operation. Examples of these models are given elsewhere (*1–3*). The use of a fixed vehicle size simplifies the network design procedure, but it precludes the simultaneous consideration of various vehicle sizes in the bus system design and thus may result in ineffective resource allocation.

Because of high labor costs, transit operators in both Europe and North America tend to use fewer and larger buses to provide the capacity required during peak-period operation. Although smaller buses cost more to operate per seat, their use may offer several advantages. Glaister (*4*) argued that using small vehicles favors the provision of higher service frequencies, lowering average wait times and increasing operation speed; the improved service levels can be expected to generate new demand for bus transit. Furthermore, smaller buses may be better suited to some types of service, such as low-demand, low-occupancy, high-quality, or special transit, as Oldfield and Bly suggest (*5*). Smaller vehicles are more acceptable to residents of certain low-density neighborhoods and tend to inflict less damage on city street surfaces. Other suggestions for using different vehicle sizes are given elsewhere (*6–9*). To the extent that a given service area includes zones of different demand densities, allowing different vehicle sizes to operate on different bus routes and offer various types of services provides the transit operator with an additional choice dimension in designing the service configuration to better meet user needs and desired service levels.

Department of Civil Engineering, University of Texas at Austin, Austin, Tex. 78712-1076.

Only in a few studies have vehicle sizes been computed explicitly. Glaister (*9*) developed a simulation model to compare system operation under two vehicle sizes, a large vehicle (88 seats) and a small vehicle (15 seats). Results of the simulation suggest that buses seating 35 to 45 riders are likely to be most suitable for service in Aberdeen. Its level of detail notwithstanding, the computer simulation model does not describe explicitly the relationship between bus size and factors such as level of demand, operator cost, and load factor. Analytic models for finding optimal vehicle sizes have been developed for this purpose.

Previous analytic models include those of Jansson (*10*), Walters (*11*), Oldfield and Bly (*5*), and Chang (*12*). Jansson argued that previous analyses overweighted the producers' costs and underestimated the users' cost. He presented a model that minimizes total social cost (operator cost, passenger waiting time, and passenger riding time), subject to a peak capacity constraint satisfying a maximum occupancy rate (the ratio of the mean passenger flow to the product of the vehicle size and the service frequency). Jansson concluded that the optimal bus size determined by minimizing social cost tends to be smaller than that under the current practice of using a given vehicle size and setting the number of buses to achieve an average occupancy rate at or below a given maximum value.

Walters presented a simpler model that examines the trade-off between waiting time and labor cost. He also suggested that bus size should be considerably smaller than typically is used in Western European and North American cities. Gwilliam et al. (*13*) and Oldfield and Bly (*5*) argued that the waiting time assumption in Walters' model is questionable and thus yields an implausible relationship between optimal bus size and demand. Oldfield and Bly's model assumes elastic demand and determines the optimal bus size by minimizing total social cost. In addition, the average passenger waiting time in their model accounts for situations in which passengers are unable to board the first bus to arrive because it is full. They concluded that the optimal size lies between 55 and 65 seats (70-seat buses are most existing systems in the United Kingdom). Current cost structures could be changed to favor operation of smaller buses, but the optimal size seems unlikely to fall below 40 seats. Chang (*12*) presented analytic models to compare vehicle size for fixed route conventional bus with that of a flexible route subscription bus system. He concluded that the optimal vehicle size is less sensitive to the demand density for flexible route service than for fixed route service.

All the previous analytic models focus on the optimization of vehicle size for an individual bus line, which is treated independently of other lines in the network. In other words, demand on a particular bus line will not be affected by the optimal bus sizes and associated route frequencies of other bus lines. This is an incorrect assumption because, in a bus system, passengers may have several

paths on which to complete their trips. Changes in bus sizes alter the route frequencies and should lead to a redistribution of passenger flows on the bus network. Jansson's and Walters's models consider demand on each given route to be known and constant. Although Oldfield and Bly's as well as Chang's models consider demand to be elastic to account for the change in route demand resulting from changes in bus size, they do not consider the systemwide effects of changes in vehicle size.

This paper presents a vehicle sizing procedure in the context of a design procedure for bus networks and service plans. Instead of assuming the demand on each bus line to be known and given, as in all previous models, the model presented here solves for the route demands by assigning the trips in a given origin-destination (O-D) demand matrix, using a transit trip assignment model. The trip assignment model also computes the maximum link flow on each bus route. The resulting maximum link flow is more reliable than the value obtained as the product of a maximum occupancy rate and vehicle seating capacity. Both the route demand and the corresponding maximum link flow then form the basis for obtaining a set of optimal bus sizes and the associated route frequencies that minimize a generalized cost function.

OPTIMAL VEHICLE SIZE FOR SINGLE ROUTE WITH GIVEN DEMAND

The well-known square-root rule for setting frequencies on bus routes is based on the minimization of the sum of operator cost and passenger waiting time. Major weaknesses of the square-root formulation are that it does not account for bus capacity constraints and that it assumes demand is independent of service frequency. In the transit industry, the frequency of service on a bus route commonly is set to achieve an applicable maximum allowed load factor (*14*) and can be written as

$$f_k = \frac{(Q_k)_{\max}}{LF_{\max} S_k} \tag{1}$$

where

f_k = route frequency for route k,
$(Q_k)_{\max}$ = maximum hourly link flow of route k,
$LF_{\max}$ = maximum allowed load factor, and
S_k = vehicle size.

According to the frequency formulation, transit operators can select the desired load factor to meet operational considerations, such as comfort. Different load factors may be set for different subsets of bus routes, depending on the type of service provided, service area, and other special considerations reflecting local political preferences. Of course, when the frequency generated from the equation is unacceptably low because of low patronage, a minimum frequency policy commonly is applied in practice, as recognized in the design procedure developed by Baaj and Mahmassani (*15*).

The approach to determining the optimal vehicle size for each individual route is similar to the generalized cost approach used to obtain the square-root expression for frequency setting. However, instead of considering the frequency as the decision variable and the vehicle size as a constant, the vehicle size is taken as the decision variable, and the frequency is set as a function of the vehicle size consistent with Equation 1.

For a given demand level on a bus route k, the optimal vehicle size is obtained by minimizing the generalized cost C_k, which consists of the operator cost C_{ko} and the user cost C_{ku} (i.e., $C_k = C_{ko} + C_{ku}$). The derivation of the optimal vehicle size is based on peak-hour operation, which is the most critical period for determining the required fleet size of the system.

Oldfield and Bly (*5*) presented a reasonable and simple approximate formulation that expresses total operator costs as a linear function of vehicle size, as follows:

$$C_{ko} = a(1 + bS_k)VM_k \tag{2}$$

where

a = constant that adjusts the overall cost level,
b = constant that captures the relative rate of increase in cost with increasing vehicle size, and
VM_k = total vehicle miles per hour operated on route k.

The total vehicle miles per hour for each route k can be expressed as

$$VM_k = f_k RTM_k \tag{3}$$

where f_k is the frequency of service on route k and RTM_k is the round-trip miles for route k.

If the function f_k is set according to the equal peak-hour load factor rule (Equation 1), the operator's cost can be expressed as

$$C_{ko} = a(1 + bS_k)RTM_k \frac{(Q_k)_{\max}}{LF_{\max} S_k} \tag{4}$$

From the passengers' point of view, the total user cost, C_{ku}, for route k consists of three components: waiting cost (WC_k), in-vehicle travel cost ($IVTTC_k$), and access cost (AC_k), as proposed by Chang (*12*).

$$C_{ku} = WC_k + IVTTC_k + AC_k \tag{5}$$

Under the assumptions that (*a*) passengers arrive at random (uniformly), (*b*) passengers can always board the first available bus, and (*c*) vehicles arrive at constant headways, the average waiting time for passengers using route k is taken as half of the route's headway. Assuming that waiting time is valued linearly (an assumption that may be relaxed if alternative value functions are calibrated from empirical behavioral data), the total waiting time for passengers using route k can be expressed as

$$WC_k = wTPT_k \frac{1}{2f_k} = wTPT_k \frac{LF_{\max} S_k}{2(Q_k)_{\max}} \tag{6}$$

where w is the value of waiting time and TPT_k is the total passenger trips (demand) per hour using route k (which is computed in the trip assignment procedure).

The expected transit passenger waiting time in an actual system depends on both the reliability of the bus schedule and the distribution of the passenger arrival times. Under the assumption of uniformly distributed random passenger arrivals at bus stops, the average passenger waiting time increases as bus headways become less regular because more passengers on average arrive during longer intervals and fewer arrive during shorter intervals (*16,17*). However, passengers may not arrive randomly in all cases. Some transit

users tend, to some extent, to coordinate their arrivals with published schedules, if available, especially for routes with long headways. Bowman and Turnquist (*18*) have derived an expression for the expected wait time when the population of users is a mixture of "scheduled timers" and "random arrivals." The resulting waiting time function is highly system dependent and should be calibrated for each system, possibly for each bus route. However, the effect of schedule timing is offset to some extent by schedule unreliability, making the half-headway assumption an acceptable compromise. More important, from a design and frequency-setting standpoint, although "scheduled timers" may not incur an actual physical wait time at the stop, they incur a schedule delay relative to the actual time they would have wanted to depart. From the user cost standpoint in a design procedure, it is this schedule delay cost that must be included in the objective function, not the actual time at the stop. Evaluating waiting time on the assumption that users time their arrivals to coincide with the schedule can seriously underestimate user costs and lead to designs that do not meet user needs. This study uses a constant waiting value, w, for different modes (e.g., at home, at bus stop, or in office). Nevertheless, the procedure presented herein can be adapted easily to any waiting cost function specified by the model user, should sufficient justification and empirical support be available.

The in-vehicle travel cost is assumed independent of vehicle size, primarily because in-vehicle travel cost savings from using smaller buses are insignificant compared with the waiting-time cost savings. In-vehicle travel cost reduction may arise mostly from the possibly different average speeds of vehicles of different sizes. Smaller buses may provide faster service for two reasons: they have better maneuverability and fewer people are getting on and off them. Because bus speed is highly dependent on traffic conditions along the route, any improvement in the in-vehicle travel time cost of smaller buses usually is limited and insignificant relative to the potential waiting-time cost savings.

Another consideration of the constant $IVTTC_k$ assumption is the difficulty and resulting complexity of incorporating $IVTTC_k$ as a function of vehicle size in the cost function. The relationship between vehicle speed and vehicle size is difficult to specify analytically, especially in light of vehicle speed variation under different traffic conditions. Furthermore, vehicles of the same size with different engines may have different acceleration and deceleration characteristics. Therefore, it hardly seems worth the effort to incorporate route-dependent and condition-dependent $IVTTC_k$.

Using the above results and assumptions, the generalized cost C_k can be rewritten as follows:

$$C_k = a(1 + bS_k)RTM_k \frac{(Q_k)_{\max}}{LF_{\max}S_k} + wTPT_k \frac{LF_{\max}S_k}{2(Q_k)_{\max}} + AC_k + IVTTC_k \quad (7)$$

Note that AC_k and $IVTTC_k$ are independent of vehicle size. The optimal bus size S_k^* for given route demand levels can be obtained by setting $dC_k/dS_k = 0$, and it can be expressed as

$$S_k^* = \frac{(Q_k)_{\max}}{LF_{\max}} \sqrt{\frac{2aRTM_k}{wTPT_k}} \quad (8)$$

The relation indicates that the optimal vehicle size for a given demand level on a route is proportional to the level of the maximum link flow, $(Q_k)_{\max}$, and varies as the square root of round-trip miles of the route, RTM_k. The optimal vehicle size is inversely proportional to the load factor, $LF_{\max}$, as well as the square root of the total number of passenger trips, TPT_k, and the value of waiting time, w.

In Equation 8, the total cost (and associated optimal vehicle size) for a given route depends on the flow level, TPT_k. However, the latter is itself the result of the users' path choice through the network, which is a function of the vehicle sizes and frequencies, not only on the given route k, but on all network routes, $k = 1, \ldots, K$. The flows, TPT_k, $k = 1, \ldots, K$ are given by an assignment procedure, reflecting a passenger path choice rule, which distributes a given peak-period O-D trip matrix to the various bus routes. In our procedure, the vehicle sizes on each route (and associated frequencies) are set on the basis of route flows that are consistent with the vehicle sizes and frequencies through the iterative application of an assignment algorithm along with the vehicle sizing formula developed in this paper. However, the vehicle sizes obtained by this procedure are not necessarily optimal for the network as a whole. In other words, we do not seek to explicitly minimize the systemwide cost, $C = \sum_{k=1}^{K} C_k$, subject to consistency with a given assignment rule. Because of the network-level interactions described earlier, the objective function is not separable on a route-by-route basis. The resulting problems would be formidable because the assignment procedure used cannot be expressed as a well-behaved mathematical formulation. Instead, we propose a practical procedure that achieves an internally consistent solution that improves on existing methods.

VEHICLE SIZING PROCEDURE

The vehicle sizing procedure starts by assigning an initial set of frequencies to the bus routes. The O-D trip demand matrix for the bus system is then assigned to the bus routes using a transit trip assignment model. The transit trip assignment model computes both TPT_k (total passenger trips per hour using route k) and $(Q_k)_{\max}$ (the highest hourly link volume of route k): TPT_k and $(Q_k)_{\max}$ are then applied in Equation 8 to determine the locally "optimal" bus size for each route. To ensure that the resulting vehicle size remains within the range of buses under consideration, minimum and maximum size constraints are imposed. The vehicle size is then used in Equation 1 to compute the route frequency for each route. Note that for less congested bus lines the peak load factor method may generate frequencies that are lower than what riders can reasonably expect. In that case, a minimum frequency policy that sets route frequencies to a preset minimum value would be used instead.

The transit trip assignment model used in this study is described by Baaj and Mahmassani (*19*) in their transit network analysis procedure, TRUST. The model considers two main criteria: the number of transfers necessary to reach the destination and the trip times incurred with alternative path choices. The transit passenger is assumed to attempt to reach his or her destination by following the path that involves the fewest possible transfers. If two or more feasible paths are available with the same number of transfers, passengers are assumed to consider only those alternatives with trip times within a particular range. A "frequency-share" rule is then applied to assign trips according to the relative frequencies of service on the alternative paths. A more detailed description of the model can be found elsewhere (*2*).

Because the frequencies change from the initial values to new values, the demand of the bus system needs to be reassigned consistently with the new frequencies, and the optimal vehicle sizes and route frequencies then need to be recomputed as well. This procedure continues until two consecutive sets of route frequencies converge. This heuristic has exhibited convergence in all test cases

conducted to date. Figure 1 shows the flowchart of the bus sizing procedure.

In summary, the procedure consists of the following steps:

Step 0. Assign an initial set of route frequencies.

Step 1. Compute TPT_k and $(Q_k)_{max}$ using the trip assignment model.

Step 2. Determine vehicle sizes using Equation 8. If the optimal vehicle size is less than the minimum vehicle size, set the vehicle size equal to the minimum vehicle size.

Step 3. Set route frequencies using Equation 1. If the resulting frequencies are less than the minimum frequency, set the frequencies to the minimum frequency.

Step 4. Check whether two consecutive sets of route frequencies converge. If yes, stop; otherwise, go to Step 5.

Step 5. Use route frequencies determined in Step 3, and go to Step 1.

The vehicle sizing model has been implemented as part of AI-BUSNET, an artificial-intelligence–based bus network design computer program that initially was developed at the University of Texas at Austin by Baaj and Mahmassani (*19*). The program is written in Lisp because its list data structure capabilities provide an effective data representation to support extensive path search and enumeration in the bus network design problem. The program runs on an Apple Mac-II with MicroExplorer, a Lisp language compiler.

ILLUSTRATIVE APPLICATION

The transit network of the Austin, Texas, urban area was selected to illustrate the above vehicle sizing procedure. The transit network consists of 40 routes with fixed schedules, operated by the Capital Metropolitan Transit Authority (Capital Metro). Express routes, UT shuttle routes, and 'Dillo routes, which reflect different service and operations concepts, are not considered in this application. Buses

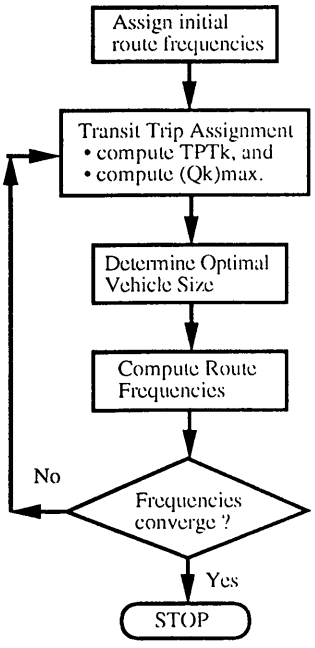


FIGURE 1 Vehicle sizing procedure.

with 37 or 43 seats are used by the system for peak-hour operation. Required inputs for this study include data on the network connectivity, nodal composition for each bus route, and a peak-hour transit O-D demand matrix. A total of 177 nodes are defined to describe the service area and associated network connectivity. Table 1 gives the numbers of the network routes, the node composition of each route, and the associated service frequency. The information is presented in list form as input to the analysis. The transit peak-hour demand matrix for the Austin area is generated using daily boarding and alighting data provided by Capital Metro. The resulting O-D trips are not necessarily those actually using the system. The system serves approximately 5,800 trips during peak-hour operation.

Table 2 summarizes the parameters used to determine optimal vehicle size and the associated route frequency for each bus route. (Values attached to the parameters are discussed later.) The coefficients, a and b, in the operator's cost function are derived from the operator costs associated with different bus sizes provided by Capital Metro. The operator cost parameters should be recomputed for other cities because wage rates and gasoline cost vary from city to city. The maximum load factor for peak-hour service is set at 1.25 (i.e., up to 10 standing passengers are allowed at any time if the bus seating capacity is 40 passengers), which is suggested by *NCHRP Synthesis of Highway Practice 69* (1980). The waiting cost coeffi-

TABLE 1 Bus Route Service Frequencies and Nodal Compositions

Route Name	Frequency	Nodal Composition
R1	7.5	(1 2 3 4 5 6 7 8 9)
R2	4.0	(13 12 11 10 1)
R3	4.0	(14 15 3 16 17 18 19)
R4	4.0	(25 24 23 22 15 21 10 20)
R5	3.0	(1 2 26 5 27 6 28 29 19)
R6	4.0	(1 2 10 30 31 32 33)
R7	4.0	(1 2 34 35 36 8 37 38 39)
R8	2.0	(40 41 42 43 44 45 46 36 8 19)
R9	1.82	(1 2 3 47 48 49 50 51)
R10	4.0	(10 1 52 53 54 55 56 57)
R11	1.5	(1 15 10 34 58 59 60 61)
R12	3.0	(2 64 65 66 67)
R13	4.0	(73 72 71 70 69 68 1 2 3)
R14	1.5	(78 77 76 75 63 21 10 74)
R15	4.0	(1 2 62 79 61 80 36)
R16	4.0	(84 83 82 81 1 2 62)
R17	4.0	(89 88 87 41 86 63 21 3 85)
R18	4.0	(90 10 20 91 43 92 93)
R19	1.5	(99 98 97 96 95 94 16 15 1)
R20	3.0	(1 15 62 59 100 101 102)
R21/22	2.14	(103 86 59 104 105 17 106 107 108)
R23	2.0	(19 109 110 111)
R25	2.0	(19 112 113 114 115)
R26	2.0	(119 118 25 117 116 1 2 26)
R27	4.0	(73 121 78 120 1 2 10 74)
R28	2.0	(78 69 54 83 66 122 123)
R29	1.5	(85 21 63 124 125)
R30	2.0	(123 128 127 126 52 2 74)
R31	1.33	(65 129 68 40 130)
R32	1.71	(36 80 131 42 132 133 134)
R33	2.0	(135 136 137 138 139 72 73)
R37	2.0	(140 21 10 141 142 45 101 143 144)
R38	2.0	(145 138 84 65 64 140 21 10 141)
R39	2.0	(36 146 46 147 148)
R40	1.71	(8 114 149 150)
R41	1.0	(151 152 153 9 8 5 2 154)
R42	2.0	(8 114 155 156 157 158)
R43	3.0	(159 160 67)
R44	1.0	(161 162 163 164 19)
R46	2.0	(19 165 166 167)

TABLE 2 Parameters Used in the Model and Values Assumed for the Application

Parameter	Definition	Value
a	coefficient on cost function	$ 2.96 / vehicle-mile
ACk	total passenger access cost of route k	-
b	relative gradient of cost function with vehicle size	0.0078
Ck	generalized cost for each route k	-
Cko	operator cost	-
Cku	user cost	-
fk	route frequency for route k	-
IVTTCk	total in-vehicle travel cost of route k	-
LFmax	the maximum allowed load factor	1.25
(Qk)max	the highest peak hour link volume on route k	-
RTMk	round trip miles for route k	-
Sk	vehicle size for route k	-
TPTk	total number of passengers using route k during the peak hour	-
vk	average bus speed for route k	12 mph
VMk	peak hour vehicle miles operated on route k	-
w	value of waiting time	$ 9, 12, 15/hour
WCk	total passenger waiting cost of route k	-

cient, w, is somewhat difficult to define. This application considers three values ($9/hr, $12/hr, and $15/hr). The minimum service frequency is set to be one bus per hour. A minimum vehicle size (10 seats) is selected when the calculated optimal vehicle size is less than that value.

Table 3 presents the resulting vehicle size and associated route frequency for each bus route in all three cases. Figure 2 shows the distribution of resulting vehicle sizes for all three waiting time values. In the case with the lowest waiting time value (w = $9/hr), 37 out of the 40 bus routes have an optimal bus size below 25 seats. Compared with the current bus system, this solution results in much lower passenger waiting cost ($7,616/hr versus $9,206/hr), with only slightly higher operator cost ($6,747/hr versus $6,6664/hr). For 15 out of 18 routes that currently operate with a frequency higher than two buses per hour, the model results provide for higher route frequencies relative to the current service. That is reasonable because the use of smaller buses usually requires higher service frequencies. Half of the 12 routes currently operating with a frequency of two buses per hour receive higher route frequencies. Only 1 of 10 routes currently operating with frequency less than two per hour receives a higher frequency in the model, suggesting that the current system uses higher minimum policy frequencies for routes with low passenger demand levels. In the two other cases with higher waiting time values, a smaller vehicle size is obtained for each bus route than in the case with the lowest waiting time value.

The result also demonstrates that the optimal vehicle sizes in the system are spread over a wide range. The use of a fixed vehicle size for the whole system is not an appropriate approach. Whereas it is infeasible to operate too many vehicle sizes in a system because of the resulting operational complexity and associated maintenance costs, meaningful benefits could be observed with a relatively small set of discrete values. For example, we reanalyzed the system under the assumption of only three commercially available vehicle sizes (37, 27, and 15 seats) and allocated those to each route using a simple nearest feasible integer heuristic. For the lowest waiting time value (w = $9/hr), the solution suggests that 32.2 percent of the savings in user costs (relative to the fixed size case) could be attained with just these three sizes. In addition, the operator cost in this case has been improved by a saving of $442/hr, as opposed to a loss of $83/hr in the optimal vehicle size case. Note that the operator cost savings are actually greater for the three sizes considered here than in the previous case because very small bus sizes are now avoided.

In general, bus systems operating a larger vehicle size have a lower operator cost. In Austin, larger vehicles (37 and 43 seats for peak-hour operation) are used on most bus routes. To provide a certain level of bus service, the bus system provides relatively high bus frequencies, resulting in higher operator cost. In addition, bus routes are operated with relatively low average load factors. Figure 3 shows that 34 out of 40 routes operate with load factors less than 0.75 in the Austin transit network. Similar situations are encountered in most North American cities. Capital Metro has recognized this fact and has been operating smaller buses on certain lower-demand suburban-oriented routes. For design purposes, vehicle sizes and service frequencies are selected to achieve the maximum allowed load factor. Therefore, only routes set at the minimum frequency because of low demand will have load factors below the maximum allowed.

CONCLUDING REMARKS

In this paper, an iterative procedure to determine the vehicle size accounting for the systemwide change in route demand resulting from changes in bus size is presented. An analytic formulation is derived to compute the locally optimal vehicle size by minimizing the total cost (operator cost and user cost) associated with each route. Bus route demand and the route's maximum link flow are critical to a determination of the optimal vehicle size. The demand and the maximum link flow are determined for each route by a trip assignment model that recognizes the operating characteristics associated with different vehicle sizes on each route.

The application shows that carefully selecting vehicle sizes will benefit both transit providers and users. In most North American cities, a large portion of routes is provided to low-demand areas to ensure users' mobility. Larger vehicles are still operated on these routes, resulting in either poor service quality or low vehicle occupancy rate. Clearly, a smaller vehicle size should be used on such routes. Because each bus system may include zones of different demand densities, various vehicle sizes should be used depending mainly on the bus route demand. However, the number of vehicle

TABLE 3 Optimal Bus Sizes and Associated Route Frequencies

Route Name	Case with w = $9/hr		Case with w = $12/hr		Case with w = $15/hr	
	fk	Sk	fk	Sk	fk	Sk
R1	7.3	28	8.5	24	9.7	22
R2	5.0	12	5.5	11	6.0	10
R3	4.6	15	5.4	13	5.8	12
R4	4.8	14	5.5	13	6.5	11
R5	3.5	14	4.1	12	4.4	11
R6	3.4	10	3.4	10	3.4	10
R7	4.6	27	5.4	23	5.9	21
R8	4.4	19	5.2	16	5.6	15
R9	1.6	10	1.6	10	1.6	10
R10	4.4	13	5.1	11	6.2	10
R11	1.0	10	1.0	10	1.0	10
R12	4.3	12	5.2	10	5.2	10
R13	6.7	29	7.8	25	8.9	22
R14	1.0	10	1.0	10	1.0	10
R15	5.1	14	6.0	12	6.6	11
R16	4.4	14	5.1	12	5.5	11
R17	5.8	15	6.6	13	7.7	12
R18	3.4	10	3.4	10	3.4	10
R19	1.2	10	1.2	10	1.2	10
R20	3.8	11	4.2	10	4.2	10
R21/22	2.9	10	2.9	10	2.9	10
R23	1.0	10	1.0	10	1.0	10
R25	3.3	10	3.3	10	3.3	10
R26	3.2	15	3.7	13	4.0	12
R27	5.7	22	6.6	19	7.4	17
R28	1.0	10	1.0	10	1.0	10
R29	1.0	10	1.0	10	1.0	10
R30	1.6	10	1.6	10	1.6	10
R31	1.2	10	1.2	10	1.2	10
R32	1.4	10	1.4	10	1.4	10
R33	2.3	10	2.3	10	2.3	10
R37	3.2	14	3.7	13	4.1	11
R38	2.9	13	3.4	11	3.7	10
R39	1.0	10	1.0	10	1.0	10
R40	1.0	10	1.0	10	1.0	10
R41	1.0	11	1.0	10	1.0	10
R42	3.1	11	3.4	10	3.4	10
R43	1.0	10	1.0	10	1.0	10
R44	1.5	10	1.5	10	1.5	10
R46	1.0	10	1.0	10	1.0	10

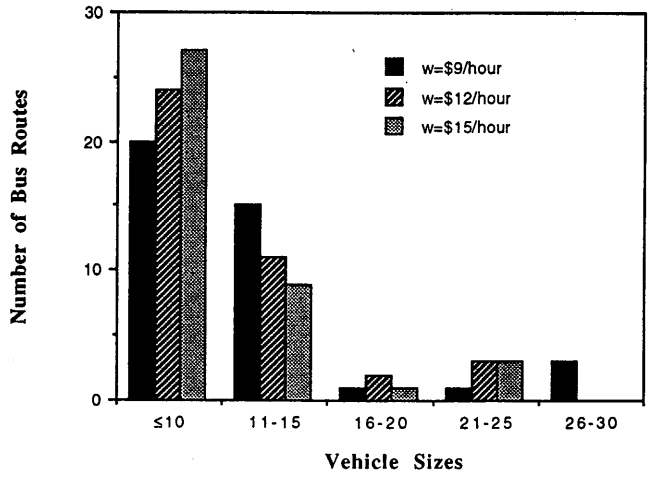


FIGURE 2 Distribution of vehicle sizes with different waiting time values.

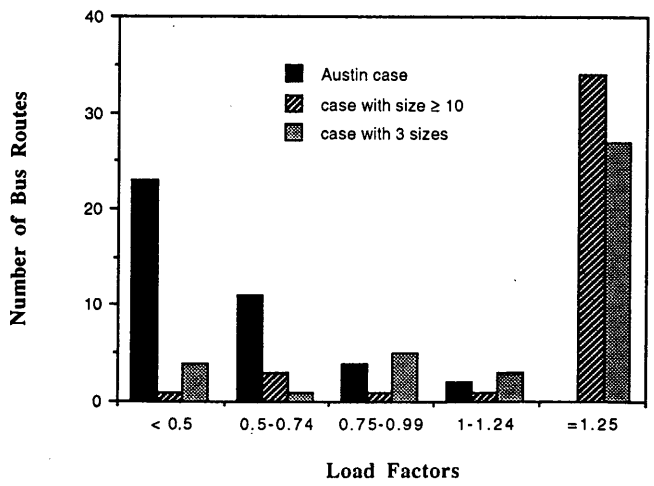


FIGURE 3 Distribution of load factors.

sizes in a system should be limited to avoid high maintenance cost and operational complexity. Incorporating variable vehicle sizes in the transit network design model will contribute to better and more realistic solutions to the problem. Although this paper demonstrates only the application of the vehicle sizing procedure to an existing bus system, the procedure has been implemented to enhance the network design procedure, AI-BUSNET.

Of course, the framework presented here incorporates a number of assumptions and relations that may be less applicable to certain locations than to others. These include the cost function components, such as relative waiting time and time cost of various operations. The methodology presented here provides a flexible framework to incorporate alternative assumptions and functional relations that may be tailored for specific cities. In future research, the passenger boarding and alighting time that affects the speed of different vehicle sizes will be considered. The optimal vehicle size that could be operated in different periods of the day will be investigated as well.

ACKNOWLEDGMENTS

The work presented in this paper is based on research supported by the State of Texas Governor's Energy Office (Oil Overcharge Funds) through the Southwest Region University Transportation Center. The authors are grateful to Hadi Baaj of Arizona State University for his help with the initial network design procedure that formed the basis of the code development for this work. Elise Miller's effort in the preparation of the Austin test data is greatly appreciated, as is the help of Kathryn Albee of Capital Metro, who provided boarding and alighting data as well as other information. The authors also wish to thank the anonymous referees for helpful suggestions.

REFERENCES

1. Van Nes, R., R. Hamerslag, and B. H. Immer. Design of Public Transport Networks. In *Transportation Research Record 1202,* TRB, National Research Council, Washington, D.C., 1988, pp. 74–83.
2. Baaj, M. H. The Transit Network Design Problem: An AI-Based Approach. Ph.d. thesis. Department of Civil Engineering, University of Texas at Austin, 1990.
3. Israeli, Y., and A. Ceder. Transit Network Design. Presented at 70th Annual Meeting of the Transportation Research Board, Washington, D.C., 1991.
4. Glaister, S. Competition on an Urban Bus Route. *Journal of Transport Economics and Policy,* Vol. 19, No. 1, 1985, pp. 65–81.
5. Oldfield, R. H., and P. H. Bly. An Analytic Investigation of Optimal Bus Size. *Transportation Research,* Vol. 22B, No. 5, 1988, pp. 319–337.
6. Walters, A. A. The Benefit of Minibuses. *Journal of Transport Economics and Policy,* Vol. 14, No. 3, 1980, pp. 320–334.
7. Mohring, H. Minibuses in Urban Transportation. *Journal of Urban Economics,* Vol. 14, 1983, pp. 293–317.
8. Bly, P. H., and R. H. Oldfield. Competition Between Minibuses and Regular Bus Service. *Journal of Transport Economics and Policy,* Vol. 20, No. 1, 1986, pp. 47–68.
9. Glaister, S. Bus Deregulation, Competition and Vehicle Size. *Journal of Transport Economics and Policy,* Vol. 20, No. 2, 1986, pp. 217–244.
10. Jansson, J. O. A Simple Bus Line Model for Optimization of Service Frequency and Bus Size. *Journal of Transport Economics and Policy,* Vol. 14, No. 1, 1980, pp. 53–80.
11. Walters, A. A. Externalities in Urban Buses. *Journal of Urban Economics,* Vol. 11, 1982, pp. 60–72.
12. Chang, S. K. Analytic Optimization of Bus Systems in Heterogeneous Environment. Ph.d. thesis. Department of Civil Engineering, University of Maryland, College Park, 1990.
13. Gwilliam, K. M., C. A. Nash, and P. J. Mackie. Deregulating the Bus Industry in Britain: the Case Against. *Transport Rev.,* Vol. 5, No. 2, 1985, pp. 105–132.
14. Furth, P. G., and N. H. M. Wilson. Setting Frequencies on Bus Routes: Theory and Practice. In *Transportation Research Record 818,* TRB, National Research Council, Washington, D.C., 1981, pp. 1–7.
15. Baaj, M. H., and H. S. Mahmassani. An AI-Based Approach for Transit Route System Planning and Design. *Journal of Advanced Transportation,* Vol. 25, No. 2, 1991, pp. 187–210.
16. Osuna, E. E., and G. F. Newell. Control Strategies for an Idealized Public Transportation System. *Transportation Science,* Vol. 6, 1972, pp. 57–72.
17. Larson, R. C., and A. R. Odoni. *Urban Operations Research.* Prentice Hall, Inc., Englewood Cliffs, N.J., 1981.
18. Bowman, L. A., and M. A. Turnquist. Service Frequency, Schedule Reliability and Passenger Wait Times at Transit Stops. *Transportation Research,* Vol. 15A, No. 6, 1981, pp. 465–471.
19. Baaj, M. H., and H. S. Mahmassani. TRUST: A Lisp Program for the Analysis of Transit Route Configuration. In *Transportation Research Record 1283,* TRB, National Research Council, Washington, D.C., 1990, pp. 125–135.

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

Real-Time Incident-Responsive System for Corridor Control: Modeling Framework and Preliminary Results

GANG-LEN CHANG, JIFENG WU, AND HENRY LIEU

An integrated optimal control model has been formulated to address the dynamic freeway diversion control process. An effective and efficient approach is developed for simultaneously solving diversion control measures, including on-ramp metering rates, off-ramp diversion rates, and green/Cycle ratios for traffic signals on a real-time basis. By approximating the flow-density relation with a two-segment linear function, the nonlinear optimal control problem can be simplified into a set of piecewise linear programming models and solved with the proposed successive linear programming algorithm. Consequently, an effective on-line feedback approach has been developed for integrated real-time corridor control. Preliminary simulation results with INTRAS for a sample network have demonstrated the merits of the proposed model and algorithm.

Real-time traffic control for freeway corridors has been the subject of increasing research in recent years because of nonrecurrent traffic congestion, especially that due to incidents. For a typical freeway corridor system consisting of a mainline freeway, a parallel surface street, and ramps, an integrated traffic control scheme should include three basic elements: (*a*) on-ramp metering, (*b*) off-ramp diversion, and (*c*) signal timing at surface street intersections. Another possible control measure, segmentwide speed limitation, has also been studied by European researchers (*1,2*). However, this measure appears impractical and is not recommended in the United States.

Although ramp metering and signal timing at street intersections have received much attention both in theoretical research and in real-world applications, relatively few attempts at real-time diversion control have been made. In fact, during a severe freeway incident, on-ramp metering usually is not adequate to relieve congestion effectively. Diverting some traffic to the parallel surface street to make full use of available corridor capacity will be necessary. However, to determine the optimal time-varying diversion flow rates when an incident is detected is a challenging task. Effective integration of on-ramp metering, off-ramp guidance, and signals on surface streets will be essential to ensure successful operation of the entire freeway corridor.

A review of the literature yielded some studies on corridor control. Cremer and Schoof (*1*) formulated a comprehensive corridor control model and proposed a heuristic on-line control algorithm on the basis of off-line solutions. Chang et al. (*3*) developed a prototype model for dynamic system-optimal control that provides coordinated operation between mainline and surface streets, but off-ramp flow diversion was not treated as a control measure. Earlier studies conducted by Gartner and Reiss (*4*) and Reiss et al. (*5–7*) made significant contributions to this topic, applying a creative multilevel control structure; however, no specific on-line diversion strategies at off-ramps were investigated. Other studies conducted by Goldstain and Kumar (*8*), Papageorgiou (*9,10*), Papageorgiou et al. (*11*), and Payne et al. (*12*) addressed only mainline freeway control.

Evidently, all three corridor control measures should be integrated in formulating corridor traffic models to achieve an optimal control. At the same time, an efficient solution algorithm must be designed for real-time applications. Only Cremer and Schoof (*1*) have developed an integrated optimal control model that includes all control variables. However, their proposed model is very difficult to solve, because it turns out to be a large-scale, nonlinear, mixed integer optimization problem, and thus real-time applications are precluded. As a result, their proposed solution algorithm is not able to handle all the control variables simultaneously. Hence, the model has to resort to a two-stage optimization procedure in which the upper level is made for route diversion and the lower level for ramp metering, speed limitation, and signaling of surface street intersections.

The need to make a decision in real time may dictate the selection of traffic models and an optimization algorithm. A certain trade-off between computational effort and an affordable level of modeling details is inevitable. Whereas analytical linearization techniques can be used to tackle nonlinear models, such algorithms generally are not efficient because of the intensive computation requirements. Hence, development of an effective algorithm is also at the core of an on-line diversion control system.

The purpose of this research is to formulate a set of linear or piecewise linear models that address all issues in an integrated incident-responsive corridor control system. The framework of a real-time corridor control system is outlined in the next section. By adopting a two-segment linear flow-density model, a set of piecewise linear optimal control models was developed, including on-ramp metering rates, off-ramp diversion rates, and green/Cycle (g/C) ratios for surface street signals. An efficient solution algorithm was developed that makes the proposed diversion control model sufficiently fast for use in real-time applications. Finally, an illustrative example conducted with a corridor simulation model (INTRAS) is outlined to demonstrate the efficacy of the proposed approach.

SYSTEM DESCRIPTION

Consider a typical freeway corridor, consisting of a unidirectional freeway segment, a parallel surface street/arterial, and a number of on- and off-ramps. The entire corridor is divided into N small segments, each having the same topological structure as shown in Figure 1 and containing only one pair of on- and off-ramps and one

G-L. Chang and J. Wu, Department of Civil Engineering, University of Maryland, College Park, Md. 20742. H. Lieu, IVHS Division, HSR-10, FHWA, 6300 Georgetown Pike, McLean, Va. 22101.

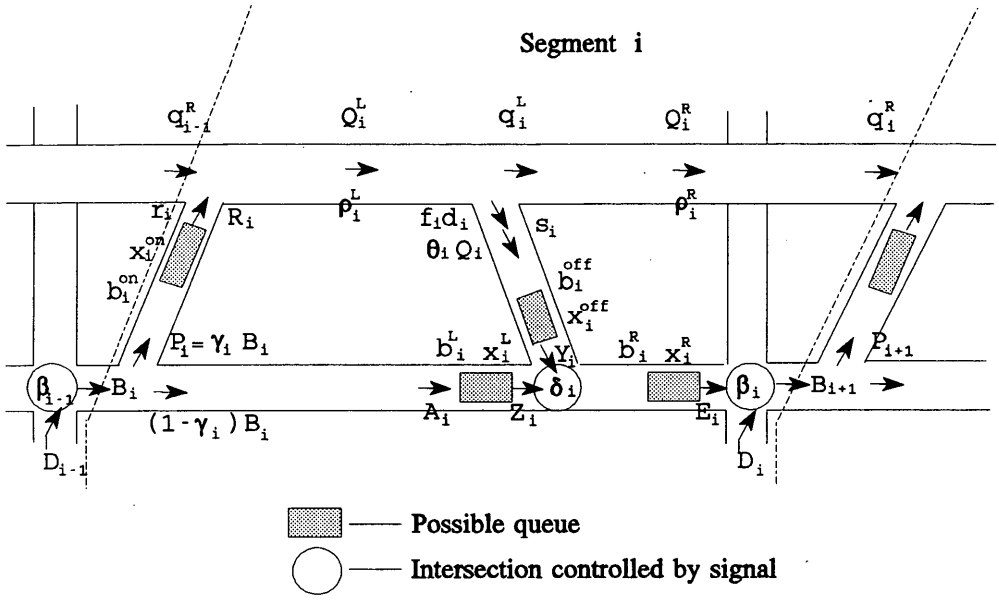


FIGURE 1 Arguments of Segment *i* and their notations.

crossing street. The off-ramp naturally divides the whole segment into two subsegments, the left part and the right part. Whereas each parallel street segment is divided into three subsections, it is assumed that (*a*) each on-ramp is close to its upstream surface street intersection and the distance between them is negligible from the congestion control perspective and (*b*) the distance between each street intersection and its adjacent upstream off-ramp intersection is short so that the link free flow time is negligible (but its waiting time may be significant because of limited street storage). Hence, for each surface street segment, the analysis of traffic dynamics may be focused mainly on two aspects: queuing at its downstream intersections and flow transition from the upstream to the downstream.

As indicated previously, metering at the on-ramps may not be sufficient if the incident is severe. Diverting freeway traffic via off-ramps may be necessary. However, the fraction of traffic to be diverted needs to be determined so that the prespecified system objective (e.g., corridor throughput) can be optimized. At the same time, signal timing at street intersections should be responsive to the diverting traffic to best serve traffic in the entire corridor. Hence, in the proposed model, we focus on formulating the interactions between those key control elements. All continuous variables are discretized into small equal intervals for analyses. The duration of each time interval is denoted as T. Notation and definitions of all relevant variables and parameters used hereafter are given in Table 1. With these defined variables, the principal issue is to solve the optimal on-line control strategies $\{R_i(k), d_i(k), \delta_i(k), \beta_i(k)\}$ according to traffic surveillance data and all other available information.

DYNAMIC TRAFFIC MODELS

Mainline Traffic Dynamics

If an equilibrium flow-density relationship prevails for each segment i, the traffic state on a segment can be represented with the mean segment density. A dynamic density evolution equation, according to the flow conservation law, can be formulated as follows:

$$\rho_i^L(k) = \rho_i^L(k-1) + \frac{T}{L_i^L l_i^L}[q_{i-1}^R(k) + e_i(k)R_i(k) - f_i(k)d_i(k) - \theta_i(k)Q_i^L(k) - q_i^L(k)] \qquad 1 \le i \le N \tag{1}$$

$$\rho_i^R(k) = \rho_i^R(k-1) + \frac{T}{L_i^R l_i^R}[q_i^L(k) - q_i^R(k)] \qquad 1 \le i \le N \tag{2}$$

where $e_i(k)R_i(k)$ and $f_i(k)d_i(k)$ express the actual flow rates at the entry on-ramp and exit off-ramp, respectively, of mainline Segment i, with $e_i(k)$ and $f_i(k)$ representing the actual system compliance parameters to the control measures $R_i(k)$ and $d_i(k)$. It is notable that $\theta_i(k)Q_i^L(k)$ is the regular flow rate exiting at off-ramp i, whereas $f_i(k)d_i(k)$ is the additional flow rate that needs to be diverted.

To capture the dynamic interrelations between adjacent segment flows, the transition flow between adjacent segments is taken as the average of the two adjacent segment boundary flows. More specifically, it is given by

$$q_i^L(k) = \frac{1}{2}\{[1 - \theta_i(k)]\,Q_i^L(k) - f_i(k)\,d_i(k) + Q_i^R(k)\} \qquad 1 \le i \le N \tag{3}$$

$$q_i^R(k) = \frac{1}{2}[Q_i^R(k) + Q_{i+1}^L(k) - e_{i+1}(k)R_{i+1}(k)] \qquad 1 \le i \le N-1 \tag{4a}$$

$$q_N^R(k) = Q_N^R(k) \tag{4b}$$

where Equation 4b, describing the mainline downstream boundary flow, is a special case.

Generally, the density dynamic equations are nonlinear because flow rate, $Q_i(k)$, is usually a nonlinear function of density, $\rho_i(k)$. That limits most of the equations' networkwide applications to off-line cases or to small networks because of considerable compu-

TABLE 1 Definition of System Variables

Network geometric and physical data	
L_i^L:	physical length (miles) of the left-section of freeway Segment i
l_i^L:	number of lanes of the left-section of freeway Segment i
L_i^R:	physical length (miles) of the right-section of freeway Segment i
l_i^R:	number of lanes in the right-section of freeway Segment i
L_i^s:	physical length (miles) of street Segment i
u_i^L, v_i^L:	parameters of flow-density relationship for the left section of freeway Segment i
u_i^R, v_i^R:	parameters of flow-density relationship for the right section of freeway Segment i
ρ^{cr}:	critical density (vehicle/lane/mile) at which freeway flow reaches its maximum
ρ^{max}:	jam density (vehicle/lane/mile) of the freeway
V_i:	normal flow speed (mph) of street Segment i
R_i^{max}:	maximum metering rate (mph) for On-ramp i
R_i^{min}:	minimum metering rate (mph) for On-ramp i
b_i^{on}:	maximum number of queue vehicles permitted on On-ramp i
b_i^{off}:	maximum number of queue vehicles permitted on Off-ramp i
b_i^R:	maximum number of queue vehicles permitted on the right section of arterial street Segment i
b_i^L:	maximum number of queue vehicles permitted on the left section of arterial street Segment i
C_i^{off}:	queue discharge rate (vehicle/green-hour) of Off-ramp i
C_i^R:	queue discharge rate (vehicle/green-hour) at the crossing intersection of arterial street Segment i
C_i^L:	queue discharge rate (vehicle/green-hour) at the off-ramp intersection of arterial street Segment i
Dynamic traffic demand	
$q_0^R(k)$:	flow rate (vph) entering the upstream end of freeway Segment 1 during interval k
$B_1(k)$:	flow rate (vph) entering the upstream end of street Segment 1 during interval k
$D_i(k)$:	flow rate (vph) approaching the corridor from the crossing street of segment i during interval k
$\theta_i(k)$:	proportion of traffic leaving freeway via off-ramp i (not including the diverted portion) during interval k
$\lambda_i(k)$:	the fraction of through traffic of $D_i(k)$
Incident data	
$\sigma_i^L(k)$:	capacity reduction factor due to an incident on the left section of freeway segment i, and $[1-\sigma_i^L(k)]100\%$ representing the reduced percentage of capacity
$\sigma_i^R(k)$:	capacity reduction factor due to an incident on the right section of freeway segment i, and $[1-\sigma_i^R(k)]100\%$ representing the reduced percentage of capacity

(*continued on next page*)

tational complexity. However, if a linear or piecewise linear $Q_i(k) \sim \rho_i(k)$ approximation is acceptable, it is obvious that the dynamic density equations (Equations 1 and 2) and the flow-density relations (Equations 3 and 4) become linear or piecewise linear, and the computational burden associated with the nonlinearity can be substantially alleviated. In fact, a two-segment linear flow-density model provides a good fit for freeway traffic operations according to recent research results (*13,14*). Figure 2 shows such a two-segment linear function, where ρ^{cr} represents the critical density at which the flow rate reaches its maximum and ρ^{max} is the jam density value.

Now suppose a two-segment linear flow-density function has been calibrated for each freeway Segment *i*:

$$Q_i^L(k) = [v_i^L(\rho) + u_i^L(\rho) \cdot \rho_i^L(k)]\sigma_i^L(k) \qquad 1 \le i \le N \tag{5}$$

$$Q_i^R(k) = [v_i^R(\rho) + u_i^R(\rho) \cdot \rho_i^R(k)]\sigma_i^R(k) \qquad 1 \le i \le N \tag{6}$$

TABLE 1 (*continued*)

Traffic Volumes (average flow rates in vph)

$q_i^L(k)$: flow rate from left section to right section of freeway Segment i during interval k

$q_i^R(k)$: flow rate from freeway Segment i to Segment i+1 during interval k

$Q_i^L(k)$: flow rate of left section of freeway Segment i during interval k

$Q_i^R(k)$: flow rate of right Section of freeway Segment i during interval k

$r_i(k)$: flow rate entering the freeway from on-ramp i during interval k

$s_i(k)$: flow rate (involving diverted traffic) leaving the freeway via off-ramp i during interval k

$B_i(k)$: flow rate entering upstream of street Segment i during interval k

$A_i(k)$: flow rate on the left section of street Segment i approaching the off-ramp junction during interval k

$Z_i(k)$: flow rate discharging from the off-ramp intersection on street Segment i during interval k

$E_i(k)$: flow rate discharging from downstream intersection of street Segment i during interval k

$Y_i(k)$: flow rate discharging from downstream Off-ramp i and merging into arterial street during interval k

$P_i(k)$: mean flow rate entering on-ramp i from the arterial street during interval k

System Parameters

$\gamma_i(k)$: proportion of the arterial street traffic entering on-ramp i during interval k

$e_i(k)$: ratio of actual flow rate entering the freeway to the metering rate for on-ramp i during interval k

$f_i(k)$: ratio of actual diverting flow rate to the calculated diversion rate for off-ramp i during interval k

$\alpha_i(k)$: platoon dispersion parameter of street segment i

$\eta_i(k)$: fraction of through traffic from street segment i to street segment i+1

$t_i(k)$: mean travel time to traverse the left section of street segment i during interval k

Control Variables to Be Solved

$R_i(k)$: metering flow rate (vph) for On-ramp i during interval k

$d_i(k)$: flow rate (vph) for freeway diversion at off-ramp i not including normal turning traffic during interval k

$\delta_i(k)$: g/c ratio for signal timing at off-ramp intersection i for the arterial traffic during interval k

$\beta_i(k)$: g/c ratio for signal timing at crossing street intersection i for the arterial traffic during interval k

State Variables

$\rho_i^L(k)$: mean density (vehicle/lane/mile) of the left section of freeway Segment i during interval k

$\rho_i^R(k)$: mean density (vehicle/lane/mile) of the right section of freeway Segment i during interval k

$x_i^{on}(k)$: average number of vehicles on on-ramp i during interval k

$x_i^{off}(k)$: average number of vehicles on off-ramp i during interval k

$x_i^R(k)$: average number of vehicles on the right section of street Segment i during interval k

$x_i^L(k)$: average number of queuing vehicles on the left section of street Segment i during interval k

where each of the coefficients $v_i^L(\rho)$, $u_i^L(\rho)$ and $v_i^R(\rho)$, $u_i^R(\rho)$ take two different values, depending on which regime the density values fall into, and $\sigma_i^L(k)$ and $\sigma_i^R(k)$ represent the corresponding capacity reduction factors due to the incident. For those segments not affected by the incident, parameters σ_i^L and σ_i^R naturally equal 1. Thus, Equations 1 through 6 constitute a set of piecewise linear models for the mainline segment density dynamics.

Ramp Traffic Dynamics

To simplify the presentation, the lengths of all the ramp links are assumed to be relatively short, and thus the traffic state at ramps can be represented with its dynamic queuing length evolution. If the travel time on any ramp is not negligible, a flow transition equation using Robertson's platoon dispersion model (*15*) can be employed

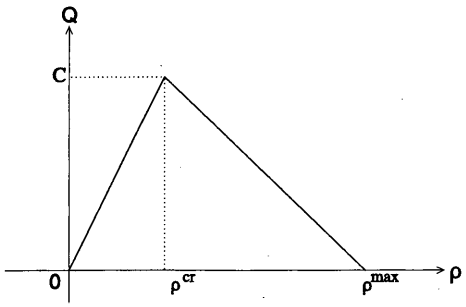


FIGURE 2 Two-segment linear flow-density model.

to capture the traffic variation as used for arterial segments. With such simplifications, the on-ramp queuing length dynamics can be formulated as follows on the basis of the conservation law:

$$x_i^{on}(k) = x_i^{on}(k-1) + [\gamma_i(k)B_i(k) - e_i(k)R_i(k)]T \qquad 1 \le i \le N \tag{7}$$

where $B_i(k)$ is the oncoming flow rate at the upstream end of surface street segment i and $\gamma_i(k)$ is the decimal proportion of $B_i(k)$.

Similarly, the queuing length dynamics for off-ramps can be given by

$$x_i^{off}(k) = x_i^{off}(k-1) + [f_i(k)\,d_i(k) + \theta_i(k)Q_i^L(k) - Y_i(k)]T \qquad 1 \le i \le N \tag{8}$$

The downstream off-ramp discharging flow, $Y_i(k)$, is affected by the signal timing at its downstream junction with the arterial. Assuming that a two-phase signal timing is applied for each off-ramp intersection and each crossing street intersection, the g/C ratios assignment $\{\delta_i(k)\}$ is the only control variable to be optimized. Given an off-ramp capacity, the average off-ramp discharging flow can be computed by

$$Y_i(k) = \min\left\{[1-\delta_i(k)]C_i^{off}, \frac{x_i^{off}(k-1) + [f_i(k)\,d_i(k) + \theta_i(k)Q_i^L(k)]\,T}{T}\right\} \qquad 1 \le i \le N \tag{9}$$

where the last item on the right-hand side expresses the average flow rate, if the queue on the off-ramp i can be cleared at the end of time interval k.

Surface Street Traffic Models

For each surface street segment i, the queues due to incidents shall be considered at both the segment's downstream off-ramp junction and the crossing street intersection. We assume that all intersections are under two-phase signal control. The g/C ratios for the arterial traffic, $1-\delta_i(k)$ and $\beta_i(k)$, are thus the two key control parameters to be optimized.

For the left-hand-side section of arterial segment i, it can be seen that the downstream flow $A_i(k)$, approaching the off-ramp junction, is determined to some extent by the upstream flow rate $[1-\gamma_i(k')]B_i(k')$ over the previous time interval k'. One of the most commonly used formulations for such a relation is the following platoon dispersion model (*15,16*):

$$A_i(k) = [1-\alpha_i(k)]B_i'(k) + \alpha_i(k)A_i(k-1) \qquad 1 \le i \le N \tag{10}$$

where

$$B_i'(k) = \{1-\gamma_i[k-t_i'(k)]B_i[k-t_i'(k)]\} \qquad 1 \le i \le N \tag{11}$$

and $t_i'(k)$ is the closest integer to the value $0.8t_i(k)/T$; $t_i(k)$ is the average travel time required for traversing surface street segment i when joining the queue within interval k; $\alpha_i(k)$ is the dynamic smoothing parameter.

With Equations 10 and 11, we can establish the interrelation between upstream and downstream flows on arterial segment i.

Similar to Equations 7 and 8, the queuing length dynamics for both possible downstream queues can be modeled as follows:

$$x_i^L(k) = x_i^L(k-1) + [A_i(k) - Z_i(k)]T \qquad 1 \le i \le N \tag{12}$$

$$x_i^R(k) = x_i^R(k-1) + [Z_i(k) + Y_i(k) - E_i(k)]T \qquad 1 \le i \le N \tag{13}$$

In the same fashion as Equation 9, discharging flows at the downstream end can be expressed with the following:

$$Z_i(k) = \min\left[\delta_i(k)C_i^L, \frac{x_i^L(k-1) + A_i(k)T}{T}\right] \qquad 1 \le i \le N \tag{14}$$

$$E_i(k) = \min\left\{\beta_i(k)C_i^R, \frac{x_i^R(k-1) + [Z_i(k) + Y_i(k)T}{T}\right\} \qquad 1 \le i \le N \tag{15}$$

Finally, the interactions between surface street flows in neighboring segments can be established through the following flow conservation relation at intersections:

$$B_{i+1}(k) = D_i(k)[1-\lambda_i(k)][1-\beta_i(k)] + \eta_i(k)E_i(k) \qquad 1 \le i \le N \tag{16}$$

where

$D_i(k)[1-\lambda_i(k)]$ = demand flow rate entering the corridor from the crossing street i during interval k,
$[1-\beta_i(k)]$ = g/C ratio for crossing street i, and
$\eta_i(k)$ = proportion of through traffic at the downstream end of surface street segment i during interval k.

FRAMEWORK OF REAL-TIME CONTROL APPROACH

Dynamic System Evolution

Using the notation defined previously, the dynamic state of the entire corridor (see Figure 3) at any time interval can be represented as

$$W(k) = F[W(k-1), C(k), G(k), H(k)]$$

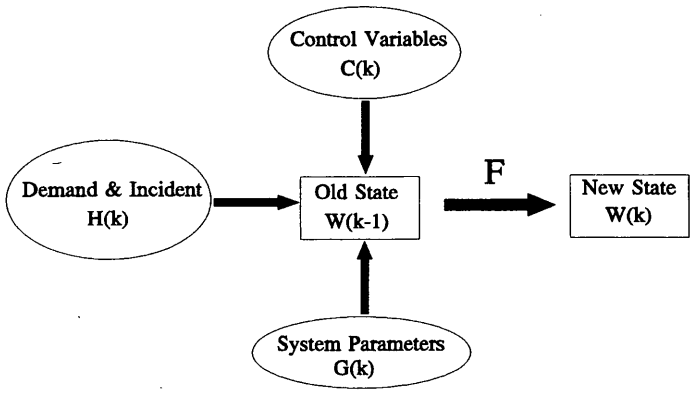


FIGURE 3 Dynamic traffic state evolution mechanism.

where

$W(k) \equiv \{\rho_i^L(k), \rho_i^R(k), x_i^{on}(k), x_i^{off}(k), x_i^L(k), x_i^R(k) \mid \text{for all } i\}$ represents the traffic state at interval k,

$C(k) \equiv \{R_i(k), d_i(k), \delta_i(k), \beta_i(k) \mid \text{for all } i\}$ denotes all control measures to be optimized,

$G(k) \equiv \{\gamma_i(k), e_i(k), f_i(k), \alpha_i(k), \eta_i(k), t_i(k) \mid \text{for all } i\}$ denotes the time-dependent system model parameters,

$H(k) \equiv \{D_i(k), \lambda_i(k), \theta_i(k), \sigma_i^L(k), \sigma_i^R(k) \mid \text{for all } i\}$ includes both the real time travel demand pattern and incident information; and

F = function determined by Equations 1 through 16.

The interrelations between principal modules as well as the control logic are shown in Figure 4. Key functions of each principal module are presented.

Optimal Control Model

The real-time incident-responsive control problem is to determine optimal control measures $C(k)$ (i.e., on-ramp metering rates, off-ramp diversion flow rates, and g/C ratios for surface street signals) for time interval k and its subsequent intervals, at the beginning of each time interval k, with the given time-varying travel demand pattern and incident information $\{H(t)\}$.

Depending on an operation agency's major concern, one may choose different control objectives, for example, to minimize the total travel time, waiting time, delay, vehicle-hours, vehicle-miles, or pollutant emissions. In this study, we select the total corridor throughput as the only measure of effectiveness, because it is the primary concern after an incident. The total corridor throughput includes vehicles exiting the corridor at surface street intersections and at the last control segment of both the freeway and the arterial. A mathematical expression of the corridor throughput is given by

$$\sum_k \left(\sum_{i=1}^{N} \{[1 - \eta_i(k)]E_i(k) + [1 - \beta_i(k)]\lambda_i(k) D_i(k)\}T + [q_N^R(k) + B_{N+1}(k)]T \right) \quad (17)$$

Assuming that all the related parameters are available, an optimal control model can then be calibrated and executed to maximize the objective function, subject to the dynamic constraints (Equations 1 through 16) and the boundary constraints given in Equations 18 through 25. Equation 18 represents metering rates, Equation 19 represents diverting flows, Equations 20 through 23 represent queuing lengths, and Equations 24 and 25 represent g/C ratios in signal timing.

$$R_i^{min} \le R_i(k) \le R_i^{max} \qquad 1 \le i \le N \quad (18)$$

$$0 \le d_i(k) \le C_i^{off} - \theta_i(k) Q_i^L(k) \qquad 1 \le i \le N \quad (19)$$

$$0 \le x_i^{on}(k) \le b_i^{on} \qquad 1 \le i \le N \quad (20)$$

$$0 \le x_i^{off}(k) \le b_i^{off} \qquad 1 \le i \le N \quad (21)$$

$$0 \le x_i^L(k) \le b_i^L \qquad 1 \le i \le N \quad (22)$$

$$0 \le x_i^R(k) \le b_i^R \qquad 1 \le i \le N \quad (23)$$

$$0 \le \delta_i(k) \le 1 \qquad 1 \le i \le N \quad (24)$$

$$0 \le \beta_i(k) \le 1 \qquad 1 \le i \le N \quad (25)$$

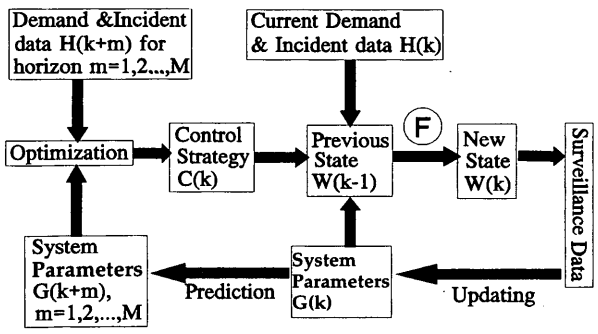


FIGURE 4 Flowchart of real-time corridor control logic.

Because the proposed model has addressed all essential aspects of corridor control, theoretically one can extend the optimal time horizon to the entire control period of interest. However, because of the computing burden and the increasing uncertainty for predicted demands, we recommend that optimal control be extended at a short

time interval preceded by a commonly used rolling time horizon method (*17–19*).

Parameter Updating and Prediction

Under nonrecurrent freeway congestion, most system parameters may vary substantially with time, especially those represented in $G(k)$. Those parameters must be identified and predicted before the optimization of the control measures $C(k)$.

A large body of dynamic prediction methods, such as time-series analysis and the Kalman filtering algorithm, which allow for performing parameter estimation and updating with the on-line surveillance data, exists in the literature. Example applications of these techniques in traffic analysis can be found elsewhere (*20,21*).

Optimization Algorithm

The above optimal control model formulated with Equations 17 through 25 is basically a nonlinear programming (NLP) problem. However, because these nonlinear constraints as well as the flow-density relation can be viewed as two connected linear functions in nature, the complex NLP becomes a series of linear programming (LP) models on alternative linear constraint regions and thus can be solved with efficient LP algorithms. To ensure that all optimal control measures can be generated in a sufficiently short time for real-time applications, a successive linear programming (SLP) algorithm was developed for this study. Step-by-step procedures for implementation follow.

SLP Algorithm

Step 1

- According to the current traffic state, $W(k-1)$, and the last control measure, $C(k-1)$, select the appropriate linear segment for use in Equations 5, 6, 9, 14, and 15
- Impose the region constraints for the equivalent LP model based on the selected segment of each two-piece function.
- Solve the LP model to produce a control solution, $C(k)$.

Step 2

Compute the resulting traffic state, $W(k)$, with this initial control measure, $C(k)$.

Step 3

Check whether any of Equations 5, 6, 9, 14, or 15 yields the identical value with either of its two functions under the projected $W(k)$ and the implemented $C(k)$. If not, this solution $C(k)$ is optimal and the search process terminates. Otherwise, go to Step 4.

Step 4

- Replace those inequality constraints for any of Equations 5, 6, 9, 14, or 15 with their complement piece of functions.
- Solve the modified LP model to generate a new $C(k)$.

Step 5

Check whether the objective function value (i.e., total throughput) has been improved after changing the constraints. If not, the current measure, $C(k)$, is the optimal solution, and the search process can be terminated. Otherwise, return to Step 2.

Discussion of SLP Algorithm

The algorithm starts with an initial LP model that is based on the current traffic state. Appropriate linear flow-density functions are then determined for Equations 5 and 6 according to the current link density. If the density is less than its critical value, ρ^{cr}, the left-segment linear function is selected; otherwise, the right-segment linear function is used. Similarly, in Equations 9, 14, and 15, the algorithm will select one of two complement functions that reduces its value with the current traffic state data. The initial LP model thus can be constructed by incorporating the boundaries of the selected linear functions (i.e., Equations 14 and 15) into its constraints.

To examine whether the initial LP formulations actually yield the optimal solution, Step 3 is taken to compare the resulting values from both of the two linear functions in Equations 5, 6, 9, 14, and 15, according to the traffic state projected in Step 2 with the initial control measure, $C(k)$. If none of the two linear functions in these equations are equal, the imposed linear constraint region contains the optimal solution, and therefore the obtained LP solution is the optimal control measure. However, very often some of the two complement linear functions may yield the same value as the given LP solution, indicating that some of the incorporated boundaries for the linear functions become binding constraints at the solution point. Apparently, the optimal control solution is not within the imposed linear region; it may lie on the boundaries or beyond the linear region. Therefore, Step 4 is executed to check whether the optimal solution point is on any of the boundaries. Through replacement of the binding inequality constraints, Step 4 and Step 5 are executed to see whether some improvement can be made. If not, the previously obtained LP solution is optimal and is located on a boundary point of the imposed linear region. The proposed algorithm repeats this procedure to successively solve a series of LP problems until the optimal control solution is reached.

The following key features make the SLP algorithm especially suitable for use in real-time implementations:

- It is convenient to implement because only LP problems need to be solved, and its formulation need not have any derivatives.
- The LP solution improves monotonically from one iteration to next.
- The algorithm often terminates within a small number of iterations and no loop may exist during the iteration process.

The SLP approach has been calibrated and tested successfully for use on an integrated real-time ramp metering control system (*22*).

Refined Real-Time Feedback Control Procedure

As indicated in Figure 4, by computing the control measures $C(k)$, the system parameters $G(k+m)$ $(m = 0, 1, \ldots, M-1)$ will all be updated and predicted with real-time surveillance data. That feedback is necessary to achieve adaptive on-line control. Such feed-

back information is critical, especially when nontrivial bias exists for previously predicted parameters.

According to the rolling time horizon logic, if the accuracy of predicted parameters is within an acceptable range, the optimization need not be executed in subsequent intervals within the time horizon. The control logic, along with the implemented rolling time horizon concept, is shown in Figure 5.

NUMERICAL EXAMPLE

To illustrate the proposed model as well as the solution algorithm for potential real-time applications, an example scenario with simulation was developed.

Simulation Tool

To evaluate the proposed real-time incident-responsive control, a traffic simulation model must be used to provide the surveillance data required by the control model. The microscopic simulation model INTRAS (*23*), developed by FHWA for freeway corridor simulation analysis, was selected for preliminary evaluation. In the numerical example, INTRAS was used to execute the control measures generated by the SLP algorithm, including on-ramp metering rates, off-ramp diversion rates, and g/C ratios for surface street signals. Interactively, the on-line surveillance data produced by INTRAS are fed back to the SLP algorithm to compute the new control measures for subsequent time intervals. The INTRAS model also provides all the most commonly used measures of effectiveness (MOEs) for evaluation.

Network and Case Design

Figure 6 shows a sample corridor network. The network consists of four identical segments. Each contains a two-lane freeway with auxiliary lanes, a three-lane arterial street, one on-ramp, one off-ramp, and one crossing surface street. All street intersections are signal controlled.

Each time interval is 3 min long. A total of 20 time intervals were used in the INTRAS simulation runs. An incident was assumed to occur on the downstream end of the second segment from time intervals 3 to 8. A detailed description of the experimental design and its implementation on INTRAS can be found elsewhere (*24*).

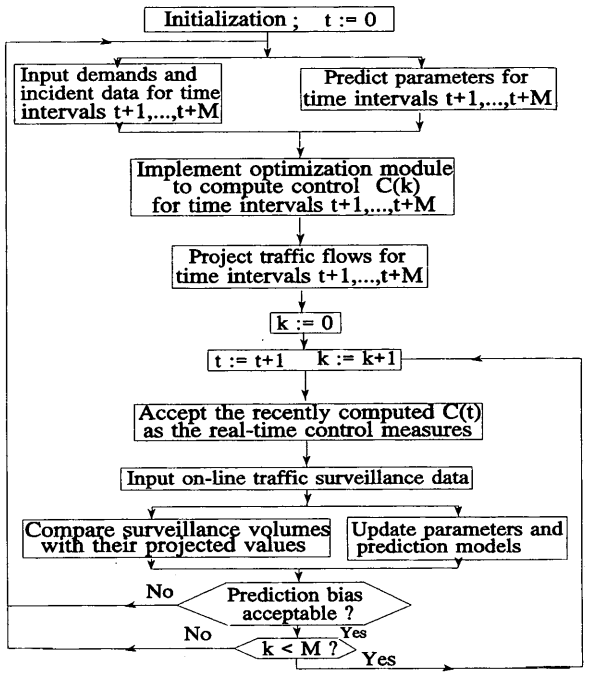


FIGURE 5 Refined real-time feedback control logic.

The following three cases with different freeway traffic loadings were simulated in the experiment: Case 1, high freeway demand (3,300 vph) with a severe incident (50 percent reduction in capacity); Case 2, moderate freeway demand (3,000 vph) with a severe incident (50 percent reduction in capacity); and Case 3, low freeway demand (2,600 vph) with a minor incident (25 percent reduction in capacity).

On surface streets, an entry volume of 1,200 vph is assumed at the upstream control boundary of the arterial street. On each crossing street, the entry flow is assumed to be maintained at 600 vph, with 30 percent of vehicles turning right to the arterial street. For evaluation of the model's effectiveness, four scenarios were developed for each case: (*a*) no control, (*b*) on-line strategy produced by the proposed model, (*c*) moderate control (Control A), and (*d*) intensive control (Control B). Control A and Control B for each case are briefly described:

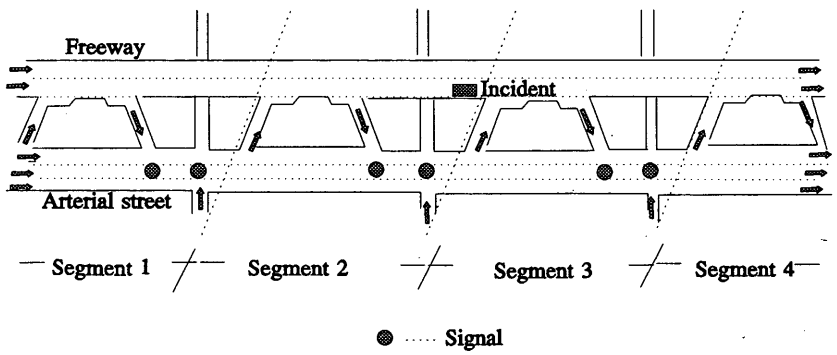


FIGURE 6 Corridor network example.

TABLE 2 Total Throughput from Simulation Results

Scenarios	Case 1		Case 2		Case 3	
	throughput	throughput increase	throughput	throughput increase	throughput	throughput increase
No control	6893	0	6616	0	6192	0
Control -A	6936	43	6624	8	6240	48
Control -B	6936	43	6610	0	6259	67
On-line control	7218	325	6849	233	6605	413

- Case 1: Control A is a moderate diversion control, whereby the first two on-ramps are closed and 10 percent of freeway vehicles are assumed to divert via the off-ramps of the first two corridor segments. The g/C ratio of traffic signals is adjusted accordingly to respond to the diverting traffic. For the intensive control case, Control B, 20 percent of freeway traffic will divert at the second two subsequent off-ramps.
- Case 2: Control A represents a moderate diversion control under which the first two on-ramps are closed and 10 percent of freeway traffic will divert via the second off-ramp. The g/C ratio of traffic signals is adjusted accordingly to respond to the diverting traffic. For the intensive control case, Control B, 20 percent of freeway traffic will divert at the second off-ramp.
- Case 3: Because of the light freeway traffic demand and the minor incident involved, only the second on-ramp will be closed for Control A. As for Control B, in addition to the ramp closure, 5 percent of freeway traffic will divert at the second off-ramp. The g/C ratios of traffic signals are adjusted to respond to the diverting traffic.

All diverted freeway traffic is assumed to reenter the freeway at the nearest downstream on-ramp beyond the incident, if the on-ramp capacity allows. If the nearest on-ramp reaches its capacity, diverted traffic will proceed on the surface street and reenter the freeway via the next available on-ramp. Note that neither Control A nor Control B, in any case, was randomly selected for evaluation. Each actually represents a reasonable control plan produced by senior freeway operation engineers. In addition, g/C ratios of all traffic signals are adjusted or optimized on the basis of expected diverting traffic patterns.

Simulation Results

Although a wide variety of MOEs, including total vehicle-miles, total vehicle-minutes, average travel speed, and total delay, are available in the output of INTRAS, most of them have some limitations for use in guiding real-time corridor control. For instance, high vehicle-miles may be accompanied by high vehicle-minutes or a low average speed. Similarly, the objective of maximizing speed or minimizing total delay may imply fewer vehicles being served. Therefore, total corridor throughput remains a more reasonable MOE for evaluation. Table 2 indicates the total corridor throughput resulting from the simulation runs for all cases and scenarios.

It is evident that the proposed on-line control algorithm produced the highest total corridor throughput in all cases (Table 2).

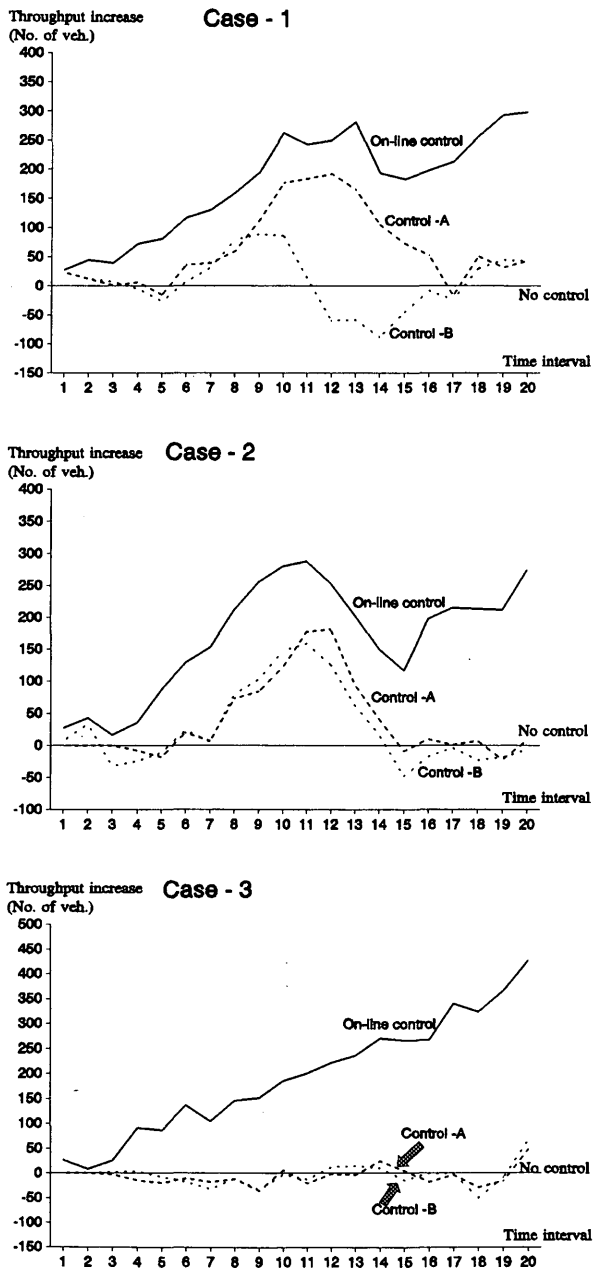


FIGURE 7 Cumulative throughput increase versus no control.

"Throughput increase" in Table 2 denotes the total corridor throughput increases for each scenario compared with the no-control scenario at the end of the 60-min control period. To better demonstrate the superior performance of the proposed approach, the time-varying cumulative throughput increase for 20 time intervals in four control scenarios is presented in Figure 7, for the three illustrative cases. Except for Case 3 (low freeway traffic demand with a minor incident) and Control B of Case 1, all control scenarios produce higher corridor throughput than the no-control case during the incident period. Among the four scenarios, the on-line control performs best during the entire control period.

CONCLUSIONS

A modeling framework, as well as formulations for its use for real-time freeway incident control, was presented. The proposed model is capable of generating the optimal control target for all available control measures, including on-ramp metering rates, off-ramp diversion fractions, and signal timing plans at surface street intersections in an entire corridor. By approximating the nonlinear flow-density relation with two linear complement functions, the study developed an efficient algorithm, named SLP, to solve the proposed integrated freeway diversion control model. Preliminary results from the numerical example with INTRAS have demonstrated the potential efficacy of the proposed modeling system as well as algorithm.

Further research along this line, conducted at the University of Maryland, incorporates an advanced signal control at surface intersections and extension of control boundaries to multiple surface streets. Thus, both freeways and surface street networks can be operated under a consistent control strategy.

REFERENCES

1. Cremer, M., and S. Schoof. On Control Strategies for Urban Traffic Corridors. In *Proceedings of IFAC Control, Computers, Communications in Transportation,* Paris, 1989.
2. Cremer, M., and S. Fleischmann. Traffic Responsive Control of Freeway Networks by a State Feedback Control. In *Transportation and Traffic Theory* (Gartner and Wilson, eds.), Elsevier, New York, 1987, pp. 213–219.
3. Chang, G.-L., P.-K. Ho, and C.-H. Wei. A Dynamic System-Optimal Control Model for Commuting Traffic Corridors. *Transportation Research C,* Vol. 1C, pp. 1–12.
4. Gartner, N. H., and R. A. Reiss. Congestion Control in Freeway Corridors: The IMIS System. In *Flow Control of Congested Networks* (A. R. Odoni et. al., eds.), Springer-Verlag, 1987.
5. Reiss, R. A., et al. Development of Traffic Logic for Optimizing Traffic Flow in an Intercity Corridor. Final report. U.S. Department of Transportation, 1978.
6. Reiss, R. A., et al. Algorithm Development for Corridor Traffic Control. In *Traffic, Transportation, and Urban Planning* (Vol. 2), George Goodwin, London, 1981.
7. Reiss, R. A., N. H. Gartner, and S. L. Cohen. Dynamic Control and Traffic Performance in a Freeway Corridor: A Simulation Study. *Transportation Research A,* Vol. 25A, 1991, pp. 267–276.
8. Goldstain, N. B., and K. S. P. Kumar. A Decentralized Control Strategy for Freeway Regulation. *Transportation Research B,* Vol. 16B, 1982, pp. 279–290.
9. Papageorgiou, M. A New Approach to Time-of-Day Control Based on a Dynamic Freeway Traffic Model. *Transportation Research B,* Vol. 14B, 1980, pp. 349–360.
10. Papageorgiou, M. A Hierarchical Control System for Freeway Traffic. *IEEE Transactions on Automatic Control,* Vol. AC-29, 1983, pp. 482–490.
11. Papageorgiou, M., J.-M. Blosseville, and H. Hadj-Salem. Modeling and Real Time Control of Traffic Flow on the Southern Part of Boulevard Peripherique in Paris; Part II: Coordinated On-Ramp Metering. *Transportation Research A,* Vol. 24A, 1990, pp. 361–370.
12. Payne, H. J., D. Brown, and J. Todd. *Demand Responsive Strategies for Interconnected Freeway Ramp Control Systems,* Vol. 1: Metering Strategies. Verac Inc., 1985.
13. Hall, F. L. Empirical Analysis of Freeway Flow-Density Relationships. *Transportation Research A,* Vol. 20A, 1986, pp. 197–210.
14. Banks, J. H. Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations. In *Transportation Research Record 1225,* TRB, National Research Council, Washington D.C., 1989, pp. 53–60.
15. Robertson D. I. *TRANSYT: A Traffic Network Study Tool.* Road Research Laboratory Report LR253, Crowthorne, England, 1969.
16. Axhausen, K. W., and H.-G. Körling. Some Measurements of Robertson's Platoon Dispersion Factor. In *Transportation Research Record 1112,* TRB, National Research Council, Washington D.C., 1987, pp. 71–77.
17. Boillot, F., et al. Optimal Signal Control of Urban Traffic Networks. *Proc. 6th International Conference on Road Traffic Monitoring and Control,* IEE, London, 1992.
18. Gartner, N. H. OPAC: A Demand-Responsive Strategy for Traffic Signal Control. In *Transportation Research Record 906,* TRB, National Research Council, Washington D.C., 1983, pp. 75–80.
19. Henry, J. J., et al. The PRODYN Real Time Traffic Algorithm. *Proc., 4th IFAC-IFIP-IFORS Conference on Control in Transportation Systems,* Baden-Baden, Germany, 1983.
20. Lu, J. Prediction of Traffic Flow by an Adaptive Prediction System. In *Transportation Research Record 1287,* TRB, National Research Council, Washington D.C., 1991, pp. 54–61.
21. Stephanedes, Y. J., E. Kwon, and P. Michalopoulos. On-Line Diversion Prediction for Dynamic Control and Vehicle Guidance in Freeway Corridors. In *Transportation Research Record 1287,* TRB, National Research Council, Washington D.C., 1991, pp. 11–19.
22. Wu, J. *Development and Evaluation of Real-Time Ramp Metering Algorithms.* FHWA, U.S. Department of Transportation, 1993.
23. Wicks, D. A., and E. B. Lieberman. *Developing and Testing of INTRAS, a Microscopic Freeway Simulation Model.* Report FHWA/RD-80/106-109, Vol. 1–4, FHWA, U.S. Department of Transportation, 1980.
24. Lieu, H. C. *An Integrated Diversion Control System for Freeway Corridors.* Ph.D Dissertation. Department of Civil Engineering, University of Maryland, College Park, 1993.

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

Fully Incremental Model for Transit Ridership Forecasting: Seattle Experience

YOUSSEF DEHGHANI AND ROBERT HARVEY

Traditionally, comprehensive multimodal regional models have been developed to conduct travel forecasts for both highway and transit projects in major metropolitan areas throughout the United States. These models generally have failed to provide accurate, detailed forecasts for existing and proposed facilities, and unrealistic expectations can be placed on these comprehensive (super) regional models. Most of the large regional models in the United States are based on scanty transit ridership information compared with the amount of data available for the predominating automobile users. Transit-component validation of the models usually has been for aggregate market shares and volumes at a few screenlines. Under these circumstances, it is no wonder that comprehensive regional models fail to provide accurate ridership forecasts for specific transit lines. The transit ridership modeling for the Regional Transit Authority (RTA) in Seattle overcomes usual limitations by relying on comprehensive regional models only for regional growth, highway congestion, and regional model coefficients. RTA modeling is structured so that transit ridership results are based on observed origins and destinations of transit users, observed transit line volumes, and a realistic simulation of observed transit service characteristics. External changes, in demographics and in highway costs, are staged into the process in distinct phases before estimating the impacts of incremental changes in transit service. The RTA transit ridership model is simple and fully incremental. The modeling system was validated on the basis of base year comparisons with transit ridership counts, and on a 1992 to 1985 backward "forecast" of transit demand.

This paper describes a fully incremental transit ridership model designed for efficient and expedient evaluation of transit project planning and ridership forecasting analyses for the Regional Transit Authority (RTA) in Seattle. The RTA model is simple and uses incremental methods to estimate new shares both for primary modes (i.e., automobile and transit) and transit submodes (i.e., automobile and walk access). The incremental form is highly desirable because it is directly based on observed data that describe current conditions, instead of relying solely on models to estimate these conditions. The model can be used to conduct systemwide or corridor-level transit planning and patronage forecasting analyses. The RTA modeling system does not require any mode choice model calibration; it is an adjunct to the existing regional model with locally appropriate time and cost coefficients.

The incremental model is more realistic than the comprehensive regional synthetic models for transit ridership forecasting analysis because it

- Is based directly on observed instead of estimated baseline travel patterns of transit users;
- Allows concentration of effort on transit network analysis for studies whose primary goals are questions about alternative transit networks;
- Is more conducive to separate evaluation of changes in population and employment, highway congestion and cost, and transit services through the three stages of the forecasting process;
- Lends itself readily to intermediate evaluation by focusing on direct comparison instead of complete simulation of travel behavior; and
- Eliminates often laborious and time-consuming calibration of subchoice models because it does not require replication of base year travel patterns.

A model validation effort was conducted to address two points in time, 1985 and 1992. It included a validation of 1992 conditions as well as a backcast from 1992 to 1985.

The paper includes a discussion of the role of regionally based synthetic models and the history of staged incremental transit modeling at Seattle Metro. The RTA three-staged fully incremental ridership forecasting model is described, and results of the base year 1992 validation and 1985 backcast analysis are presented. Finally, some conclusions and incremental model limitations, as well as a few areas of future research, are offered.

Y. Dehghani, Parsons Brinckerhoff, Inc., 999 Third Avenue, Suite 801, Seattle, Wash. 98104. R. Harvey, Regional Transit Authority, 821 Second Avenue, Seattle, Wash. 98104.

ROLE OF REGIONALLY BASED SYNTHETIC MODELS

Traditionally, synthetic models have been used to predict multimodal travel demand on various highway and transit facilities. The conventional four-step synthetic method entails using separate models for (*a*) determination of total person trips in each zone, (*b*) distribution of total interzonal trips, (*c*) prediction of share of travel by each mode, and (*d*) estimation of demand volumes on transit and highway facilities. Supplementary subarea models and procedures are usually used to generate detailed link-specific travel demand forecasts. The subarea (synthetic) models use incremental methods to the extent that they directly use available base year traffic counts in their application phases.

In the Puget Sound region, there are about 15 separate but interdependent transportation models. Only one, maintained by the Puget Sound Regional Council (PSRC), is a regional synthetic model, complete with feedback loops on land use. The remaining subarea models provide focused analysis on local transportation supply issues, primarily those related to street and highway capacity.

Unrealistic expectations are often placed on comprehensive (super) regional models. At its best performance, a regional model can be expected to generate reasonably reliable travel forecasts only along supercorridors and among very large districts. The inability of regional models to produce detailed project-specific travel demand forecasts probably has been a major factor in the proliferation of subarea models.

Previous research findings indicate that generation of reasonable land use forecasts is possible only at a superdistrict level (*1*). Consequently, breaking a geographic area into several smaller districts does not necessarily lead to a more accurate regional model. Past research (*2,3*) also indicates that the larger (level) district regional models will facilitate efficient integration of appropriate algorithms to allow full interactions and equilibrium among land use, travel time, and cost variables, as well as resulting travel demand.

The practical function of superregional models appears to be as a base for input to more focused application models, instead of for direct application to transportation studies. An auxiliary function of the superregional models has been to systematize a regional information data base, including land use and demographic data and forecasts. Direct application of the models to transportation studies increasingly has been limited to "big picture" questions on regional air quality, regional travel demand, or long-term land use visions. Because of the limitations of supermodels, the need to develop simple models that are operationally more efficient and sensitive to project settings—as well as able to produce more realistic detailed travel demand forecasts on both transit and highway facilities—has become more evident than ever.

HISTORY OF STAGED INCREMENTAL TRANSIT MODEL AT SEATTLE METRO

Work by Brand and Benham (*4*) led, in 1985, to the Metro staff's consideration of a "quick-responsive incremental travel demand forecasting method" based on the concept of staged forecasting analysis. In 1986 Metro installed "logit mode-choice equations for pivot-point analysis" on EMME/2 software (R. Harvey, unpublished data, 1986). These equations were translated from descriptions by Ben-Akiva and Atherton (*5*), Koppleman (*6*), Nickesen et al. (*7*), and many others.

In 1988, Metro clarified the relationship between its incremental transit forecasting model and the regional model at PSRC (R. Harvey, unpublished data, 1988). At that time, the method included (*a*) four distinct stages for ridership forecasting analysis, (*b*) an incremental mode-of-access component, (*c*) the use of regional person trip tables to represent growth (in lieu of a Fratar-type calculation), and (*d*) direct use of the regional model coefficients on travel time and cost variables (R. Harvey, unpublished data, 1989).

In 1991 a team of Metro staff and Parsons Brinckerhoff consultants updated the process for the Regional Transit Project, resulting in a *Travel Forecasting Methodology Report* (*8*) in October 1991. Changes included (*a*) synthetic access-mode and automobile-occupancy submodels with borrowed and adjusted coefficients, (*b*) return to a Fratar-type matrix balancing for growth, (*c*) consolidation of cost and highway time impacts in the staged forecasting analysis, (*d*) an increase in the number of zones, and (*e*) more emphasis on trip purpose in the model structure. The 1991 version of the RTA model was a combination of incremental approaches previously used by Seattle Metro and J. M. Ryan of Parsons Brinckerhoff, Inc. Before the Seattle application (*8*), Ryan had used incremental methods for ridership forecasting analysis in a number of cities in the United States, including San Francisco (*9*), Baltimore (*10*), and Honolulu (*11*), for evaluation of major transit investments.

In 1993, the process was again refined, reflecting a renewed commitment to integration with the regional model. Transit operators completed a new set of comprehensive ridership surveys and counts, providing a new base for the model. Refinements included (*a*) use of regional model coefficients for consistency, (*b*) return to regional trip tables for consideration of regional growth, (*c*) addition of a fully integrated incremental model to represent transit and automobile submodes, and (*d*) further refinement of the zone structure. An updated *Travel Forecasting Methodology Report* (*12*) summarizing these changes and the new transit surveys was published in November 1993.

Presently, there are well-established markets for park-and-ride and group ride activities in the Seattle area. Potential difficulties with the use of an incremental transit access component, usually considered to be related to zero or 100 percent shares in the cells, are not problematic with the RTA model application. The following factors allow the RTA model to avoid the problem:

- There are more than 50 park-and-ride lots within the RTA area.
- The automobile-access definition used from the surveys included all automobile access to transit.
- There is extensive peak-period coverage with local bus service throughout the RTA area. Almost all park-and-ride service is provided by groups of separate local routes that come together at lots before beginning the express portion of the trip.
- Mode-of-access shares were calculated from the aggregation of survey data to larger districts, especially at the attraction ends.
- A boundary has been used (i.e., 10 to 90 percent) for calculation of the access shares. The precaution is both practical and reasonable because of the four considerations just noted.

The reasons for changing the transit access submodel to an incremental form again in 1993 related primarily to difficulties encountered in trying to match base access shares to important markets, such as downtown Seattle, with a synthetic component (*8*). The availability of a new set of access-mode share data from the 1992 surveys suggested that an incremental approach would be preferable.

STAGED INCREMENTAL FORECASTING ANALYSIS

Underlying methods and assumptions used in the 1993 RTA three-stage fully incremental ridership forecasting model are now described.

Incremental Logit Model Equations

The incremental form of the logit model is derived from the standard logit formulation. Ben-Akiva and Lerman state:

> . . . using elasticities is one way to predict changes due to modifications in the independent variables. For the linear-in-parameters multinomial logit model there is a convenient form known as the *incremental logit* which can be used to predict changes in behavior on the basis of the *existing choice probabilities of the alternatives and changes in variables* that obviates the need to use the full set of independent variables to calculate the new choice probabilities. (*13*)

Mode-specific constants in a synthetic model theoretically represent the effects of unmeasurable attributes and usually capture more than two-thirds of explanatory power in logit models (*14,15*). In actuality, these constants are quite large, and they compensate for all types of errors in synthetic models, even network coding idiosyncrasies. They are used as overall adjustment factors to move the model results close to targeted regional totals; they typically range

as high as 50 to 150 min. of equivalent in-vehicle time. Without these constants, synthetic models could never replicate even the regional totals for a base year. The mode-specific constants fall out of the computations in the incremental logit model.

Recursive Logit Model

The recursive "nested" form of the logit model is less restrictive; therefore, it is more attractive than the simultaneous structure to travel demand practitioners. However, there is no convincing statistical evidence or professional consensus on using a particular recursive (nesting) structure.

In the absence of a theoretically sound behavioral theory to describe mode choice formation and a consensus on the form of a recursive logit model, the RTA uses an implicit recursive structure only, because of computational convenience in using the incremental logit model to estimate new shares for both the primary and subchoice modes. The RTA model also uses a coefficient of 1.0 for the LogSum variable, which is consistent with the PSRC simultaneous logit model forms.

For an incremental logit model application, primary modes (i.e., transit and automobile) are represented by subchoices. For the transit mode, the subchoice is between access to transit by walking or by automobile. For the automobile mode, the subchoice is between single and multiple occupancy for commute trips. For noncommute trips, all automobile submodes are combined into a single automobile mode.

LogSum Variable

The natural logarithm of the denominator of the standard logit model is a single "inclusive" index, I_m, (*16*), indicating the desirability of main mode m, taking into account the attributes of access modes. This index is often called LogSum and calculated from

$$\text{LogSum} = \log \{\text{SUM}_j^m [\exp(V_j)]\}$$

where V_j is the utility of mode i in choice set m ($j = 1, 2, 3, \ldots, i, \ldots, m$) and contains measurable components of transportation systems, such as travel time and cost as well as socioeconomic attributes of trip makers.

Derivation of Changes in LogSum Variable

Contrary to a synthetic subchoice model, new shares for submodes are computed using incremental methods. That requires derivation of an appropriate formula to compute the difference in the values of the LogSum variable for submodes (e.g., DIFF LogSum^i for the mode of access). The derivation process starts by using the definition of difference in the LogSum values and ends up with a simple formula, as follows:

$$\text{DIFF LogSum}^m = \ln \{\text{Sum}_i^n [S_i * \exp (\text{DIFF } V_i)]\} \quad (1)$$

where

DIFF LogSum^m = difference (future − base year) in LogSum term for mode m,
V_i = utility of submode i (e.g., walk or drive access attributes) within subchoice n (i.e., automobile or transit),
S_i = base year observed share of using submode i (e.g., walk or drive access), and
DIFF V_i = difference (future − base year) in the utility (e.g., travel time) of submode i.

Model Specification and Coefficients

The RTA model includes

- Transit travel time and cost (i.e., in and out-of-vehicle times and transit fare) variables in the utilities of the transit submodes, walk and drive access; and
- Automobile travel time and cost (i.e., parking and automobile operating) variables in the utilities of the automobile submodes.

The cost variable is normalized with respect to zonal median income. This composite variable is constructed by dividing the automobile cost components (i.e., sum of automobile operating, parking, and ownership costs) and transit fares by the ratio of zonal median income over the base year regional median income. The PSRC mode choice model coefficients are used in the incremental mode choice models.

Base Mode Shares

Application of the incremental logit model requires a reasonable estimate of existing shares for each alternative mode. The census journey-to-work data provide an excellent source of automobile, carpool, and transit shares for commute trips. Even with those data, however, there are many zone-to-zone interchanges with no reported shares. Base mode shares, therefore, are computed by aggregating shares to 26 summary districts at the work ends only. The shares at home ends are calculated at a 219-FAZ (PSRC forecasting analysis zones) level.

For derivation of the base year park-and-ride shares, a procedure similar to that just mentioned is used to aggregate the shares. Specifically, base year park-and-ride shares are calculated at 26-district-to-219-FAZ interchanges using the transit on-board origin and destination data.

Surveys Conducted in 1985 and 1992

The 1985 survey conducted by Seattle Metro and the 1992 surveys conducted by the four transit operators (Metropolitan King County, Pierce County, Everett Transit, and Community Transit) provided a complete cross section of representative transit trips for two separate years. The 1985 survey was limited to only one county (King); the 1992 surveys covered the three-county RTA area shown in Figure 1 (see Table 1).

Transit operators also provided detailed ridership counts by route and time of day, which were the basis for expanding the surveys to 100 percent of the transit travel.

Time of Day and Trip Type Hierarchy

For the project planning studies, the RTA assumes that most questions to be addressed by the modeling effort would require tests of alternate transit service instead of alternate external environments. Variables affecting ridership are more related to time of day than to trip purpose for these questions. For example, both fares and service

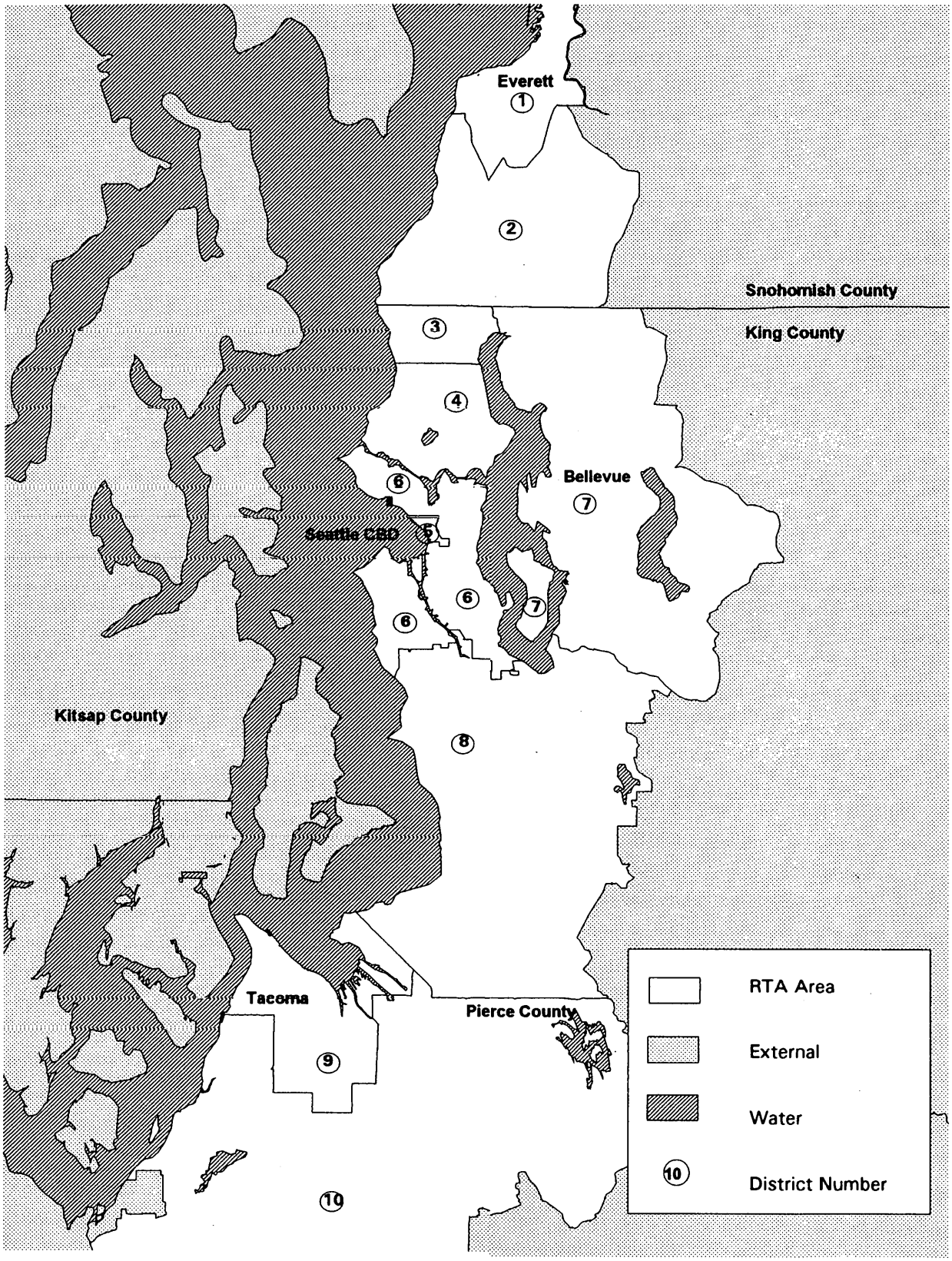


FIGURE 1 Ten districts within the RTA area.

TABLE 1 On-Board Transit Surveys

	King County	Seattle CBD	Pierce County	Pierce Seattle Express	Everett Transit	Community Transit
Month Conducted	May-92	Feb-93	Sep-92	Apr-92	Nov-92	Nov-92
Responsible Agency	Metro	Metro	Pierce Transit	Pierce Transit	Everett Transit	Community Transit
Approximate Response Rate	40%	20%	30%	60%	50%	50%
Percent of 3-County Ridership	75%	8%	8%	1%	2%	6%

vary by time of day, not by trip purpose. In fact, service variability by time of day is quite extreme in the Seattle area. The RTA model simulates afternoon peak and off-peak transit travel patterns.

Rider response to on-board survey questions on trip purpose is not as strictly controllable as are travel diary surveys or interview-based surveys. Therefore, the RTA model uses a simple categorization of trips, "commute" versus "noncommute."

RTA Staged Forecasting Analysis

Stage 1: Changes in Demographics

The RTA model uses PSRC trip tables to change surveyed transit demand from a base year to a forecast year. Because there are many mismatches due to the occurrence of zeros within any two trip tables, some aggregation is necessary to ensure reasonable application of cell-to-cell growth factors. The RTA model calculates factors at the level of 219-FAZs. The RTA modeling effort will retain the Fratar method as a backup to using regional trips from the PSRC trip distribution model. The calculation is

$$\text{Stg1Trn} = \text{SurvTrn} \times \frac{\text{Trips}^f}{\text{Trips}^b}$$

where

Stg1Trn = Stage 1 transit trip forecasts (737 × 737 zones),
SurvTrn = base year surveyed transit trips (737 × 737 zones),
Trips^f = forecast-year PSRC travel demand (219 × 219 FAZs), and
Trips^b = base-year PSRC travel demand (219 × 219 FAZs).

The results of the Stage 1 analysis are the transit trips for a future year assuming nothing changes but population and employment. Secondary impacts of growth on transit demand, such as increased highway congestion, are not accounted for in Stage 1.

Stage 2: Changes in Highway Congestion and Cost

Stage 2 considers influences on mode choice due to changes in highway congestion, automobile costs (including parking costs), transit fares, and income.

In all of the ridership analysis done in the Puget Sound region, transit fares have been held constant across alternative transit networks. Should that approach change, it would be advantageous to shift consideration of transit fares to Stage 3, where the fare policy could vary with each transit network.

PSRC is responsible for all regional highway modeling. RTA patronage forecasts use PSRC estimates of highway travel times. The times are tabulated in the form of 219 × 219 FAZ-to-FAZ matrices for each highway network. When a transit alternative significantly affects the highway system (e.g., taking of freeway lanes for transit facilities), additional PSRC future highway networks and congestion analysis are required.

Stage 2 transit trip forecasts are calculated using the following incremental logit equation:

$$\text{Stg2Trn} = \frac{\text{Stg1Trn}}{S_t + (1 - S_t) * [\exp(B * \text{DIFF LogSum}^a)]}$$

where

Stg2Trn = Stage 2 transit trip forecasts,
Stg1Trn = Stage 1 transit trip forecasts,
S_t = observed transit shares from census data for base year,
B = LogSum variable coefficient (equal to 1.0) for the automobile subchoice, and
DIFF LogSum^a = difference in the LogSum values due to changes in highway congestion and costs (future − base year) [census data (for the baseline share), highway skims and costs, and fares are used in Equation 1 to estimate DIFF LogSum^a representing drive alone and group ride submodes].

Stage 2 transit share forecasts (Stg2Shr) are calculated as follows:

$$\text{Stg2Shr} = \frac{\text{Stg2Trn} * S_t}{\text{Stg1Trn}}$$

Results of Stage 2 analysis are the transit trips for a future year, having accounted for factors external to the transit service itself. The results serve as a platform for analysis of ridership on alternative transit networks.

In most project planning ridership forecasting, Stages 1 and 2 need not be calculated as often as Stage 3. Only when a transit alternative is presumed to have a strong effect on land use or the regional highway network, for example, would the entire process have to be cycled through. Guidelines published by FTA (*17*) discourage such cycling in the evaluation of transit investments.

Stage 3: Changes in Transit Service

In the third and final stage of the forecasting analysis, incremental changes in the transit level of service are taken into consideration. The change is reflected in resulting relative values of the LogSum^t variable using the base year and future transit networks. Stage 3 transit ridership forecasts, Stg3Trn, are calculated as follows:

$$\text{Stg3Trn} = \frac{\text{Stg2Trn} * [\exp (B * \text{DIFF LogSum}^t)]}{\text{Stg2Shr} * [\exp (B * \text{DIFF LogSum}^t)] + (1 - \text{Stg2Shr})}$$

where

B = Logsum variable coefficient (equal to 1.0) for the transit subchoice, and

DIFF LogSum^t = difference in LogSum values due to changes in transit level of services (future − base year). Base year observed shares for park-and-ride and changes in transit level of services are used in Equation 1 to estimate DIFF LogSum^t representing walk- and automobile-access submodes.

RTA ridership analysis involves preparation of summary information on the three-stage incremental forecasting process for each alternative plan. Sample trip ends for p.m. (noon-to-midnight) origin districts (see Figure 1 for the district definition) are indicated in Table 2. Information presented in Table 2 facilitates separate examinations of the potential impacts of incremental change in each variable at each stage of the ridership forecasting analysis.

MODEL VALIDATION AND BACKCAST RESULTS

RTA model validation analyses were conducted for both the base year 1992 and the 1985 backcast.

Route-Level Validation Results

Figure 2 shows the model's replication of route-level boardings for 1992. The surveyed and expanded transit trips were assigned to the model network to validate boardings and transfer penalties. Figure 2 shows a regression of total boardings on 342 lines against the automated passenger counter (APC) boardings on these lines. The R^2 of 0.91 and standard deviation of 369 daily boardings indicate a remarkably close match.

No boarding counts by route are available for 1985 because the APC system was not operational at that time. Boardings for Pierce County and Snohomish County routes are from driver counts.

Observed versus Estimated 1985 Backcast Results

Table 3 compares observed and estimated 1985 transit trips. A comparative analysis is possible for trips within King County (excluding intra–central business distinct) because of the availability of observed 1985 transit trips from the King County 1985 transit survey. No comparison can be made for other counties because no survey was conducted in 1985 for those transit markets. Overall, the RTA

TABLE 2 Sample Build-Up/Down Analysis: 1992 to 1985 p.m. Daily Transit Trips by p.m. Origins

PM Origins	1992 Observed Trips	1985 Stage 1 Trips	1985 Stage 2 Trips	1985 Stage 3 Trips	Stage 3 % Change From 1992
1 Everett	4,060	3,160	3,260	3,880	-4.4%
2 SW Snohomish County	1,620	1,160	1,390	1,410	-13.0%
3 Shoreline	920	830	900	930	1.1%
4 North Seattle	21,370	20,120	18,880	18,470	-13.6%
5 Seattle CBD	53,270	45,010	46,620	46,700	-12.3%
6 South Seattle	25,470	25,000	26,880	27,490	7.9%
7 Eastside	4,840	3,350	3,640	3,770	-22.1%
8 South King County	6,560	5,620	6,270	6,360	-3.0%
9 Tacoma	7,570	7,000	7,880	7,680	1.5%
10 Pierce County	2,150	1,850	2,190	2,170	0.9%
Total (Noon-to-Midnight)	**127,830**	**113,100**	**117,910**	**118,860**	**-7.0%**
% Change Relative to 1992 Observed Trips	0.0%	-11.5%	-7.8%	-7.0%	
% Change Relative to Previous Step in Build-Up/Down Analysis	----	-11.5%	4.3%	0.8%	

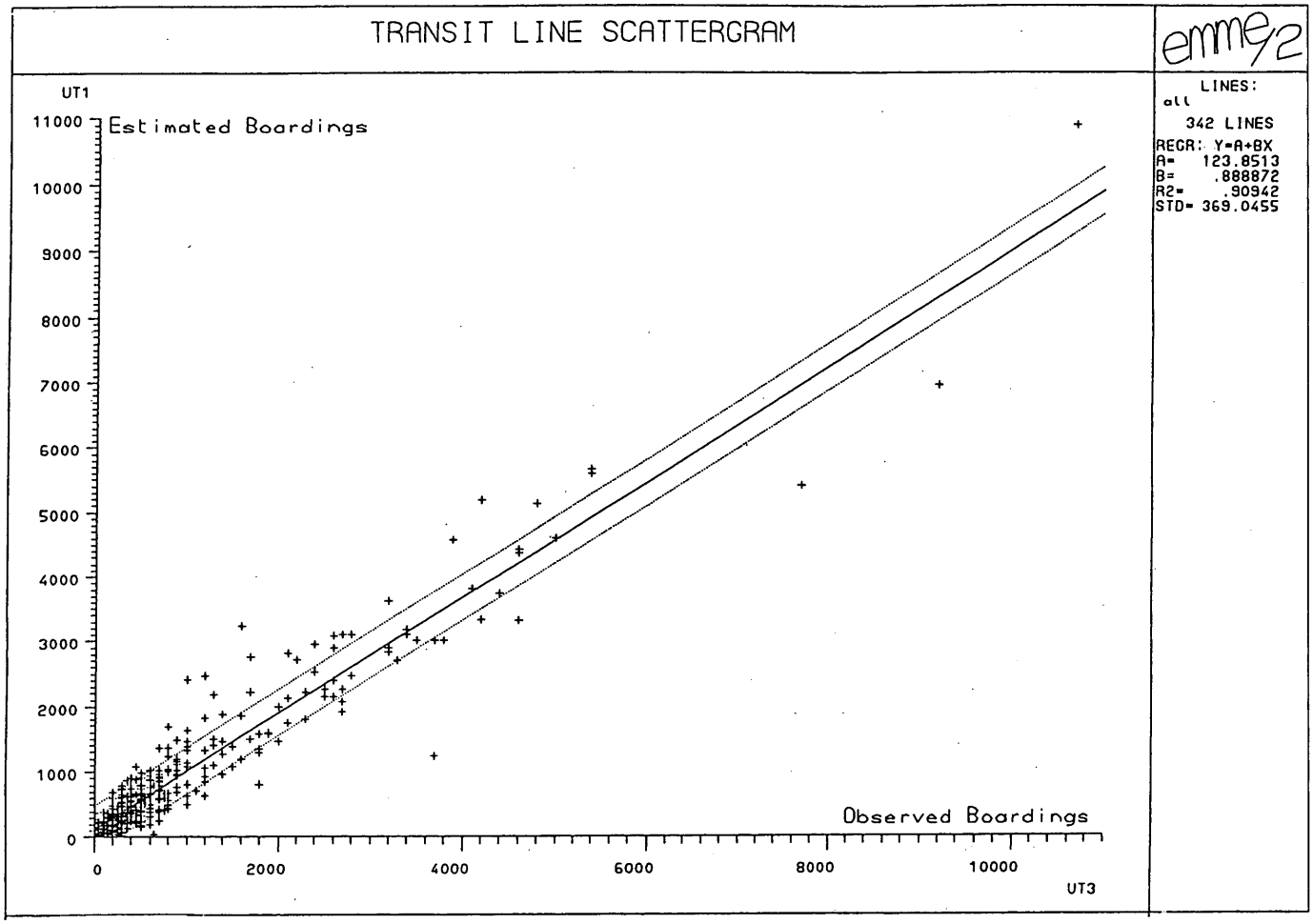


FIGURE 2 Daily transit line boarding comparison for 1992.

model has produced accurate 1985 backcasts for about three-fourths of the markets with over 1,500 daily transit trips (at least in King County). Those results are based on using the EMME/2 matrix balancing method to generate Stage 1 forecasts. Use of regional trip distribution estimates from the PSRC model resulted in worse 1985 backcasts for most markets (*12*). Comparative analyses of the EMME/2 matrix balancing (Fratar) method and trip distribution gravity model will be the subject of a future paper by the authors.

Results from Highlighted Changes in Transit Service, 1992 to 1985

In evaluating RTA model performance, one useful criterion is whether the model replicated ridership response to measurable changes in the transit systems between 1992 and 1985. The existing fully incremental RTA model has been responsive to measurable changes in transit systems between 1992 and the 1985 backcast year. There have been only a few changes in transit service between 1985 and 1992. Distinct and measurable changes in transit service between 1985 and 1992 include

- Change in park-and-ride express bus services from Snohomish County to Seattle,
- Change in park-and-ride express bus services from Pierce County to Seattle,
- Opening of the Downtown Seattle Transit Tunnel, and
- Introduction of the U-PASS Program in the University of Washington district.

The U-PASS program is a transit pass that makes transit virtually free for University of Washington students and some staff. These changes should have caused an increase in 1992 ridership relative to 1985 within these markets. The RTA model has responded correctly to the changes, as reflected in the resulting 1985 transit trip estimates (see Table 4). In summary, the RTA model has estimated

- 23 percent fewer p.m. peak automobile-access transit trips between downtown Seattle and Snohomish County in 1985;
- 36 percent fewer p.m. peak automobile-access transit trips between downtown Seattle and Pierce County in 1985; and
- 13 percent fewer intra–Seattle central business district, off-peak, noncommute trips in 1985.

Additional results pertaining to changes in intercounty park-and-ride express service during the 7-year interval are summarized in Table 5.

TABLE 3 Comparative Analysis of 1985 Estimated and Observed Daily Transit Trips by Origin and Destination (King County Districts Only)

Origin District		Destination District 1	2	3	4	5	6
1 Shoreline	Estimated	201					
	Observed	264					
2 North Seattle	Estimated	933	16,449				
	Observed	**1,052**	**14,125**				
3 Seattle CBD	Estimated	1,675	10,923	n/a			
	Observed	**1,669**	**11,667**	n/a			
4 South Seattle	Estimated	551	10,897	19,265	25,415		
	Observed	**733**	**9,472**	**21,632**	**24,554**		
5 Eastside	Estimated	194	1,759	4,469	2,070	2,511	
	Observed	**175**	**2,074**	**5,211**	**1,734**	**3,346**	
6 South King County	Estimated	95	1,153	4,980	3,352	772	6,028
	Observed	**164**	**1,239**	**5,223**	**3,500**	**753**	**6,869**

* Numbers on the observed rows represent 1985 transit trips from Metro King County on-board Survey.

TABLE 4 Model Performance in Response to Highlighted Transit Changes Between 1992 and 1985

Highlighted Change	Transit Trip Type	To	1992 Observed Trips	1985 Estimated Trips	1985 % Change From 1992	1985 Actual Change
Increase in Cross-County Express Bus Service to New Park-and-Ride	PM Peak Park-and-Ride	Snohomish County	2,500	1,900	-24%	-30%
	PM Peak Park-and-Ride	Pierce County	1,000	600	-40%	-50%
Opening of Downtown Transit Tunnel	Off-Peak Non Commute	Seattle CBD (Intra-Trips)	8,900	7,720	-13%	Not Available

TABLE 5 Model Performance in Response to Intercounty Park-and-Ride Express Service (Daily Transit Volumes)

Screenline Location	1992 Observed	1992 Estimated	1985 Observed	1985 Estimated
Pierce County Line:	2,600	2,700	1,000	1,000
Snohomish County Line:	8,700	9,000	5,800	5,600

CONCLUSIONS AND AREAS OF FUTURE RESEARCH

The incremental model presented in this paper is simple and can be easily applied to other projects for a more efficient and accurate transit ridership forecasting analysis. Implementation of the RTA fully incremental model became possible because of the availability of new surveys covering well-established markets of all transit riders, including park-and-ride users in the Seattle area. The integration of incremental subchoice models should be a noticeable improvement compared with traditional synthetic methods. Initial results from the validation analyses have clearly demonstrated responsiveness of the RTA model to changes in transit service, although limited, between 1985 and 1992.

The incremental RTA transit model is more efficient for transit planning analysis, because it

- Is simple and is directly based on observed travel, not estimated travel;
- Is an adjunct to the existing regional model and requires no model calibration;
- Has been responsive to highlighted changes in transit service from 1992 to 1985;
- Has reproduced observed travel patterns for park-and-ride transit users;
- Concentrates efforts on transit network analysis for studies whose primary questions are about alternative transit networks;
- Highlights error sources effectively whether in networks or in trip data; and
- Is a cost-effective and transparent staged forecasting process.

Incremental Model Limitations

The incremental model also has some limitations, because it

- Requires observed baseline travel pattern of transit trips;
- Is applicable only to areas with relatively good existing transit coverage;
- Would require a synthetic submodel for areas without well-established park-and-ride markets or transit in general;
- Requires availability of a regional model for nontransit input data and for interfaces with highway analysis;
- Is not well-suited for comprehensive analysis of major structural changes, such as land use visions involving feedback loops to transportation investments; and
- Requires good coordination with regional modeling staff and local traffic modeling staff for evaluation of transit improvement impacts on highway facilities.

Areas of Future Research

Presently, the incremental method is not useful for long-range multimodal corridor studies, comparing simultaneous transit and highway improvement strategies. Research should be directed toward developing methods such as the gradient approach suggested by Spiess (*18*) for estimation of base year origin to destination trip tables, possibly for all modes, from the existing actual counts of passengers and vehicular traffic. Such counts are usually collected by transit operators, cities, counties, and state transportation departments. Availability of base year trip tables for both transit and automobile modes will extend application of the incremental method not only to traffic forecasting but also to multimodal modeling analyses. Use of incremental models will not only simplify the existing travel demand modeling practices but also greatly enhance efficient generation of detailed project- and subarea-specific travel demand forecasts.

Currently, the incremental method depends on trip-based definition instead of activity-based definition. That limitation can be rectified by incorporating pertinent findings from new research into existing regional synthetic models before incremental methods are applied.

Finally, research also should be directed toward either limiting or eliminating conventional zone definition in transportation modeling and forecasting processes. Trips or activities should be geocoded to their actual surveyed household and destination locations rather than using traditional origin and destination zones. That concept will allow the transformation of modeling from a matrix-calculation environment to the calculation of incremental equations directly on the survey records, including use of trip-specific, as opposed to zone-specific, data for the level-of-service attributes. Limited experiments by the authors on a no-zone concept in incremental transit modeling have been encouraging but require additional research on representation of access-mode choices.

ACKNOWLEDGMENTS

The authors would like to thank James M. Ryan and Gordon W. Schultz of Parsons Brinckerhoff, Inc., for their contribution during the 1991 phase of mode choice model development for the Seattle Regional Transit Project. We would also like to thank the system forecasting team, specifically, David Phillip Beal of RTA and Cathy J. Strombom of Parsons Brinckerhoff, Inc., for their contribution and support during the second phase of mode choice, model development, and application. This work was funded by Washington State Department of Transportation, through its High-Capacity Transit Program, and by Metropolitan King County.

REFERENCES

1. Talvitie, A., M. Morris, and M. Anderson. Assessment of Land-Use and Socioeconomic Forecasts in the Baltimore Region. In *Transportation Research Record 775*, TRB; National Research Council, Washington, D.C., 1980, pp. 38–41.
2. Talvitie, A. Planning Model for Transportation Corridors. In *Transportation Research Record 673,* TRB, National Research Council, Washington, D. C., 1978, pp. 106–112.
3. Talvitie, A., and Y. Dehghani. Comparison and Evaluation of Case Study Alternatives for a Light-Rail System and Its Possible Land-Use Impacts in Buffalo, N.Y., Region. *Transportation Research Forum,* Vol. 2.2, Sept. 1985.
4. Brand, D., and J. L. Benham. Elasticity-Based Method for Forecasting Travel on Current Urban Transportation Alternatives. In *Transportation Research Record 895,* TRB, National Research Council, Washington, D.C., 1982, pp. 32–37.
5. Ben-Akiva, M., and T. Atherton. Methodology for Short Range Travel Demand Predictions. *Journal of Transport Economics and Policy,* Vol. 7, 1977.
6. Koppelman, F. S. Predicting Transit Ridership in Response to Transit Service Changes. *Journal of Transportation Engineering,* Vol. 109, No. 4, July 1983.
7. Nickesen, A., A. H. Meyburg, and M. A. Turnquist. Ridership Estimation for Short-Range Transit Planning. *Transportation Research B,* Vol. 17B, 1983.

8. *Travel Forecasting Methodology Report—Draft.* Parsons Brinckerhoff, Inc., Seattle, Wash., Oct. 1991.
9. *Patronage Forecast Methodology—Colma BART Station.* Parsons Brinckerhoff, Inc., and COMSIS Corporation, Feb. 1987.
10. *Service and Patronage Impact Assessment Methods Report—Central Light Rail Line Extensions.* Parsons Brinckerhoff, Inc., and COMSIS Corporation, Nov. 1989.
11. *Service and Patronage Forecasting Methodology—Honolulu Rapid Transit Development Project, Alternatives Analysis and Draft Environmental Impact Statement, Task 5.* Parsons Brinckerhoff, Inc., and COMSIS Corporation, Dec. 1989.
12. *Travel Forecasting Methodology Report—Final Draft.* Regional Transit Authority and Parsons Brinckerhoff, Inc., Seattle, Wash., Nov. 1993.
13. Ben-Akiva, M., and S. Lerman. *Discrete Choice Analysis—Theory and Application to Travel Demand.* MIT Press, Cambridge, Mass., 1985.
14. Dehghani, Y. *Prediction, Models and Data: An Analysis of Disaggregate Choice Models.* Ph.D. dissertation. State University of New York at Buffalo, Buffalo, April 1980.
15. Talvitie, A., Y. Dehghani, et al. *Refinement and Application of Individual Choice Models in Travel Demand Forecasting.* Technical Document, Vol. 1. State University of New York at Buffalo, May 1981.
16. McFadden, E., A. Talvitie, et al. *Demand Model Estimation and Validation.* Urban Travel Demand Forecasting Project Final Report, Vol. 5., University of California, Berkeley, 1977.
17. Procedures and Technical Methods for Transit Project Planning. FTA, 1992.
18. Spiess, H. *A Gradient Approach for the O-D Matrix Adjustment Problem.* Transportation Research Center, University of Montreal, Montreal, Canada, May 1990.

Opinions and views presented in this paper are solely the authors' and do not necessarily represent official policy of RTA or its member agencies.

Publication of this paper sponsored by Committee on Public Transportation Planning and Development.

Will Multimodal Planning Result in Multimodal Plans?

JAMES L. COVIL, RICHARD S. TAYLOR, AND MICHAEL C. SEXTON

As the multimodal planning and programming processes that are encouraged by the Intermodal Surface Transportation Efficiency Act of 1991 are developed, potential effects of open competition upon the mix of project types need to be recognized. Inherent differences between modes, as well as between different types of projects within a mode, mean that a comprehensive evaluation process will be necessary. Further, analytical processes alone cannot be relied on in weighing the relative merits of competing projects. Instead, judgments about the values attached to a variety of evaluation parameters will have to be made. The way that is done clearly will have profound effects on the mix of projects that survive the planning and programming process. To get the proper mix, some bias was introduced into what initially was intended to be an unbiased evaluation of project worthiness. That is, a high value was placed on the social, energy, and environmental qualities if certain candidate projects were to compete successfully against projects that had superior transportation mobility and cost-effectiveness characteristics. Consequently, for the foreseeable future, a combination of analytical processes and value judgments will be necessary in developing multimodal plans that encompass the full range of modes and project types.

Intermodal Surface Transportation Efficiency Act (ISTEA) requirements stipulating that every state implement a multimodal planning process are well known to transportation planners. ISTEA is, without question, the most profound transportation legislation that Congress has enacted since the legislation that produced the Interstate system. It is also most significant that the U.S. Department of Transportation chose to give considerable flexibility to each state and metropolitan planning organization concerning compliance with ISTEA. The federal rulemaking process is refreshing because it embodies the philosophy that "one size does not fit all" and acknowledges that processes adopted by individual jurisdictions should not be forced into a rigid, "no deviation allowed" format. The federal process is also challenging because each jurisdiction must select multimodal planning and programming processes that are effective and practical for it. The task is made difficult because our profession has limited experience with multimodal planning, whereby choices are made between a variety of transportation alternatives.

For many of us, ISTEA affords a most welcome and long-awaited environment in which to conduct multimodal planning and programming. However, there are some potential pitfalls that could be encountered unless we identify them now and take steps to avoid them.

PRE-ISTEA PLANNING PHILOSOPHY

It is critical to understand the past to plan for the future. In fact, it is the past that explains why the old ways of doing transportation planning and programming produced results that are not entirely satisfactory, and why we transportation planning professionals see opportunities in ISTEA.

Not very long ago, it was recognized that, in terms of ground transportation, highways dominated the transportation system. Instead of undertaking comprehensive multimodal planning, the nation chose to do primarily modal planning, whereby candidate projects compete with similar projects within a particular mode. Certain kinds of projects did not emerge as winners in the new planning and programming processes. For example, given open competition, most rail-highway grade crossing projects were not winning out against capacity-enhancement highway projects, nor were most bicycle projects or highway safety projects considered high priorities.

It became clear that if the United States were to have a transportation system that met a wide range of transportation needs, special provisions would be needed to recognize the value of each kind of transportation project in the planning and programming process. Value judgments were made that said a certain portion of funds would be used for different kinds of projects. The process led to more and more categorical programs. Finally, multimodal plans arose by structuring the fund allocation process to yield a variety of project types. That did not necessarily result in "balanced" or optimal multimodal plans; although plans contained certain elements for each mode, they were in reality a collection of modal plans.

ISTEA REQUIREMENTS AND PROGRAMMING CATEGORIES

ISTEA indicates clearly that any new plans are to be multimodal in nature, and much attention has been focused on the expanded funding flexibility provided by ISTEA. However, we still have what amounts to categorical programs. ISTEA provides for safety and transportation enhancement set-asides and maintains the separate bridge program. These features clearly inhibit the extent to which projects will be allowed to compete on their own merits with other types of projects.

Neither ISTEA nor the final rules prescribe how the multimodal planning process is to be structured other than through specifications for public involvement and the consideration of specific factors in evaluating projects. Nevertheless, there are strong indications that previous programming processes, in which suballocations of funds to different types of projects are made to ensure that some projects of each type are actually selected, will not be permitted.

EXAMPLES OF MULTIMODAL PLANNING

TRB has sponsored several efforts to identify good examples of multimodal planning. At a TRB conference in Seattle in 1993,

Wilbur Smith Associates, P. O. Box 92, Columbia, S.C. 29202-0092.

Michael Meyer reported he was able to find only two examples of "illustrations of close-as-you-get multimodal planning." One example was the I-15 Corridor Alternative Analysis in Salt Lake City (*1*). The cited process involved a project level analysis wherein more than 50 performance and impact measures were developed for 12 highway and transit alternatives.

The second study cited by Meyer was the Maryland Commuter Assistance Study, a study of 14 corridors to determine "how best to move people given the varied nature of commuter problems statewide" (*2*). Alternative improvements included express bus service, highway access control, roadway widening, shoulder bus lanes, exclusive bus roadways, high-occupancy vehicle (HOV) lanes, commuter rail, and light and heavy transit.

Although it was not noted in Meyer's paper, the process used by the Metropolitan Transportation Commission (San Francisco Bay Area) is probably one of the most noteworthy of the multimodal/intermodal trade-off analyses. Technical aspects of the process involve an initial screening of potential projects on the basis of selected criteria. Projects that pass the screening are processed using a scoring system that includes performance-based standards. Finally, projects are subjected to various "programming criteria to ensure that the program of projects increases mobility, cleans the air, leverages the most State and Federal resources, and is equitable" (*3*).

Another regional planning process that has been cited as state-of-the-practice in multimodal evaluations (unpublished report of NCHRP Project 20-5) is the Hali 2000 Study (*4*). It was conducted to update Oahu's Long-Range Transportation Plan; it reviewed all major travel corridors and addressed the full range of transportation alternatives, including transportation system management, HOV, bus, light rail, rapid transit, and highway improvements. The evaluation matrix contained a mixture of more than 60 quantitative and qualitative factors. Evaluation criteria focused on (*a*) cost-effectiveness, (*b*) community and institutional acceptance, and (*c*) measures of effectiveness related to transportation goals and objectives.

The cited examples suggest that in the foreseeable future multimodal processes will require a combination of analysis and judgment to produce multimodal transportation plans.

PUBLIC INVOLVEMENT

Whatever multimodal planning process eventually is adopted, ISTEA requires that it be a much more open process than some planning agencies may have undertaken in the past. We need to consider carefully how that might influence the content of the multimodal plans we are to prepare.

Ohio was awarded one of the six grants provided by FHWA for the development of a "model intermodal planning process." As consultants to Ohio Department of Transportation (ODOT), Wilbur Smith Associates participated in a two-part series of outreach meetings. What the company heard from the general public, public officials, and special interest groups was very revealing. For example:

> In rural areas, the dominant message was "highways, highways, highways". . . . In contrast, ODOT's metropolitan customers were far more interested in other modes—particularly public transportation and rail. (*5*)

A similar experience in Des Moines, Iowa, involved the primary question of how to move people and goods between the suburbs and downtown. Because there is a history of using highways to solve problems in the community, "pro-highway" people tended to take the outcome of the study for granted and not attend study meetings. However, "no growth" and transit proponents were less apathetic. Only by conducting home telephone surveys did we determine that support at public hearings was skewed toward certain modes.

Although 71 meetings were held in connection with the "Access Ohio" project, the overwhelming majority of input received dealt with the transportation of people. Although letters of invitation were sent to freight transportation interests, typically they either did not attend or, if they did, they did not assert their positions. Only by direct contact with the freight interests was significant input obtained. Their lack of participation appears related to the historical overemphasis on highways during previous planning efforts and their reluctance to discuss private business in public. It is true that "people vote, freight does not," yet planning processes should adequately consider freight because of its importance to the country's economy.

A multimodal planning process should recognize the considerable differences in messages received as part of a public involvement program and ensure that they are put into proper context.

CONCERNS THAT MUST BE FACED

Some believe that the current modally oriented approach to transportation planning is the preferred approach. They reason that each mode is so different that mixing them together is technically impossible.

Let's assume that we have determined we want to do truly multimodal planning. Further, let's assume that the multimodal planning process involves throwing all transportation projects into a common pot and requiring them to compete on an unbiased basis with other transportation projects. If we decide that multimodal planning is to be conducted in this manner, what are the potential challenges?

What Would Be the Balance Between Freight and Passenger-Oriented Projects?

There is the possibility that projects that are concerned primarily with the transportation of people will dominate the program of selected projects, because of the emphasis they receive in public outreach processes. Historically, transportation agencies have had little experience with freight transportation, and there often is an attitude that freight transportation concerns should be addressed by the private sector.

Evidence of this problem already has surfaced in Florida and California, where metropolitan planning organizations have shown preference for local roadway and signalization projects over port access projects or deemed port projects "not a proper use" of federal monies, even though ISTEA specifically mentions port access. Whereas government officials state in public that all projects are weighed equally, privately they admit that public pressure for local improvements means more at decision time. (*6*).

What Will Be the Balance Between Rural and Urban Projects?

Given the great differences in the intensity of use, there is a potential that, in taking a "common pot" approach, urban projects will

dominate. Many rural projects have been justified in the past primarily on the basis of providing minimum access to all parts of an area, and federal funding programs have been designed accordingly. ISTEA anticipated the problem and included provisions that guarantee funds to areas with a population of 5,000 or less, on the basis of previous secondary funding. Thus, regardless of the transportation benefits of the project, geographical allocations are required.

How Would Pedestrian and Bicycle Projects Fare?

Relatively low utilization of bicycle and pedestrian modes could result in such projects being at a disadvantage in the programming and prioritizing processes. This may be why ISTEA still has vestiges of the old categorical funding program, specifically requiring that plans include bicycle and pedestrian elements, including provisions for enhancement projects that encompass bicycle elements.

Would Safety Projects Be Able To Complete Well with Other Projects?

Certainly, safety projects have significant benefits; however, they never got much attention until categorical programs that focused attention on them were instituted. Apparently, those who wrote ISTEA thought that such projects might get less attention if a true multimodal approach were adopted. ISTEA includes requirements for a 10 percent set-aside of surface transportation programming funds for safety construction projects. Indeed, ISTEA gives even more attention to safety projects by requiring implementation of a safety management system.

What Would Be the Balance Between Transit and Highway Projects?

In terms of balancing transit and highway projects, it is less clear what a true multimodal process will yield. A major influence is the Clean Air Act Amendments of 1990. In addition, funding flexibility in ISTEA provides that approximately $103 billion of the $151 billion provided by ISTEA can be spent on transit. These provisions could shift the balance toward transit projects, particularly in the larger urban areas.

Some argue that the environmental justification for this shift is minimal. In testimony before the U.S. House of Representatives Public Works and Transportation Subcommittee on Surface Transportation, Pennsylvania Secretary of Transportation Howard Yerusalim commented on a Bay Area $11 billion investment: "Massive shifts in transportation investment from highways to transit . . . only works at the margins of the clean air problem." Further, he stated that many people

> promote Transportation Control Measures (TCM)—things like ridesharing programs, transit improvements, park-and-ride facilities, and bike/pedestrian programs—as an answer to air quality . . . in fact, there is evidence that the impact of traditional TCMs such as these is so small, that it is below the accuracy of our measuring ability. (*7*)

A similar experience in the Puget Sound area indicated that extensive use of TCMs and the construction of an $11.5 billion rail transit system could achieve only a 2 percent reduction in motor vehicle travel.

Would the Balance Between Bridges and Other Highway Elements Shift?

There are different kinds of projects within each mode. Under ISTEA, there are constraints on the amount of competition to which bridge projects can be exposed because the Bridge Replacement and Rehabilitation Program has been continued. That is, completely open competition will not occur under ISTEA. One wonders what would happen if bridge projects did not enjoy this special recognition.

What Would Be the Balance Between Port, Rail, and Aviation Projects Relative to Highway and Transit Projects?

ISTEA essentially addresses only highway and transit modes and some intermodal facilities. Public funding for port and rail projects often has been limited because such projects are considered commercial undertakings. Overall, there is less reluctance to fund airport projects with public funds, despite the commercial features that are apparent. If projects for these three modes were required to compete with highway and transit projects for the same funding, a change in the balance would be almost sure to occur, simply because we currently fund these projects in very different ways.

How Will the Balance Be Affected by Conflicting Goals?

Clean air concerns will drive alternative analyses toward more fuel-efficient modes or modal options. The urban transit-automobile relationship already has been mentioned. On the freight side, water and rail present viable alternatives from a clean air perspective, but not from a service-oriented shipping community viewpoint. Alternatively, longer combination highway vehicles could promote fuel efficiency, among other production enhancements, but they raise considerable safety concerns.

At the heart of this issue are two key public policy questions that are often in conflict. First, should government merely respond to market demands for certain transportation improvements, or should government force the public to alter existing travel preferences to create greater efficiencies? Second, should transportation planning be used to solve various social ills?

LESSONS LEARNED THUS FAR

Because this is a fairly new undertaking for most of us, the lessons we have learned have been a bit limited. Nevertheless, they are quite profound and include, at a minimum, the following key points:

- There are few working examples of successful multimodal planning processes.
- The old approach of fund allocations to each mode and to project-type category programs should not be used.
- The public involvement process cannot be relied on completely to reflect all of the value systems that should be embodied in a multimodal process.
- Most of us are relatively inexperienced with approaches that include private interests in the multimodal planning process.
- Because past planning and funding patterns were heavily slanted toward highway solutions, historical trends might be mean-

ingless in identifying future demands and solving transportation problems.

- If a balanced approach to multimodal planning is to be achieved, multimodal alternatives must be compared truly and an unbiased technical evaluation of each mode's potential contribution conducted.
- Technical analyses are only part of the answer. Value judgments are another crucial element.

REFERENCES

1. *Draft Environmental Impact Statement, I-15/State Street Corridor.* Report FHWA-UT-EIS-90-02-D. U.S. Department of Transportation, Wasatch Front Regional Council of Governments, and Utah Department of Transportation, 1990.
2. *Maryland Statewide Commuter Assistance Study, Summary Report.* Maryland Department of Transportation, Baltimore, 1990.
3. *Resolution No. 2457.* Metropolitan Transportation Commission, San Francisco, Calif., July 22, 1992.
4. *HALI 2000 Study Alternatives Analysis: Final Report.* Wilbur Smith Associates, Columbia, S.C., June 1984.
5. *Access Ohio: Reaching New Horizons in the 21st Century, the Outreach Program.* Ohio Department of Transportation, Columbus, Ohio, Feb. 1993.
6. A Year Later, Freight Projects Losing Out as Transit Rakes in Highway Bill Money. *Traffic World,* Feb. 15, 1993, pp. 12–18.
7. *ARTBA Newsletter.* American Road and Transportation Builders Association, Washington, D.C., June 22, 1993.

Publication of this paper sponsored by Committee on Statewide Multimodal Transportation Planning.