1220

TRANSPORTATION RESEARCH RECORD

Forecasting

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1989

Transportation Research Record 1220
Price: $12.00

modes
1 highway transportation
2 public transit

subject areas
12 planning
13 forecasting

TRB Publications Staff

Director of Publications: Nancy A. Ackerman
Senior Editor: Edythe T. Crump
Associate Editors: Naomi C. Kassabian
Ruth S. Pitt
Alison G. Tobias
Production Editor: Kieran P. O'Leary
Graphics Coordinator: Karen L. White
Office Manager: Phyllis D. Barber
Production Assistant: Betty L. Hawkins

Printed in the United States of America

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

Forecasting.
p. cm.—(Transportation research record, ISSN 0361-1981 ; 1220)
Papers from the 68th annual meeting of the Transportation Research Board.
ISBN 0-309-04814-1
1. Traffic assignments—Mathematical models—Congresses.
2. Traffic estimation—Mathematical models—Congresses.
I. National Research Council (U.S.). Transportation Research Board.
Meeting (68th : 1989 : Washington, D.C.) II. Series.
TE7.H5 no. 1220
388 s—dc20
[388.3'14'0112] 89-14408
CIP

Sponsorship of Transportation Research Record 1220

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

Transportation Data, Economics, and Forecasting Section
Chairman: Edward Weiner, U. S. Department of Transportation

Committee on Passenger Travel Demand Forecasting
Chairman: James M. Ryan, Comsis Corporation
Moshe E. Ben-Akiva, Daniel Brand, Jeffrey M. Bruggeman, Robert M. Donnelly, Christopher R. Fleet, David A. Hensher, Joel L. Horowitz, Ron Jensen-Fisher, Peter M. Jones, Frank S. Koppelman, David L. Kurth, T. Keith Lawton, David M. Levinsohn, Hani S. Mahmassani, Fred L. Mannering, Eric J. Miller, Michael Morris, Eric Ivan Pas, Joseph N. Prashker, Phillip R. Reeder, Aad Ruhl, Earl R. Ruiter, G. Scott Rutherford, Gordon W. Schultz, Peter R. Stopher, Kenneth E. Train

James A. Scott, Transportation Research Board staff

The organizational units, officers, and members are as of December 31, 1988.

NOTICE: The Transportation Research Board does not endorse products or manufacturers. Trade and manufacturers' names appear in this Record because they are considered essential to its object.

Transportation Research Board publications are available by ordering directly from TRB. They may also be obtained on a regular basis through organizational or individual affiliation with TRB; affiliates or library subscribers are eligible for substantial discounts. For further information, write to the Transportation Research Board, National Research Council, 2101 Constitution Avenue, N.W., Washington, D.C. 20418.

Transportation Research Record 1220

Contents

Foreword

This Record contains a research paper on highway traffic assignment techniques based on an improved behavioral model of drivers' route choice and a paper on dynamic assignment in three-dimensional time space. Additional subjects treated in this Record are (a) modeling demand diversion to less congested routes within a corridor, (b) the adequacy of conventional models in trip generation analysis, (c) assessment of the convergence properties of four traffic assignment algorithms, (d) network evaluation, (e) trip generation rate analysis for a state, and (f) the updating of a regional travel forecast.

The paper on techniques of traffic assignment, by Antonisse et al., presents models that are based on evidence of drivers' varying valuations of a number of characteristics of roads, are probabilistic, and are based on more variables than used in previous models. The models contrast with conventional traffic assignment procedures, which are typically based on a single measure of travel impedance. Hamerslag's paper on dynamic assignment assesses traditional assignment models, where cars are assigned to a route and are present on all links of that route simultaneously. A three-dimensional model is described, and the effect of improved capacity on congestion downstream is shown.

Demand diversion to less congested routes within a corridor is of increasing importance as a result of increasing traffic volumes, congestion, and delays. The paper on demand diversion, by Stephanedes et al., develops and describes two models for diversion at the trip origin and at freeway entrance ramps. The various factors contributing to diversion are also described.

In the areas of trip generation and traffic assignment, Monzon et al. assess the adequacy of conventional linear regression models in trip generation analysis. Simulation experiments designed to examine whether model coefficients can be accurately estimated by least-squares estimation when the dependent variable is a nonnegative integer are described. Horowitz examines the convergence properties of four popular traffic-assignment algorithms. The paper evaluates the algorithms according to (a) errors associated with insufficient iterations, (b) arbitrary selection of a starting point, (c) inexact theory, and (d) small variations in data. A third paper in the traffic assignment area, by Barbour and Fricker, investigates how the node-balancing solution for a network is affected by choice of method, which by implication also means choice of criterion. The paper further discusses the two categories of techniques developed to balance the network: algorithms and mathematical programming formulations. A comparison was also made between these procedures and the maximum-likelihood method.

Interregional stability of household trip generation rates for the state of New Jersey is reported and summarized in the paper by Walker and Olanipekun. Stratification schemes are tabulated and analyzed to determine the most appropriate basis for making disaggregate trip rate comparisons between regions.

The paper by Kollo and Purvis describes the process of updating a regional travel-forecasting model for the San Francisco Bay Area in terms of providing a data base for model estimation and validation. The 1981 Bay Area travel survey and the 1980 Census Urban Transportation Planning Package were used.

Highway Assignment Method Based on Behavioral Models of Car Drivers' Route Choice

ROBERT W. ANTONISSE, ANDREW J. DALY, AND MOSHE BEN-AKIVA

This paper proposes a highway traffic assignment technique based on an improved behavioral model of drivers' route choice as developed in a recently completed study in the Netherlands. Route choice models are developed from data collected in three corridors in the Netherlands. The models presented here, which are based on evidence of drivers' varying valuations of a number of road characteristics, are (a) probabilistic and (b) based on more variables than were used in previous models. They contrast with the underlying route choice models of conventional traffic assignment procedures, which are typically based on a single measure of travel impedance (e.g., travel time, generalized travel cost). A key feature of the models developed in the present study is that they are based on data describing the actual routes chosen by individual drivers. The paper describes how these models are used to generalize assignment methods through the exploitation of a multiclass-user technique. In an uncongested network, several routes typically will be predicted to be used between a given origin and destination; as congestion increases, so will the diversity of routes used. Several models appropriate for use in varying circumstances of data availability are presented and compared. Model inputs (e.g., road attribute data) are described, and practical implications of the underlying structural assumptions are discussed. Spatial transferability of the models is appraised on the basis of the differing results obtained for the three corridors studied. Finally, advantages and limitations of application of the proposed assignment method compared with conventional procedures are discussed.

A central element of the traffic assignment procedure is a model of the traveler's decision about which route to take given the origin, destination, and mode of travel of a trip. The problem of route choice for a traveler might be stated as follows: Given the other characteristics of the trip to be made—purpose, time, origin, destination, and mode, for instance—choose the "best" route through the transportation network in terms of some criterion. This best route is most often thought of as the one that minimizes travel disutility. Existing traffic assignment models often assume single measures of travel disutility such as travel time or distance, or some simple formula of generalized travel cost.

In reality, the problem of route choice faced by an automobile driver is very complex because of

1. The large number of possible alternative routes through even modestly sized road networks, and
2. The complex patterns of overlap between the various route alternatives.

Realistic replication of the human decision process in route choice—which synthesizes many factors about the trip and the various possible routes in making a choice—with a mathematical model is difficult at present because of the limited understanding of the route choice phenomenon, as well as limited techniques and computational resources.

The primary interest of studying car drivers' route choice is in improving traffic assignment procedures. In particular, accurate predictions of the usage of proposed new infrastructure are essential to the evaluation of the need for that infrastructure. The results of the route choice study suggest that current methods may underestimate the traffic attracted to major new roads. Secondarily, understanding route choice is valuable in attempting to redirect traffic streams so as to make the best possible use of existing roads. The current study is one of very few directed to a better fundamental understanding of this important aspect of behavior and the implementation of that understanding in practical planning methods. The overall objectives were twofold:

1. To improve current understanding of drivers' route choice preferences, and
2. To develop a practical traffic assignment model that reflects this choice process with greater sophistication.

FACTORS AFFECTING ROUTE CHOICE BEHAVIOR

A major task completed during the first phase of this project was an extensive literature review of factors affecting drivers' route choice preferences. Ben-Akiva et al. (*1*) synthesized the results of this review as a set of hypotheses that may be broken down into the following three categories:

1. Drivers' knowledge about alternative routes: Several authors hypothesize that drivers plan their trips in a hierar-

R. W. Antonisse, Operations Directorate, Massachusetts Bay Transportation Authority, Room 4730, Ten Park Plaza, Boston, Mass. 02116. A. J. Daly, Hague Consulting Group, b.v., Surinamestraat 4, 2585 GJ The Hague, Netherlands. M. Ben-Akiva, Department of Civil Engineering, Massachusetts Institute of Technology, Room 1-181, 77 Massachusetts Avenue, Cambridge, Mass. 02139.

chical fashion, building up from lowest-level (local) roads near the origin of the trip to expressways at the highest level, which they use for the bulk of travel, and back to local streets at the end of the trip (*2–4*). Knowledge may be lacking of local road alternatives to expressway portions of trips (*5–7*). Drivers often are unable to evaluate simple characteristics of paths and thus are unable to find the quickest or shortest route (*2, 8, 9*).

2. Decision processes: Various hypotheses assert that drivers either plot out their entire route before departure or make decisions at road junctions as they encounter them independently from previous decisions (that is, they follow a Markov process), or else they use some combination of these two approaches (*10*).
3. Route attributes and preferences: Specific attributes of routes to which drivers are attracted include travel time (*11–14*), distance (*14*), number of traffic signals (*5*), scenery (especially for nonobligatory trips, such as social or recreational ones) (*6,15*), time or distance on limited-access highways (*15*), safety (*11,15*), commercial development, congestion (*15,16*), road quality, and road signing (*17*).

Most of these hypotheses are not reflected in existing traffic assignment models.

NEW MODEL OF ROUTE CHOICE BEHAVIOR

The earlier work on this project documented by Ben-Akiva et al. (*1*) also included the conceptualization of a two-step model of route choice that (a) narrows down the large number of possible route alternatives to a choice set of a few alternatives and (b) chooses a route from this choice set based on the characteristics of the trip, driver, and attributes of the available alternatives. Survey data were collected for a sample of drivers observed to travel between the cities of Utrecht and Amersfoort, including information on the driver and on the trip itself (including the route actually chosen on the survey day). A network model of the corridor was used to generate sets of alternative routes for the sampled drivers, and a large number of route choice models was tested.

The empirical evidence of the first phase of this study showed that factors other than time and distance play a significant role in interurban route choice. For example, several road attributes that one normally associates with major highways—large capacity, restricted access, high hierarchical level, and high speed limit—were found to positively attract route choice. Traffic signals, on the other hand, were found to have a negative effect.

The estimation results demonstrated the feasibility of the two-stage approach to modeling route choice and produced a model that reflects the hypothesized structure underlying route choice behavior. Finally, a number of market segmentation tests demonstrated that trip purpose, frequency, and length can have important influences on route choice.

OBJECTIVES

The results of the second phase of this project are presented. They are based on a new data collection effort that began in 1980 in two other road corridors in the Netherlands. The primary objectives of the second phase were

1. To test the transferability of both stages of the modeling process as developed in the first phase (the method used to generate the set of alternative routes and the choice model) to the other corridors, and
2. To simplify the choice model as a way to enhance the applicability of the model in a wider geographical area and under conditions of limited road network data.

Drivers were surveyed in two different corridors in the Netherlands in the spring of 1980—one between Amsterdam and Purmerend and the other between Arnhem and Apeldoorn. All the corridors offer a number of viable route alternatives for the many trips between the two cities defining each study area. In each case, a cordon of roadside survey points was laid out across the corridor. At some survey points, return-mail questionnaires were handed out to drivers, whereas at other points license plate numbers were recorded and registered owners of the vehicles were sent a return-mail survey form at home. The surveys asked respondents to trace the route they took on a map provided for the day of the sighting. The questionnaires also asked a range of questions about trip and personal characteristics: purpose at origin and destination, frequency of this trip, age, profession, and so on. Meanwhile, network data were collected from engineering sources.

ROUTE CHOICE MODEL FOR TRAFFIC ASSIGNMENT

This section describes the basic methodological requirements and data and computer needs for forecasting route choice behavior using the new approach.

Methodology

Travel behavior in general and route choice behavior in particular can be considered as choosing between discrete, mutually exclusive alternatives. Discrete choice analysis attaches expressions of attractiveness or utility to each of the available choice options. The utility expression of each alternative generally incorporates information on the attributes that may either add to or detract from its attractiveness. It is then assumed that the decision maker will choose the alternative that is most attractive.

With the primary problem in this case being highway route choice, the two major steps in determining behavior are

1. Identifying a set of route alternatives that the driver can choose among, and
2. Making the choice from this set on the basis of the type of driver and trip conditions and the various attributes of the route alternatives.

Because it would be prohibitively time-consuming and behaviorally unrealistic to evaluate the attractiveness of all possible routes between the origin and destination, a method is applied to narrow down the vast number of route possibilities to a few alternatives that may be considered in greater detail.

Once a set of options has been identified, it is necessary to measure the relevant attributes of those options that affect their attractiveness. A choice model is used to relate the probability of choosing each available alternative to its attractiveness, which, in turn, is based on the attributes of the alternatives. For the predictive tool to be successful in forecasting travel behavior under a wide range of circumstances, the choice model must be responsive to how changing travel conditions and varying perceptions affect the relative attractiveness of the available travel options.

Generation of Route Alternatives

As discussed above, the first stage of the route choice modeling process involves the generation of a set of candidate route alternatives from the myriad of feasible paths through the road network. The technique developed in this study is called the "labeling" approach because descriptive labels are attached to the selected route candidates. Each of these labeled routes is optimal with respect to some criterion from among all possible routes between the given origin-destination pair. For example, "quickest," "shortest," and "most scenic" might be criteria used to define three candidates from all route possibilities. The criteria to be used may be extracted from hypotheses regarding influences on route choice behavior and could be considered to constitute a model of drivers' perceptions of a road network.

So that these labels can help determine specific paths through the given network, a quantitative descriptor based on available network data must be selected to measure a route in terms of the label criterion. Labeled paths are defined by an impedance function that depends on one or more link attributes. A separate function is specified for each label criterion to be used. Determining the labeled path for a particular criterion is then simply a matter of calculating the associated impedance for all links in the network and executing a minimum-path algorithm that can efficiently generate labeled paths for a large set of origin-destination pairs. Observed chosen route data are required in selecting the most reasonable set of labels to apply in forecasting route choice. The selected set of impedance functions maximizes the frequency of observed routes included in the set of the corresponding labeled paths.

At this point, it is useful to describe the network data available for this study. Two types of data were used in this analysis: a basic network data base system and sets of extra, detailed link attributes. The Dutch Ministry of Transport maintains a computerized "Basisnetwerk" system consisting of many node and link records that represent the national highway network. This system is used extensively in the Ministry's planning and management functions. Node records include the junction's geographic location. Link records include A- and B-nodes, distance, speed code, and road hierarchy level as attributes. These basic attributes supply sufficient information for the generation of a few important labels.

A large number of detailed road link attributes was gathered for the detailed study area within each of the data collection corridors. Example attributes include road surface type, width of roadway, number of lanes in each direction, zoning type of adjacent land, and presence of various types of facilities along the roadside. A large number of alternative labeled paths could then be generated for any driver traveling either entirely or partially through the detailed study area.

In the first phase of the project, 10 labels were selected for application. These same labels were designated for application to the two new study areas in the second phase. The labels chosen and associated quantitative descriptors are described briefly as follows:

- Minimize time: travel time is calculated from information on the distance and average speed of the link.
- Minimize distance: the distance from the link records is applied.
- Maximize travel on scenic roads: the measure of impedance for the route is time spent driving on roads adjacent to nonscenic land uses—city center, dense residential, or industrial—as determined from percentage of link distance through these types of land use, which is available from detailed attributes.
- Minimize number of traffic signals: for each link, the number of traffic signals was calculated using detailed attribute information and the following formula:

$$\text{No. signals} = \text{no. signals along link} + 0.5\ (\text{total signals at nodes})$$

- Minimize travel on congested roads: detailed attributes allowed calculation of volume:capacity (V/C) ratios for road links in the Phase 1 study area. The descriptor is time spent on roads with high V/C ratios. Unfortunately, link volume data were not available for either the Arnhem-Apeldoorn or the Amsterdam-Purmerend study area, and this label had to be dropped from the analysis of joint data.
- Maximize use of expressways: links were classified as expressways if the network speed code was the maximum, that is, 100 km/hr (approximately 60 mph). Time spent on nonexpressway roads was used as a measure of link impedance in this case.
- Maximize travel on high-capacity roads: the impedance measure is time spent on low-capacity roads, that is, roads that either are less than 9 m (approximately 30 ft) wide or have less than two lanes in either direction.
- Maximize travel in commercial areas: again using land use data from the detailed attributes, time spent in noncommercial areas was calculated on the basis of the distance traveled in any land use area other than cities or industrial areas.
- Maximize road quality: for every link in the study area, a road quality rating is available on a scale of 1 (best quality) to 3 (worst quality). Time spent on poor-quality roads—those with a rating of 2 or 3—was measured.
- Hierarchical travel: each link includes an attribute for road hierarchy level. This label favors travel on the highest-level roads—generally limited-access highways. Two impedance measures were used: (a) time spent on roads of the lowest hierarchy (local roads) and (b) time on roads of moderate hierarchy (main roads of regional importance).

With the label descriptors determined, the next step is the specification of the impedance functions. In the case of the "minimize time" and "minimize distance" labels, this is simply the measure itself. For the other labels, however, it was possible for the optimal route of the criterion to deviate unreasonably from the minimum time path. To mitigate this prob-

TABLE 1 LABEL IMPEDANCE FUNCTIONS AND FINAL COEFFICIENT VALUES

Label Criterion	Link Impedance Function (to be minimized)	Initial Coefficients	
		Utrecht-Amersfoort	Amst-Purm & Arnhem-Apel
Min. Time	TIME		
Min. Distance	DISTANCE		
Max. Scenic	TIME + β_1(NON-SCENIC TIME)	2.0	2.0
Min. Signals	TIME + β_2(# SIGNALS)	30 sec.	5 min.[1]
Min. Congestion	TIME + β_3(HIGH V/C TIME)	3.0	not used
Max. Expressways	TIME + β_4(NON-EXP. TIME)	3.0	3.0
Max. Capacity	TIME + β_5(LOW-CAP. TIME)	1.5	2.0
Max. Commercial	TIME + β_6(LOW-COMM. TIME)	1.5	1.5
Max. Road Quality	TIME + β_7(LOW-QUAL. TIME)	2.0	2.0
Hierarchical Travel	TIME + β_{81}(HIER. 1 TIME) + β_{82}(HIER. 2 TIME)	5.0, 100	5.0, 100

[1] The value of 5 min. was not tested for Utrecht-Amersfoort. However increased values (above 30 sec.) did not show a large loss of coverage on that data.

lem, the impedance functions were specified as a weighted sum of the primary criterion measure (e.g., scenic time) and total travel time. The only remaining task is the assignment of relative weights to the two component measures. Table 1 shows the impedance functions, specified with weighting coefficients, as applied in this study.

The label parameters were optimized by finding the set of parameter values that maximized the number of observed chosen routes for the area under study that are matched or "covered" by the label set. A straightforward computer algorithm is used to compare each chosen route with the set of labels developed for that origin-destination pair and determine the existence of a match. Labels were introduced into the label set one at a time, optimizing the label's coefficient(s). A final sensitivity analysis ensured that changing any one parameter did not reduce the total matching score of the set of labels. The final Utrecht-Amersfoort label coefficients are shown in Table 1. A similar analysis was carried out in the other two study areas, using the Utrecht-Amersfoort coefficients as initial values. These values are also shown in Table 1, indicating the minimal changes between the areas in this respect.

Conditional Route Choice Models

After the decision maker's set of alternative routes has been generated, the next step is specification of the utility functions for each of the alternatives. Utility functions include a systematic component—an expression of how independent variables affect the attractiveness of the alternatives scaled up by their respective estimatable parameters—and a random term that accounts for the variability in choice behavior independent of the options' attributes. The systematic utility expressions are usually specified as linear combinations of the independent variables.

The "maximum likelihood" method is used to estimate the values of the utility function parameters. Simply put, this method determines the values of all parameters for which the observed choices are most likely to have occurred. The two most commonly applied probabilistic discrete choice models, logit and probit, differ in their assumptions about how the random term of the utility function is distributed [see, for example, Ben-Akiva and Lerman (*18*)].

The different mathematical properties of the random variables mean that each method has its advantages and disadvantages. Logit-form models are generally more flexible in the feasible number and structure of alternatives in the choice problem, and their parameters can be estimated with considerably less computational burden than those of probit-form models. Logit-form estimation programs are widely available. The package used in the route choice study is ALOGIT (*19*).

Two types of independent variables are considered for inclusion in the utility functions: (a) "level-of-service" attributes for each route including, for example, measures of travel time, distance, and travel cost; and (b) dummy variables that take a value of either 1 or 0 depending on whether an alternative meets certain conditions. In this analysis, dummy variables are used to indicate whether the route corresponds exactly to one or more of the labels considered.

Although these variables are objective measures of the routes themselves, different drivers perceive these attributes differently. The most common bases for these differences in perception may be the characteristics of the drivers themselves (e.g., age or profession) and characteristics of the trip being made (e.g., its purpose and the frequency with which it is made). An attempt is made to capture these differences in perceptions through estimation of models for various segments in the population and examination of the variance in the respective model parameter estimates.

Traffic Assignment

The methods outlined in the preceding sections can be applied by an adaptation of a "multiclass-user" (MCU) procedure. In a standard MCU method, classes are defined a priori as using paths that are minimal with respect to a class-specific impedance function. In the models described in this paper, the assignment procedures define the classes as users of each of the labeled routes. Because the usage of these routes is not known a priori and is dependent on the features of the routes, additional steps have to be introduced into the assignment procedure to apply the model. The procedure advocated is outlined in Figure 1.

An important feature of the procedure outlined is the integration of the new choice modeling approach developed in this study with the "capacity-restraint" methods that have been the subject of many previous studies. This integration means that previously developed algorithms, techniques, and so on, can be retained and current methods can be seen as independent improvement that loses none of the previous gains.

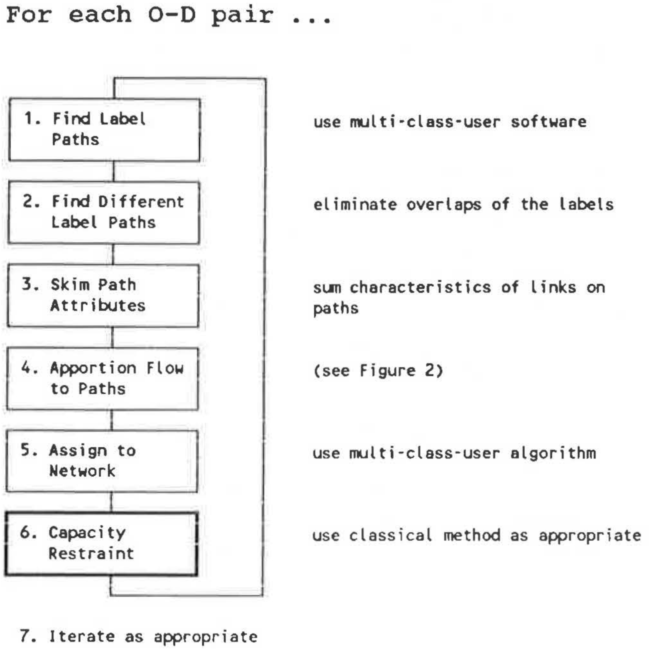
For each O-D pair ...
1. Find Label Paths
use multi-class-user software
2. Find Different Label Paths
eliminate overlaps of the labels
3. Skim Path Attributes
sum characteristics of links on paths
4. Apportion Flow to Paths
(see Figure 2)
5. Assign to Network
use multi-class-user algorithm
6. Capacity Restraint
use classical method as appropriate
7. Iterate as appropriate

FIGURE 1 Assignment procedure (overview).

The procedure involves six steps for each origin-destination pair for which a positive traffic flow is predicted.

Note that the first, third, fifth, and sixth steps, which are the most demanding in terms of computer processing, are standard MCU assignment steps and are already provided in standard packages. The second step is a simple programming task.

The fourth step in the process shown in Figure 1 is novel and is illustrated in greater detail in Figure 2. For each origin-destination pair, an apportionment is made by the model to each of the labels.

Figure 2 provides for a matrix of size (labels * segments) to be calculated for each origin-destination pair. It may be helpful to note that this procedure would be equivalent to a simple MCU procedure if labels and segments were identified, that is, if the matrix was simply 1.0 on the diagonal and zero elsewhere. The computation necessary to calculate and apply the matrix is not excessive.

In summary, an assignment procedure is proposed that requires comparatively minor extensions to existing software. Execution of this procedure requires little more computer time than a standard MCU method. The procedure is organized as a generalization of existing capacity restraint procedures, thus offering an advance without eliminating the possibilities resulting from previous studies.

MODELING RESULTS

This section summarizes the major quantitative findings from this project. The first subsection discusses results from the choice set generation using the labeling approach described above. The following subsections report and evaluate the final choice models that consider sets of six or fewer route alternatives.

Label Set Coverage of Observed Chosen Routes

With the primary objective of development of an assignment tool that can be applied in all three areas studied in this project, the parameters of the full set of nine labels are developed by maximizing matches of the chosen routes in all three study areas. More manageable six-label and four-label sets were developed for use in the choice modeling stage of the analysis. The labels included in these reduced sets were selected in part on the basis of the expected availability of the link attribute data required for generation of the label.

A computer network analysis package, SATURN (*20*), was used to build the labels between all chosen origin-destination pairs in the three study area networks. A separate computer program was written to compare each observed chosen route with the set of corresponding labeled routes and to summarize the match results. The label parameters developed in the first phase of this project (see Table 1) were used as initial values. Parameters were adjusted one by one, keeping the others fixed, in the direction that increased the number of matches to chosen routes.

Table 2 shows the match results for the initial full-label set and the six-label set using both initial and final label coefficient values. Each row in the main body of the table refers to one label and reports the coefficient value(s) and the set

For a given O-D pair ...

$$\begin{array}{l|ccccccc|l}
\text{Volumes:} & N_1 & N_2 & \cdots & N_i & \cdots & N_s & & \\
 & \text{Seg.1} & \text{Seg.2} & \cdots & \text{Seg.i} & \cdots & \text{Seg.s} & & \text{Flow on paths:} \\
\hline
\text{Path 1} & p_{11} & p_{21} & \cdots & p_{i1} & \cdots & p_{s1} & & V_1 = \Sigma^s{}_{i=1} \; N_i \cdot p_{i1} \\
\text{Path 2} & p_{12} & p_{22} & \cdots & p_{i2} & \cdots & p_{s2} & & V_2 = \Sigma^s{}_{i=1} \; N_i \cdot p_{i2} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
\text{Path j} & p_{1j} & p_{2j} & \cdots & p_{ij} & \cdots & p_{sj} & & V_j = \Sigma^s{}_{i=1} \; N_i \cdot p_{ij} \\
\vdots & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
\text{Path r} & p_{1r} & p_{2r} & \cdots & p_{ir} & \cdots & p_{sr} & & V_r = \Sigma^s{}_{i=1} \; N_i \cdot p_{ir} \\
\hline
\end{array}$$

Where: N_i is the number of vehicles for this O-D for segment i (input data);

p_{ij} is the probability of choosing path j for segment i (derived from model and path attributes);

V_j is the predicted volume for path j for this O-D (output to assignment stage).

FIGURE 2 Path apportionment.

of match scores for that single label in the various study areas—Utrecht-Amersfoort (full-label set only), Arnhem-Apeldoorn, and Amsterdam-Purmerend. Three types of match scores are reported for each label in the table:

1. Absolute matches: the total number of chosen paths matched by this individual label for corresponding origin-destination pairs.
2. Incremental matches: the percentage of chosen routes not matched by previous labels but matched by this label for corresponding origin-destination pairs.
3. Marginal matches: the percentage of chosen routes matched by this label and not matched by any other label in the table.

It is clear from Table 2 that a significant gain in coverage of the observed route choices can be realized by including additional criteria besides "minimize time." A comparison of the coverage in percentage terms of the time label alone versus the final six-label and full-label sets yields the results for the three study areas shown in Table 3, in which a four-label set—comprising time, distance, signals, and hierarchy labels—is also presented. Table 3 also shows the decreasing "rate of return" from increasing the size of the label set. Note that the apparent lower coverage of the full-label set relative to the six-label set for the Amsterdam-Purmerend area is explained by the use of the initial, nonoptimal set of coefficient values.

A sensitivity analysis of the label coefficients near their initial values showed that the matching rates generally remained stable. Nevertheless, some gains were made possible for the Arnhem-Apeldoorn and Amsterdam-Purmerend study areas by adjusting the parameters for the "minimize signals" and "maximize capacity" labels. These adjustments are reflected for the six-label set in Tables 2 and 3 and account for the apparent decrease in coverage shown for Amsterdam-Purmerend in the latter table when progressing from the six- to the nine-label set. The match score results in Table 2 as well as data availability considerations were used to decide which labels were to be eliminated to form the reduced sets.

The analysis found that many aspects of the labeling methodology were transferable between the three areas studied. The values of the label parameters, when optimized on chosen routes for the three study areas, also agreed very closely for most of the labels, as can be seen from Table 1.

Choice Modeling Results

Extensive discrete choice modeling was conducted on sets of six and fewer labels. Most of the modeling was done on the combined set of chosen route data from the Arnhem-Apeldoorn and Amsterdam-Purmerend study areas. Alternative specifications tested the explanatory power of various combinations of level-of-service variables as well as various forms for the constants in the utility functions of the alternatives.

A number of model runs explored the effects of applying separate models for various subgroups in the population. Information from the survey responses was used to assign individual drivers to categories of trip length, trip frequency, and trip purpose. A surprising result from the choice modeling analysis was the relatively significant effect of geographical area that could not be explained in terms of differences in trip purpose, length, or frequency profile for the study areas.

Tables 4 and 5 show, respectively, the six-label and four-

TABLE 2 NUMBERS OF CHOSEN ROUTES MATCHED BY SIX AND NINE LABELS

Label	β Coef.	Match Scores							
	Value(s)	Utrecht-Amersfoort		Arnhem-Apeldoorn			Amsterdam-Purmerend		
		Absolute	Incremental	Absolute	Incremental	Marginal	Absolute	Incremental	Marginal
			%		%	%		%	%
Labels not requiring β parameters									
Time		1505	69.9	1659	56.1	0.3	1236	67.4	1.4
Distance		462	1.6	1905	13.6	12.1	357	0.5	0.5
Labels requiring β parameters: initial values									
Scenic	2.0	770	5.7	1288	1.3	1.1	115	0.3	0.2
Signals	30 sec.	851	2.7	1639	0.1	0.1	1233	0.1	0
Capacity	1.5	1058	2.5	1497	0.3	0.3	1170	3.1	0.4
Hierarchy	5.0, 100	712	3.3	1299	0.5	0.5	1246	6.1	6.1
Total 6 labels		1846	85.8	2179	72.0		1436	77.5	
Labels not included in models									
Quality	2.0	1677	0.4	1435	1.6	1.6	1237	0	0
Commercial	1.5	1506	0	1664	0.1	0.1	887	0.8	0.8
Expressway	3.0	501	0.1	1499	0.0	0.0	1197	0	0
Total 9 labels		1857	86.3	2179	73.7		1436	78.3	
Labels requiring β parameters: final values									
Scenic	2.0			1288	1.3	1.1	115	0.3	0.2
Signals	5 min.			1639	0.3	0.3	1233	4.5	3.4
Capacity	2.0			1497	1.8	1.8	1170	3.4	0.8
Hierarchy	5.0, 100			1299	0.5	0.5	1246	5.1	5.1
Total 6 labels				2179	73.6		1436	81.2	

TABLE 3 COMPARISON OF LABEL SET COVERAGE

Study Area	Time Only (%)	Four Labels[a] (%)	Six Labels (%)	Full-Label Set[b] (%)
Utrecht-Amersfoort	69.9	N/A[c]	85.8	86.3
Arnhem-Apeldoorn	56.1	70.7	73.6	73.7
Amsterdam-Purmerend	67.4	77.6	81.2	78.3

[a] Time, distance, signals, and hierarchy labels make up this set.
[b] Initial, not optimal, label coefficient values were applied for Arnhem-Apeldoorn and Amsterdam-Purmerend corridors.
[c] N/A = not applicable.

TABLE 4 ESTIMATION RESULTS FOR REFERENCE SIX-LABEL MODEL

Variable	Coef. Estimate	Standard Error	T-Ratio
Estimated with separate label-specific dummies by area			
Total travel time (minutes)	-0.0979	0.053	-1.8
Total distance (kilometers)	-0.895	0.099	-9.1
Scenic time (minutes)	0.0767	0.025	3.1
Number of traffic signals	-0.138	0.041	-3.4
Expressway distance (km)	0.108	0.022	4.8
High road quality distance (km)	0.365	0.075	4.9
Low road hierarchy time (min)	-0.0877	0.017	-5.3
Overall label-specific dummy variables			
Minimum time route	0.119	0.064	1.9
Minimum distance route	0.881	0.083	10.7
Maximum scenic route	-0.0527	0.088	-0.6
Minimum signals route	0.722	0.081	8.9
Maximum capacity route	0.571	0.077	7.4
Hierarchical travel route	0.888	0.076	11.7

	First Run	Second Run
Total number of observations:	3052	3052
Likelihood with zero coeffs.:	-3459.6	-3459.6
Final Likelihood:	-1038.7	-1363.4
$p^2(0)$:	0.700	0.606

label models resulting from this analysis. These models include seven generic level-of-service variables, and Tables 4 and 5 show the coefficient estimates for all variables included in the alternative utility functions. The label-specific dummy variables take a value of 1 if the indicated label is matched by the route alternative in question and 0 otherwise.

Initial models with generic level-of-service variables and label-specific dummies applicable across study areas failed to yield significant, correctly signed level-of-service coefficients (e.g., total travel time). Thus a two-step estimation process was used to develop the models presented in Tables 4 and 5. First, a model specification with separate sets of label-specific dummies for each area was estimated, producing significant level-of-service coefficient values of correct sign and relative magnitude. Because this model cannot be applied generally with respect to geographic area, a second estimation is required, yielding values for a single set of label-specific dummies while constraining the level-of-service coefficients to the values obtained in the previous estimation.

Evaluation of Choice Models

The choice models of Tables 4 and 5 can be compared in several terms, including data requirements, chosen route coverage, goodness of fit, and values of model coefficients.

TABLE 5 ESTIMATION RESULTS FOR REFERENCE FOUR-LABEL MODEL

Variable	Coef. Estimate	Standard Error	T-Ratio
Estimated with separate label-specific dummies by area			
Total travel time (minutes)	-0.198	0.061	-3.3
Total distance (kilometers)	-0.577	0.101	-5.7
Scenic time (minutes)	0.145	0.031	4.7
Number of traffic signals	-0.0849	0.052	-1.6
Expressway distance (km)	0.0936	0.031	3.0
High road quality distance (km)	0.206	0.072	2.8
Low road hierarchy time (min)	-0.0824	0.018	-4.4
Overall label-specific dummy variables			
Minimum time route	0.448	0.060	7.4
Minimum distance route	1.64	0.115	14.2
Minimum signals route	1.10	0.093	11.8
Hierarchical travel route	1.47	0.098	15.0

	<u>First Run</u>	<u>Second Run</u>
Total number of observations:	2635	2635
Likelihood with zero coeffs.:	-2434.5	-2434.5
Final Likelihood:	-726.5	-1087.8
$p^2(0)$:	0.702	0.553

In terms of data requirements, both models require information for the seven level-of-service attributes for all alternative routes. The only additional difference between the six-label and the four-label models is that the former requires the road capacity data necessary to generate the capacity label. In this analysis, capacity was calculated on the basis of numbers of lanes and road width. The four-label model requires somewhat less computation to run because only four labels must be generated for all relevant origin-destination pairs as opposed to six for the other model.

Considering chosen route coverage, the six-label model was based on approximately 3 percent more (3,667 versus 3,563) chosen route observations than the four-label model. This is because the inclusion of two extra labels in the model specification allowed the analysis of the behavior of an additional sample of drivers to take place—namely, those 104 drivers who were observed to choose a "maximum scenic" or "maximum capacity" route that did not overlap the other four labeled routes.

The likelihood and ρ^2-statistics of each model indicate how well the implied predictions for models about route choice fit the observations for the available sample of drivers. Strictly speaking, the ρ^2-values for these two models are not comparable because they were not estimated on the same set of observations. Nevertheless, keeping these reservations in mind, the ρ^2-statistic of the six-label model apparently indicates somewhat better fit to the data for the two study areas—Arnhem-Apeldoorn and Amsterdam-Purmerend.

The coefficient estimates of the level-of-service variables for both models all have the intuitively correct sign. For example, one would expect increasing travel time to lead to decreasing attractiveness of the alternative, and indeed the travel time coefficient has a negative sign. Similarly, road quality, scenic time, and distance on expressways are all hypothesized as positive qualities of a route, and these variables have positive signs. Another measure that may be used to appraise the reasonableness of a model is its implied "value of time," which is calculated here by determining the ratio between the time and distance coefficients and factoring in an assumed operating cost per unit distance.

Assuming a marginal cost of driving of $0.15 per mile (gasoline costs about $2.90 per U.S. gallon in the Netherlands at current exchange rates), the implied values of time for a major nonexpress road that is neither scenic nor of high quality are $0.61 per hour and $1.92 per hour for the six- and four-label models, respectively. For a minor road that is neither scenic nor of high quality, the respective values are $1.16 per hour and $2.72 per hour. Although all these estimates appear to be on the low side, the values of the four-label models agree more closely with other sources of value-of-time estimates.

In conclusion, the differences between the two final models are not very great. The six-label model (Table 4) is based on more observations and shows better fit to the observed data, whereas the four-label model requires fewer data to operate and has a more reasonable implied value of time. If a choice were to be made between application of one model or the other, the six-label model would be recommended unless capacity data were difficult to come by or the value of time were perceived as too low based on other studies.

In practice, several of the variables used in these models are not likely to be available for the networks to which the models are to be applied. For these circumstances, reduced models were developed in which the requirements for data were substantially reduced or omitted, for example, scenery, road quality, and traffic signals. These models are based on four or even three labels. The loss of explanatory power of these reduced models compared with the models of Tables 4 and 5 is the inevitable consequence of the omission of the relevant variables. Fortunately, some variables other than time and distance, such as hierarchical level, speed limit, and capacity, are generally available in the Netherlands.

Other Results

Apart from the variables incorporated in the models presented in Tables 4 and 5, several other variables were considered for inclusion in the models. Some of these could be eliminated because of their excessively high correlation with variables already included in the models, others because they were not found to significantly influence route choice. In particular, income-dependent effects were carefully tested, but no significant influences could be found.

Further tests were made of differences in behavior among drivers traveling for various purposes, making trips of varying lengths, or traveling with various frequencies. Although some differences of these types were found, they were much smaller than the differences with respect to geographical area.

In general, despite the differences between areas just mentioned, a substantial degree of transferability was found among the three areas for which data were available. As noted above, the labeling procedure was transferable without problems; the choice models lost explanatory power in the transfer but still gave useful and reliable results.

Structural tests were also made on the models estimated. Again some evidence was obtained of failure of the independence assumption on which the logit model is based, but this was not sufficiently serious to cause the structure to be abandoned. Moreover, there was no simple way in which the structural divergence could be approximated.

CONCLUSIONS AND RECOMMENDATIONS

A method has been developed for describing the general route choice behaviors of car drivers. The method is based on the "labeling" of alternative routes that provide realistic possibilities for each driver's journey. A probability model then represents the choice among these route alternatives.

The method is based on a fundamental reassessment of the choice processes that lead to the selection of routes and the analysis of the choices actually made by nearly 7,000 drivers observed in three corridors in the Netherlands.

Several road characteristics other than time and distance are found to be important in influencing route choice. Of particular relevance to policy is the finding that characteristics associated with major roads (restricted access, high speed limit, high capacity, hierarchical status) are strongly positive in influencing route choice. Scenery (positive) and traffic lights (negative) are also found to be relevant.

Even under uncongested circumstances, several routes are used for a given journey. The models estimated identify these routes and predict the proportion of vehicles that will use them. The fact that these predictions are based on models formulated by observing behavior rather than on an arbitrary basis as in some algorithms in current use gives much more confidence in their use.

Application procedures have been developed for the models. These procedures take into account the existing sophisticated methods for the treatment of capacity constraint. The application of the route choice models would add little to the computer time needed to make an assignment and would require little additional software.

Reduced models have been developed to be applied in circumstances of reduced data availability.

Further development of route choice analysis is required to account for two important aspects:

1. The information available to the driver is not currently modeled. Apart from fixed sign posting, interest in formulating policy on the dynamic provision of information is growing, and it is important to know the extent to which drivers might be influenced by methods of providing it.
2. Cost is incorporated into the models only weakly, through the distance variables. The policy under consideration includes "road pricing," whereby drivers would pay much more directly for the use of roads; the influence of such measures on route choice, however, needs to be investigated.

A third aspect that might be considered is the apparent safety of one route compared with another and how that affects route choice.

ACKNOWLEDGMENTS

This project was conducted jointly by Hague Consulting Group and the Transportation Systems Division of the Netherlands

Ministry of Transport (the project's sponsor). The authors wish to acknowledge the staff of the Ministry who contributed to the study. They also wish to thank Theo Bergman, Lionel Silman, Steve Pitschke, and Rohit Ramaswamy, who contributed greatly to the earlier phases of this study.

REFERENCES

1. M. Ben-Akiva, M. J. Bergman, A. J. Daly, and R. Ramaswamy. Modelling Inter-Urban Route Choice Behavior. In *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*, Kijkduin, The Netherlands, VNU Science Press, Utrecht, 1984, pp. 299–330.
2. H. Aashtiani and W. Powell. Route Choice Models in the Traffic Assignment Process. Massachusetts Institute of Technology, Cambridge, Mass., 1978.
3. N. Hidano. Driver's Route Choice Model: An Assessment of Residential Traffic Management. Presented at the WCTR, Hamburg, West Germany, 1983.
4. P. H. L. Bovy. Het Kortste-Tijd Routekeuzecriterium: Een Empirische Toetsing. Presented at the Colloquium Vervoerplanologisch Speurwerk, The Hague, The Netherlands, 1979.
5. M. H. Ueberschaer. Choice of Routes on Urban Networks for the Journey to Work. In *Highway Research Record 369*, HRB, National Research Council, Washington, D.C., 1971, pp. 228–238.
6. F. Tagliacozzo and F. Pirzio. Assignment Models and Urban Path Selection Criteria: Results of a Study of the Behavior of Road Users. *Transportation Research*, Vol. 7, 1973, pp. 313–329.
7. C. C. Wright. Some Characteristics of Driver's Route Choice in Westminster. *Proc., PTRC Summer Annual Meeting*, PTRC Education and Research Services, Ltd., London, 1976.
8. J. E. Burrell. Multiple Route Assignment and Its Application to Capacity Restraint. In *Fourth International Symposium on the Theory of Traffic Flow*, Karlsruhe, West Germany (W. Leutzbach and P. Baron, eds.), 1968.
9. R. B. Dial. Probabilistic Assignment: A Multipath Traffic Assignment Model Which Obviates Path Enumeration. Ph.D. dissertation. University of Washington, Seattle, 1970.
10. R. Hamerslag. Onderzoek naar Routekeuze met Behulp van een Gedisaggregeerd Logitmodel. *Verkeerskunde*, No. 8, 1979, pp. 377–382.
11. H. Morisugi, N. Miyatake, and A. Katoh. Measurement of Road User Benefits by Means of a Multi-Attribute Utility Function. *Papers of the Regional Science Association*, Vol. 46, 1981, pp. 31–43.
12. V. E. Outram. Route Choice. *Proc., PTRC Summer Annual Meeting*, PTRC Education and Research Services, Ltd., London, 1976.
13. E. J. Lessieu and J. M. Zupan. River Crossing Travel Choice: The Hudson River Experience. In *Highway Research Record 322*, HRB, National Research Council, Washington, D.C., 1970, pp. 54–67.
14. D. L. Trueblood. Effect of Travel Time and Distance on Freeway Usage. *Bulletin 61*, HRB, National Research Council, Washington, D.C., 1952, pp. 18–35.
15. M. Wachs. Relationship Between Driver's Attitudes Toward Alternative Routes and Driver and Route Characteristics. In *Highway Research Record 197*, HRB, National Research Council, Washington, D.C., 1967, pp. 70–87.
16. R. M. Michaels. The Effect of Expressway Design on Driver Tension Responses. *Public Roads*, Vol. 32, No. 5, 1962.
17. H. J. Wootton, M. P. Ness, and R. S. Burton. Improved Direction Signs and the Benefits for Road Users. *Traffic Engineering and Control*, 1981.
18. M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, Mass., 1985.
19. Hague Consulting Group. *ALOGIT User Documentation*, The Hague, The Netherlands, 1988.
20. J. D. Bolland, M. D. Hall, and D. van Vliet. SATURN: A Model for the Evaluation of Traffic Management Schemes. *Working Paper 106*. Institute for Transportation Studies, University of Leeds, Leeds, England, 1979.

Demand Diversion for Vehicle Guidance, Simulation, and Control in Freeway Corridors

YORGOS J. STEPHANEDES, EIL KWON, AND PANOS MICHALOPOULOS

Rapidly increasing traffic volume, congestion, and excessive delay are making the management, control, and guidance of traffic flow one of the most critical transportation problems in urban freeway corridors. Modeling demand diversion to less congested routes within a corridor is a necessary part of demand modeling efforts for improved simulation and control, as are guidance-navigation systems in real time. Models for describing diversion at the trip origin and diversion at freeway entrance ramps are discussed. Data collected in a major metropolitan area have shown that diversion at the origin is a function of trip time, route length, and the number of intersections along the trip. However, trip time is the dominant determining factor and can be employed to estimate the decision in the absence of additional information. Diversion at freeway entrance ramps depends on the perceived trip time on the freeway and arterial and the perceived waiting time at the ramp queue. The data confirm that socioeconomic indicators do not play a role in the diversion decision. The purpose of developing these models is for dynamic simulation, on-line freeway corridor control, and demand forecasting suitable for guidance and navigation.

Rapidly increasing traffic volume and the ensuing congestion and excessive delay are making the management and control (guidance) of traffic flow one of the most critical transportation problems on urban freeways. To remedy the problem, corridor management seeks to divert freeway drivers away from the congested segments of freeway corridors to alternative routes within a corridor, such as adjacent arterials. The diversion can occur at the beginning of the trip, before entering the freeway ramp, or on the freeway.

Demand diversion, generally caused by excessive delay and ramp queues, is a major problem (*1–3*) that has not been effectively considered in real-time control systems, although recently an effort has been made to address the problem (*4*). The major difficulty lies in the rapidly changing traffic flow conditions; furthermore, substantial instrumentation is required to collect data for modeling traffic diversion. Determination of realistic control policies and effective guidance-navigation schemes for the freeway corridor should include diversion as an integral part. Existing literature (*1,2*) suggests that there is a lack of an on-line demand predictor suitable for real-time control for interconnected ramps and arterials. However, existing demand diversion models are not suitable for effective real-time freeway control strategies. Current demand diversion models are based on assumptions that are considered unrealistic (*5*), such as user-optimized equilibrium flow patterns, perfect knowledge of traffic conditions ahead, and infinite storage capacity on surface streets. Diversion is only a part of the more general demand prediction problem.

Modeling of demand diversion is addressed in this paper; this modeling was needed to develop a reliable prediction algorithm suitable for implementing real-time control policies (*6,7*) in freeway corridors. Within this context, diversion is an essential element necessary for proper estimation of traffic demand as well as for determination and simulation of the optimal control strategy or guidance plan. The diversion models presented here can be used with a demand predictor (*7*) to simultaneously determine ramp demands and diversion volume as part of an integrated corridor simulation-control-guidance process in real time.

The models should be appropriate for employment in guidance-navigation systems that use information on current traffic conditions for selecting optimal routing in real time. In such systems the models are needed to estimate the impact that the guidance-navigation information has on drivers. In particular, guidance-navigation systems are expected to respond to drivers' queries by providing information on freeway and arterial delays, freeway ramp queues, and the resulting ramp delays as freeway conditions and ramp metering rates change with time.

A critical review of the most widely accepted research on the diversion problem is presented first. This review includes a summary of model features that emphasize effectiveness and drawbacks of each approach from the limited tests found in the literature. Subsequently, two utility-based demand diversion models are developed, one for the diversion at the trip origin and one for the diversion at freeway ramps. The models are tested with data from the I-35W freeway corridor in the south area of the Twin Cities—Minneapolis and St. Paul, Minnesota.

Consistent with expectations, the model specifications indicate that trip time is the dominant factor determining diversion at the trip origin, whereas route length and the number of intersections along the trip also play significant roles. Diversion at freeway entrance ramps depends on the perceived trip time on the freeway and arterial and the perceived waiting time at the ramp queue. The data confirm that socio-economic indicators do not play a role in the diversion decision. Further, for commuter trips shorter than 1 hour, freeway drivers consider only one diversion alternative, a preferred arterial, and do not divert to downstream ramps. The diversion models require only limited data for implementation.

Department of Civil and Mineral Engineering, University of Minnesota, Minneapolis, Minn. 55455.

BACKGROUND

Freeway corridor models have considered diversion within the context of control and assignment by determining the long-term equilibrium flow pattern that satisfies Wardrop's principle within a given time slice or by employing self-assignment (i.e., assuming that omniscient drivers can find the quickest route at each decision point of their trip). Although some researchers determined the flow pattern through a combination of models, others sought to increase computational efficiency and avoid potential modeling inconsistencies by developing a single modeling approach. Further, earlier methods (*8*) may, by assumption, limit diversion to occur only at the trip origin, whereas more recent methods offer the flexibility of allowing diversion at multiple points during the trip.

Diversion methods that are based on a combination of models are older. Lieberman (*9*) developed a freeway corridor simulation program, SCOT, by combining DAFT, a macroscopic corridor simulation model, with UTCS-1. Traffic flow on nonfreeway links is treated as a collection of individual vehicles, each processed every second of simulated time; in contrast, the freeway flow is described macroscopically, which permits the grouping of vehicles into platoons and the use of a coarser time step. With the origin-destination (O-D) demand matrix or turning movements at each node specified by the user, traffic is routed following the minimum-time path, which the user recalculates successively by selecting the time interval.

Another composite model that incorporates diversion within a freeway corridor simulation, CORQ1C, was proposed by Orthlieb and May (*10*). CORQ1C allows diversion from the freeway to arterials only for the "flexible" users whose destinations are within the corridor boundaries, whereas other users have fixed O-D routes. The model combines FREQ3 and TRANSYT5 to simulate the diversion following a linear-programming decision process that selects the optimal ramp metering rates. The corridor assignment associated with the optimal rates maximizes the total trip time savings for the flexible users of the freeway. In each 15-min time slice, after all fixed-route demand has been distributed, the decision process incrementally assigns the optimal flexible-route demand subject to corridor capacity constraints. For each optimization increment, a constant value of time savings for the flexible users is estimated from simulating the previous traffic loadings in the corridor. After each optimization, the resulting optimized volume is assigned and the new value of time savings is found. This method assumes that diversion is possible only at the trip origin.

In contrast to the earlier methods, FREQ7PE (*11*) is based on a single program rather than a combination of programs. At each 15-min time slice the method calculates the optimal ramp metering rates for the given O-D ramp volumes that optimize freeway objectives. The resulting diversion is determined by estimating the equilibrium flow pattern in the corridor for each time slice. An iterative assignment procedure is performed until the travel time difference between any alternative routes for each O-D pair is within an acceptable range. Evidently this procedure allows for diversion at several alternative ramps.

The models in the CORQ (*12*) family use a form of microassignment corridor technique with the given O-D zone demand divided by 15-min time slices. For each time slice, a minimum-time path is constructed for all O-D pairs, and an incremental assignment is performed by iteratively updating the link cost. The remaining demand at each time slice is stored at the upstream node of a link where it is queued and assigned at the next time slice with the new demand and updated minimum-time path. CORCON (*13*) extended the minimum path assignment algorithm of CORQ by incorporating turn prohibitions and a traffic diversion procedure from the queueing link to the nonqueueing alternative on the basis of travel cost (time) difference. However, both models assume unlimited queue storage capacity of arterials and drivers' perfect knowledge of the existing traffic condition in the network. Although these assumptions are not realistic, the ability to determine and keep track of queues is an advantage over the previous methods.

INTEGRATION-1 (*14*) is a microscopic corridor simulation model, which, unlike previous methods, considers the behavior of traffic flow in terms of individual vehicles that have self-assignment capabilities. The model is not based on the time-slice approach; rather, it assigns individual vehicles sequentially to a network that is already loaded with any previous departures that have not reached their destination. The turning movement of each vehicle at each node and instant is dictated by the minimum-path tree table existing at that instant and is recalculated every 6 sec. The main difference between CORQ and INTEGRATION-1 is that CORQ considers vehicle flow rates for an entire time slice, whereas INTEGRATION-1 treats individual vehicles on a continuous basis. The departure times of all trip demands are given, and drivers are assumed to have full knowledge of the existing traffic conditions on the entire network.

In addition to the above methods, a number of models developed for network simulation implicitly consider diversion. Of these, TRAFLO (*15*) and SATURN (*16,17*) are worth mentioning because of their extensive use by government and private organizations. These composite models implicitly consider diversion in the larger context of simulation and assignment. In particular, TRAFLO combines an equilibrium assignment model with four different simulation models that estimate the expected performance of the assigned flows. However, the assignment model does not have the feedback function that can employ the refined travel time and queue size estimates to update and correct the initial traffic assignment assumptions. Although SATURN adopts an iterative procedure to correct and update the network parameters for the assignment, it currently uses all-or-nothing assignment; further, it assumes a cyclic flow profile, only suited for signalized arterials. Such assumptions limit its applicability for freeway corridor analysis.

To effectively control the traffic flow in a corridor, the estimation of the time-dependent flow pattern of the diverting traffic is of critical importance. As the above review indicates, existing diversion methods determine the equilibrium flow pattern satisfying Wardrop's principle at each time slice either macroscopically or by employing the self-assignment technique, thus assuming that drivers can find the quickest route at each decision point with perfect knowledge of traffic conditions ahead. However, it has been argued that Wardrop's principle is not applicable to the dynamically changing traffic environment mainly because of the human nature of drivers; that is, drivers are not well informed or are not sufficiently skilled to choose the best route (*5*).

Understanding commuter reactions to ramp control strategies and guidance-navigation information on freeway and arterial trip characteristics is essential in estimating and controlling corridor flow to decrease congestion. This paper proposes a utility-based approach for the dynamic diversion problem, which, when combined with an appropriate filter, will more realistically model the commuter diversion process for simulation, control, and guidance-navigation in congested freeway corridors.

For the purposes of this analysis we assume that diversion occurs at two points: the trip origin and the entrance to the freeway ramp. Although diversion can occur at any point during the trip, all intermediate decision points were included in the stated two because of time and data limitations and the need to immediately employ a diversion model that addresses the points where most drivers make a route diversion decision. The following sections summarize the model formulation and the parameter estimation results.

MODEL FORMULATION

The structure of the overall diversion-control-guidance modeling approach can be analyzed at several levels of detail. At the most general level, it may be pictured as a sequential process (Figure 1) with the freeway corridor performance sector acting as a link between traffic diversion and changes in freeway controls and guidance-navigation information. For instance, at the trip origin, trip makers select either the freeway or the arterial route, depending on their corresponding perceived trip times, which are functions of known variables such as volume and capacity. Their perception is enhanced with the updated information they receive from radio and TV and from guidance-navigation systems, if such are in operation.

Even though the initial route of choice may be the freeway, at the entrance ramp the freeway commuter can still decide

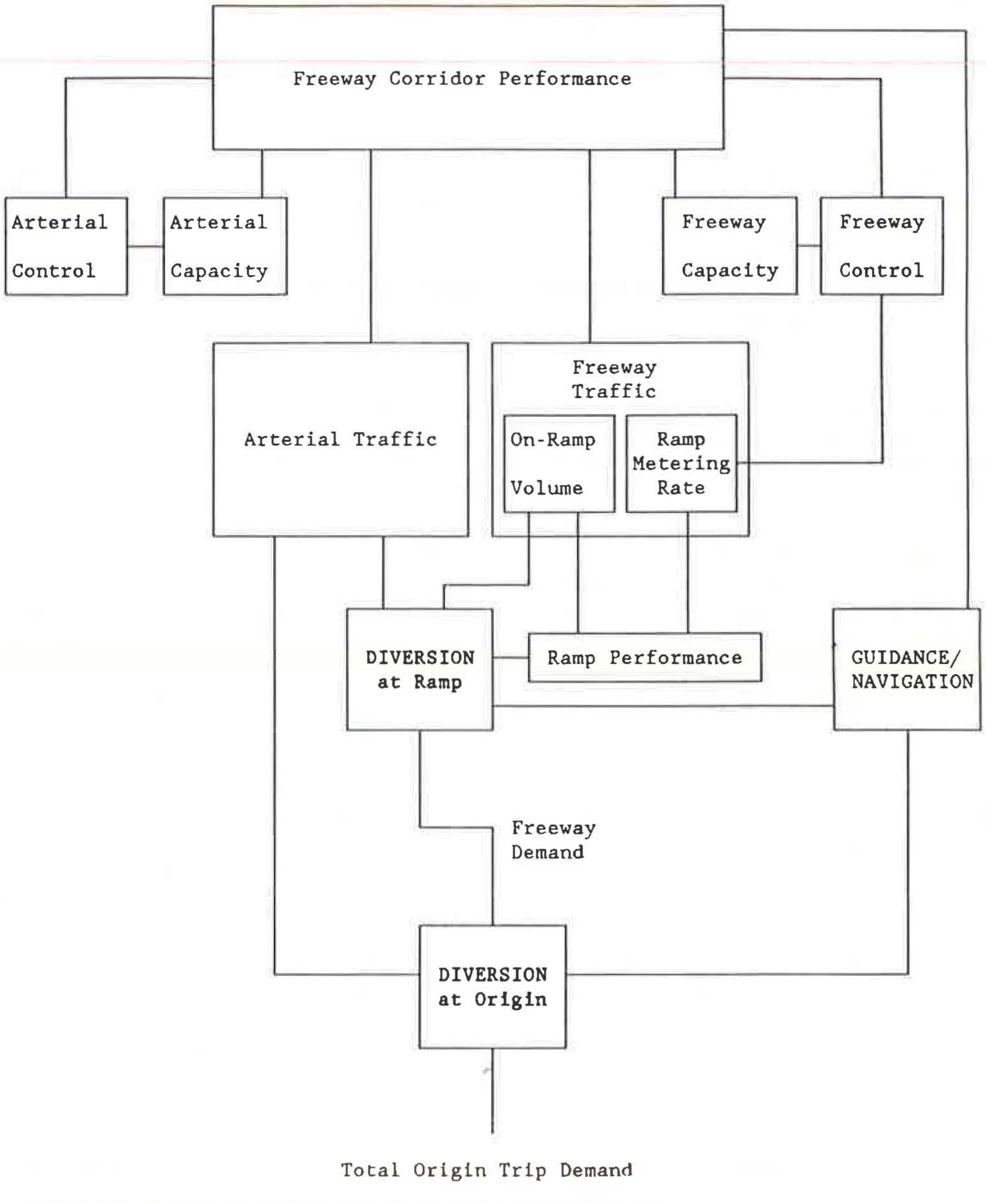

FIGURE 1 Demand diversion, control, and guidance-navigation.

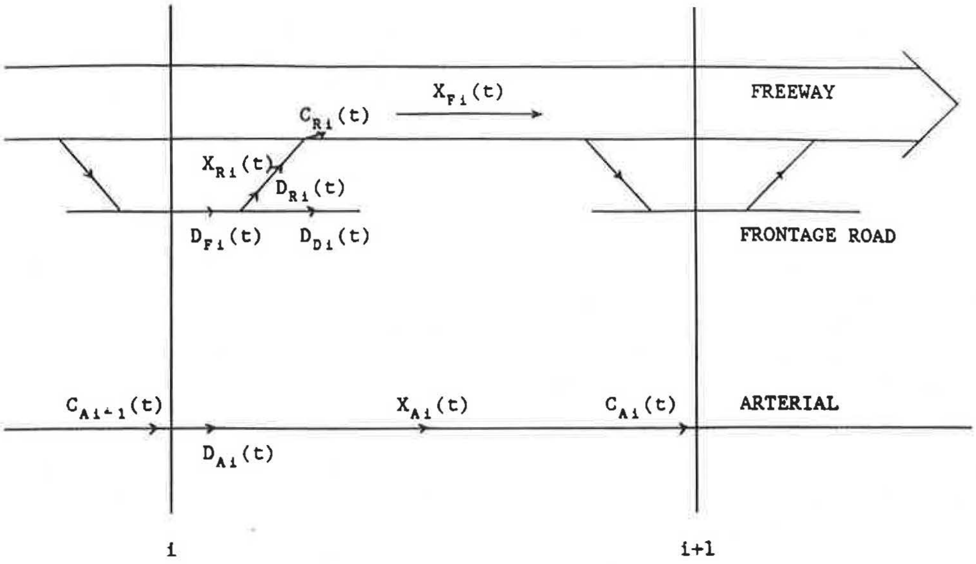

FIGURE 2 Sample freeway corridor.

not to enter; rather, the vehicle can divert to an arterial street depending on the ramp traffic situation. This decision is again enhanced by any additional information the driver has been receiving from guidance-navigation or other communication systems. The diversion decision of each driver affects the overall volume on the freeway and arterial(s) and, thus, the performance of the freeway corridor. In turn, the corridor performance is used as a basis for setting the control in the corridor, such as ramp metering rate and arterial signals, further affecting the corridor performance.

Communications and guidance-navigation systems pick up the current performance information and transmit it to the drivers, who can update their diversion decisions, and the process continues full circle. Therefore, the traffic diversion process reflects the short-term reaction of traffic flow to the control and guidance schemes, and the resulting congestion patterns in the dynamically changing traffic environment.

Tracing the diversion, control, and guidance-navigation interactions through time is done on the basis of component equations that are used to model the diversion, filter the traffic flow measurements, and set the desired control and guidance strategy. In this paper, we focus on the development of the diversion equations.

Because of the limitations of the existing models, dynamic freeway diversion equations were developed that fulfill the requirements of the time-sensitive approach followed in this work. Assuming for the purposes of this discussion that the freeway model and all other component equations are complete, the diversion equations apply the conservation principle to the freeway ramp and adjacent arterial(s) to determine the traffic volume as a function of known inputs and outputs and the state of the system. As a reference to the diversion equations, Figure 2 shows an example corridor system consisting of a freeway with an entrance ramp, a frontage road, a parallel one-way arterial street, and cross streets connecting the arterial with the freeway entrance ramp. For simplicity, the frontage road is used only for the diversion from the ramp, and the diverted traffic volume directly joins the arterial flow.

Applying the conservation principle to the ramp and arterial link, respectively, for a suitably small length of time slice t, the state evolution equations for the ramp and the arterial of corridor component $(i, i + 1)$ can be written:

$$X_{Ri}(t) = X_{Ri}(t - 1) + D_{Ri}(t) - C_{Ri}(t) \qquad (1)$$

$$X_{Ai}(t) = X_{Ai}(t - 1) + I_{Ai}(t) - C_{Ai}(t) \qquad (2)$$

where

$X_{Ri}(t)$ = number of vehicles on ramp i at time slice t,
$D_{Ri}(t)$ = vehicles entering ramp i at t,
$C_{Ri}(t)$ = vehicles exiting ramp i at t,
$X_{Ai}(t)$ = vehicles on arterial link $(i, i + 1)$ at t,
$I_{Ai}(t)$ = vehicles entering arterial link $(i, i + 1)$ at t, and
$C_{Ai}(t)$ = vehicles exiting arterial link $(i, i + 1)$ at t.

Then, on the basis of the concept of utility, the input volumes for the entrance ramp and the arterial link are

$$D_{Ri}(t) = D_{Fi}(t) * P_{Ri}(t) \qquad (3)$$

$$I_{Ai}(t) = C_{Ai-1}(t) * Q_i(t) + D_{Ai}(t) + D_{Di}(t) \qquad (4)$$

where

$D_{Fi}(t) = D(t) * \exp[V_F(t)]/\Sigma\exp[V'(t)]$
= freeway trip demand at trip origin at t, (5)

$P_{Ri}(t) = \exp[U_R(t)]/\Sigma\exp[U'(t)]$
= portion of $D_{Fi}(t)$ entering ramp i at t, (6)

$D_{Ai}(t) = D(t) * \exp[V_A(t)]/\Sigma\exp[V'(t)]$
= arterial trip demand diverted from origin at t, (7)

$D_{Di}(t) = D(t) * \exp[V_F(t)]/\Sigma\exp[V'(t)] * [1 - P_{Ri}(t)]$
= diverted volume at entrance ramp to arterial at t, (8)

$Q_i(t)$ = portion of $C_{Ai-1}(t)$ entering arterial link $(i, i + 1)$ at t,

$D(t)$ = total demand originating from this corridor section at t,

$V(t)$ = utility of freeway (V_F) or arterial (V_A) route for diversion at origin at t (see Table 2), and

$U(t)$ = utility of entering ramp (U_R) or diverting (U_D) to arterial at entrance of freeway ramp at t (see Table 4).

The above model assumes that the state evolution is first order, with the diverting volume estimated from disaggregate data collected in the study area. The exit volume $C(t)$ can be estimated as a function of the link volume and the physical characteristics of the link, or, in real-time application, the actual measured volume can be used to update the model parameters using filtering techniques (*18*). The model implicitly assumes that time slice t is suitably small or the link is relatively long.

Using the proposed model, the optimal control in the freeway corridor minimizing total system travel time for the given time period can be formulated as follows:

find optimal control policy $u(t)$ to minimize

$$\sum_{t=0}^{T}\left(\delta t * \{X_A[t, u(t)] + X_R[t, u(t)] + X_F[t, u(t)]\}\right) \qquad (9)$$

(subject to corridor flow standards and management constraints)

where

$X_F(t)$ = number of vehicles in freeway section at time slice t,

δt = size of time slice, and

T = number of time slices in optimization period.

We are now validating the proposed model using corridor traffic data. In this paper we report the estimation results of the utility functions for the diversion decisions. The comprehensive validation results will be presented in a forthcoming paper.

PARAMETER ESTIMATION

Route Diversion at Trip Origin

Before their departure, commuters make their initial decision on which route to take for their trip to work. In general, this decision considers two major determining factors—the set of alternative routes for the trip and the characteristics of each route. In this work we assume that the set of possible trip routes consists of a freeway and an arterial. Our extensive surveys indicate that very few commuters (less than 3 percent) seriously consider a third alternative and, even then, they select that alternative only in low-likelihood circumstances (e.g., in a severe snowstorm).

We estimated the route diversion at the origin by specifying a binary logit model for the freeway and arterial alternatives. For this model, we define the freeway alternative (and, similarly, the arterial) as a trip route that is at least 80 percent freeway. Model variables can be of two types—trip related and socioeconomic. The three trip-related variables are

- travel time (T) in minutes, the one-way trip time in the vehicle;
- route length (L) in miles, the one-way trip distance; and
- number of intersections (I), the number of intersections crossed by the vehicle along the one-way trip. (If the exact number is not available, a range of values can be used; e.g., suggested range is low at $I < 15$, medium at $15 < I < 45$, high at $45 < I$.)

Management and control policies can directly affect the above variables. For instance, for the same trip route, changes in ramp metering rates and in the number of freeway lanes available will affect the travel time. Similarly, ramp closings and construction detours will increase the route length. Reduced access at intersections will decrease the number of intersections experienced by the trip maker on the priority access road. Of course, changes that are of a more substantial nature, such as bridge reconstruction, a new bypass, or a new ramp, may develop new alternatives for a subset of drivers; in such cases, the new values for the above variables must be entered in the diversion specification.

Drivers are expected to know the value of each of the above variables for the two major commuting alternatives. Such values rarely change, but when they do, updated information is likely to become widely known to commuters because it is routinely communicated through newspapers, radio, and television. Up-to-the-minute information on changes resulting from unforeseen events, such as freeway incidents, is also commonly available through special radio or TV announcements and would be part of guidance-navigation systems in urban areas. Real-time information on incidents is smoothed by the departing driver depending on the planned trip departure time, a subject that we are currently analyzing.

In addition to the above trip-related variables, we tested annual household income, a socioeconomic variable proposed by Abu-Eisheh and Mannering (*19*) for the route choice process. However, we did not expect, and our tests did not indicate, this variable to play a role in the diversion.

A questionnaire survey of 500 households was conducted, and individual characteristics were recorded for the commuters with trips originating in the south I-35W corridor in November 1987 (see Figure 3 for an illustration of this freeway corridor crossing the Twin Cities in a north-south direction). Following data treatment, 105 employees having a common destination were selected as the sample commuters. All commuters in the sample had the choice of driving in a northerly direction using a predominantly freeway route or an adjacent, one-way arterial (Park Avenue, see Figure 3). Although the data treatment resulted in a decreased sample size, the improved quality of the treated sample contributed to an increased significance and robustness of the estimated model parameters.

Each employee was asked to draw his or her freeway and arterial routes on the map, indicating the initial choice and expected travel time for each route under normal conditions. From this information, the detailed trip characteristics including route length, number of intersections, and number of turns were obtained. Further, the socioeconomic characteristics of each driver were provided from the questionnaire (Table 1). Sample commuters were evenly distributed in the study area, and most sample characteristics were almost-Gaussian distributed (number of intersections was missing the left tail).

Three disaggregate models to estimate diversion at the trip origin were derived from the Twin Cities data (Table 2). Model M1 tests the hypothesis that trip time affects the diver-

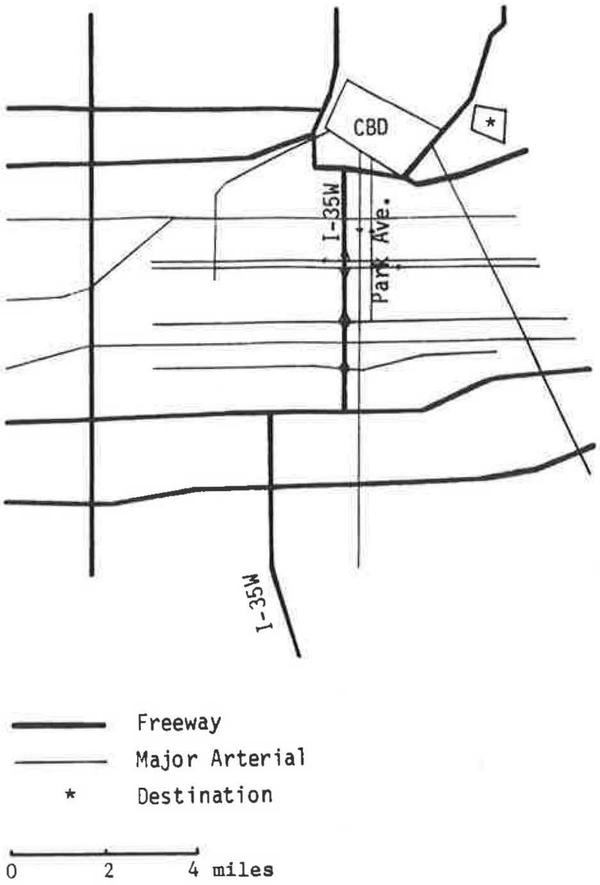

FIGURE 3 I-35W study area.

sion decision, and the estimation results indicate that it is, indeed, a significant factor (99 percent significance level). Model M2 indicates that, although travel time is a dominant factor in diversion, the number of intersections and route length also play a highly significant role. All coefficients have the expected sign; further, model M3 confirms our expectation that the socioeconomic indicator (income) does not influence the diversion.

The estimation results show that when freeway and arterial are the two alternatives, and collection of data on number of intersections and route length is exceedingly costly, model M1 can be employed to estimate diversion at the origin based only on travel time. Model users who wish to employ model M2, either to gain the additional predictive power or because it is necessary for policy purposes, do not have to know the exact number of intersections along the commuter trip. As indicated at the bottom of Table 2, an approximation of the value of this variable can be used to facilitate model implementation.

Ramp Diversion

Commuters approaching the freeway entrance ramp can opt to divert to an alternative route before entering the ramp. Their decision depends on the set of available alternatives and the traffic conditions at the ramp. Our surveys indicate that, for the corridor under study, only a small percentage (less than 4 percent) of drivers divert to a downstream ramp, while the vast majority of diverting drivers select the arterial option. Based on this finding, we have limited the set of route alternatives at the ramp entrance to two (freeway and arterial) and estimated the diversion by specifying a binary-choice logit model.

Model variables can be of two types—trip related and socioeconomic. However, based on our conclusions from modeling the diversion at the origin, and after confirming those conclusions with the data we collected for the ramp diversion, we eliminated all socioeconomic variables from the ramp diversion model. Our final hypothesis included four trip-related variables:

- Freeway travel time (*FTT*) in minutes, the one-way trip time from the point of entering the freeway proper to destination.
- Arterial travel time (*ATT*) in minutes, the one-way trip time from the point of diverting at the ramp entrance to destination.
- Waiting time (*WT*) in minutes, the one-way waiting time at the freeway ramp prior to entering the freeway proper.
- Total travel time (*TTT*) in minutes, equals $WT + FTT$ for the freeway alternative; if the driver diverts, $TTT = ATT$.

Freeway management and control strategies can directly affect the above variables. For example, ramp metering rates have an immediate effect on *WT* and an indirect effect on the main traffic stream on the freeway. Drivers approaching the ramp perceive changes in *WT* by considering the queue length but can only guess about any changes in *FTT* and *ATT* by considering the traffic situation (such as speed and density) in the vicinity of the ramp entrance. Lane closings and maintenance work can affect both *FTT* and *ATT*, but such information is either known to drivers at the origin or not known at all. Additional information on these conditions can be provided through other means of communication, including routine radio announcements and new guidance-navigation systems.

A return-mail questionnaire survey of 600 drivers actually commuting via the I-35W freeway corridor was conducted at three northbound freeway entrance ramps in November 1987. From the 195 usable responses, data were obtained on driver individual characteristics such as trip origin and destination, departure and arrival times, maximum tolerable waiting time and queue size before diverting to the arterial route, travel time of the alternate route from the diverting point, and socioeconomic information (Table 3).

Two disaggregate models of the diversion at the freeway ramp were derived from the data of the corridor sample (Table 4). Models D1 and D2 test the hypothesis that trip time affects the decision to divert at the ramp. Although the results from model D1 indicate that trip time is a significant factor (99 percent level), inspection of model D2 indicates that this variable should be treated as alternative specific rather than generic—a consideration that improves the estimation power of the model from 59 to 71 percent.

All estimated coefficients have the expected sign and a high statistical significance. The specifications reflect our belief that, for commuting trips of the nature encountered in the Twin Cities, the competition between freeway and arterial times should not follow a linear rule. In particular, the diversion should be highly sensitive to trip times that are very short

TABLE 1 SUMMARY STATISTICS OF SAMPLE COMMUTERS AT ORIGIN

		Group 1	Group 2
Sample size		74	31
Annual household income($)		39000	34000
Age		39.5	35.0
Years in area		7.8 yrs	6.0 yrs
Primary Route	Route type	Freeway	Arterial
	Route length	7.2 mi.	6.6 mi.
	Travel time	17.0 min.	20.1 min.
	# Intersections	18	40
	# Turns	5.4	7.4
Alternate Route	Route type	Arterial	Freeway
	Route length	8.8 mi.	7.8 mi.
	Travel time	23.3 min.	21.6 min.
	# Intersections	58	24
	# Turns	6.9	8.7

TABLE 2 ESTIMATED LOGIT COEFFICIENTS FOR DIVERSION AT ORIGIN

Variable	Model M1	Model M2	Model M3
Constant (freeway only)	0.238 (0.83)[*]	-0.348 (-0.95)	-0.432 (-1.13)
Travel Time (min.)	-0.260 (-4.46)	-0.212 (-3.58)	-
Travel Time * Annual Household Income ($1000)	-	-	-0.00513 (-3.46)
Number of Intersections * Route Length	-	-0.00401[**] (-2.59)	-0.00462 (-2.86)
Sum of Chosen Probabilities	73.6	76.7	76.5
Sum Prob. Ratio	0.72	0.75	0.75
Initial Log Likelihood (L_o)	-70.7	-70.7	-70.7
Final Log Likelihood (L_θ)	-44.3	-39.5	-39.6
$\rho^2 = 1 - [L_\theta / L_o]$	0.37	0.44	0.44

[*] t-statistic

[**] For DI between 15 and 45,
if DI < 15 then -0.0232
if DI > 45 then -0.0362
where DI = Intersections in Arterial Route - Intersections in Freeway Route.

TABLE 3 SUMMARY STATISTICS OF SAMPLE COMMUTERS FOR RAMP DIVERSION

Location	Sample Size	Access Time to Ramp (min)	Freeway Travel Time (min)	Arterial Travel Time (min)	Max. Wait Time on Ramp (min)	Max. Queue Size on Ramp (no. of cars)
51st Street	64	5.4	18.8	26.3	7.5	13.0
46th Street	88	6.7	17.5	23.2	5.8	17.9
35th Street	43	7.1	17.2	23.7	8.3	16.9
Total or average	195	6.4	17.8	24.2	6.9	14.8

TABLE 4 ESTIMATED LOGIT COEFFICIENTS FOR RAMP DIVERSION

	Model D1	Model D2
Constant*	-0.751 (-4.03)***	-2.31 (-6.42)
TTT	-0.123 (-5.52)	-
DTT/FTT*	-	-4.71 (-5.29)
DTT/ATT**	-	-18.60 (-7.13)
Sum of Chosen Probabilities	76.7	92.3
Sum Prob. Ratio	0.59	0.71
Initial Log Likelihood L_o	-180.2	-180.2
Final Log Likelihood L_θ	-158.6	-115.3
$\rho^2 = 1 - [L_\theta / L_o]$	0.12	0.36

DTT = ATT - (WT+FTT)

* Freeway only, else 0.

** Arterial only, else 0.

*** t-statistic

but not as sensitive to those that are long. The improved estimation and statistical performance (in terms of *t*-statistic and ρ^2-value) of model D2 is not surprising because commuters are known to attach different values to their time, depending on whether they are traveling on a route where the speed is expected to be high (the freeway) or on one where no such expectation exists (the arterial).

Although implementation requires only limited data, an additional analytical step is needed before the above models become fully operational in a real-time traffic environment. In particular, relationships should be developed between the value of the model variables, which are perceived by drivers, and the value of variables that could be routinely measured by traffic engineers. For instance, a specification should be developed for the relationship between the length of the ramp queue (or the number of cars in queue) and the ramp waiting time perceived by approaching drivers. Such relationships are now under development.

SUMMARY OF THE RESULTS

The rapid increase in the volume of traffic, congestion, and excessive delays is making the management, control, and guidance of traffic flow one of the most critical transportation problems in urban freeway corridors. Modeling demand diversion to less congested routes within a corridor is part of demand modeling efforts for improved simulation and control as well as guidance-navigation systems in real time. In this paper two such diversion models were developed. The first model described the diversion at the trip origin, and the second, the diversion at freeway entrance ramps.

From a survey of approximately 1,100 commuters in the south I-35W corridor of the Twin Cities Metropolitan Area, two logit specifications were estimated. The data indicated that diversion at the origin is a function of trip time, route length, and the number of intersections along the trip. However, trip time is the dominant determining factor and can be employed to estimate the decision in the absence of additional information. Diversion at freeway entrance ramps depends on the perceived trip time on the freeway and arterial and the perceived waiting time at the ramp queue. Further, the data confirmed that socioeconomic indicators do not play a role in the diversion decision. It was also determined that, for commuter trips shorter than one hour, freeway drivers consider only one diversion alternative, that is, a preferred arterial, and do not divert to downstream ramps.

Although the models were based on data collected from only three freeway ramps in a specific metropolitan area and have not yet been transferred to other areas, it is expected that, for trips of a similar nature, the behavioral principles underlying the models generally would be applicable to other areas as well. Ongoing work seeks to validate the models and further extend them to make them operational in a real-time environment in conjunction with demand predictors under development. The purpose of developing these models is for dynamic simulation, on-line freeway corridor control, and demand forecasting suitable for guidance and navigation.

ACKNOWLEDGMENT

This work was partly supported by the National Science Foundation and the Center for Transportation Studies of the University of Minnesota.

REFERENCES

1. *Demand Responsive Strategies for Interconnected Ramp Control Systems.* FHWA, U.S. Department of Transportation, 1984.
2. S. Yagar. The Future Freeway Related Traffic Control Systems. Issues in Control of Urban Traffic Systems. *Proc. Found. Conf.*, 1981, pp. 215–222.
3. AASHTO. *Briefs of Research Problem Statements Considered by the AASHTO Select Committee on Research for the FY 1985 Program for the NCHRP.* Washington, D.C., 1983.
4. N. Gartner and R. Reiss. Congestion control in freeway corridors: The IMIS system. In *Flow Control of Congested Networks.* NATO ASI Series, Vol. F38 (A. Odoni et al., eds.), Springer-Verlag, Heidelberg, 1987, pp. 113–132.
5. R. Hall. Traveler Route Choice: Travel Time Implications of Improved Information and Adaptive Decisions. *Transportation Research* Vol. 17A, No. 3, 1983, pp. 201–214.
6. P. Michalopoulos, G. Stephanopoulos, and G. Stephanopoulos. An Application of Shock Wave Theory to Traffic Signal Control. *Transportation Research*, 1980.
7. Y. Stephanedes, P. Michalopoulos, and R. Plum. Improved Estimation of Traffic Flow for Real-Time Control. *Transportation Research Record 795*, TRB, National Research Council, Washington, D.C., 1981, pp. 28–39.
8. W. Taylor. Optimization of Traffic Flow Splits. *Highway Research Record 230*, HRB, National Research Council, Washington, D.C., 1968, pp. 60–77.
9. E. Lieberman. Simulation of Corridor Traffic: The SCOT Model. *Highway Research Record 409*, 1972, pp. 34–45.
10. M. Orthlieb and A. May. *Freeway Operation Study, Phase IV. Report 74-75: Freeway Corridor Control Strategies.* Institute of Transportation and Traffic Engineering, University of California, Berkeley, 1975.
11. D. Roden, W. Okitsu, and A. May. *FREQ7PE—A Freeway Corridor Simulation Model.* Institute of Transportation Studies, University of California, Berkeley, 1980.
12. S. Yagar. CORQ—A Model for Predicting Flows and Queues in a Road Corridor. *Transportation Research Record 533*, TRB, National Research Council, Washington, D.C., 1975, pp. 77–87.
13. B. Allen, S. Easa, and E. Case. Application of Freeway-Corridor Assignment and Control Model. *Transportation Research Record 682*, TRB, National Research Council, Washington, D.C., 1978, pp. 76–84.
14. M. Van Aerde, J. Voss, A. Ugge, and E. Case. Integration-1: A Model for Evaluating Integrated Traffic Networks. *3rd Canadian Seminar on Systems Theory for the Civil Engineer*, Montreal, Quebec, Canada, 1988.
15. E. Lieberman and B. Andrews. TRAFLO: A New Tool to Evaluate Transportation System Management Strategies. *Transportation Research Record 772*, TRB, National Research Council, Washington, D.C., 1978, pp. 9–15.
16. M. Hall, D. Van Vliet, and L. Willumsen. SATURN—A Simulation-Assignment Model for the Evaluation of Traffic Management Schemes. *Traffic Engineering Control*, 1980, pp. 168–176.
17. D. Van Vliet. SATURN—A Modern Assignment Model. *Traffic Engineering Control*, 1982, pp. 578–581.
18. I. Okutani and Y. Stephanedes. Dynamic Prediction of Traffic Volume Through Kalman Filtering Theory. *Transportation Research* Vol. 18B, No. 1, 1984, pp. 1–11.
19. S. Abu-Eisheh and F. Mannering. Discrete/Continuous Analysis of Commuters' Route and Departure Time Choices. *Transportation Research Record 1138*, TRB, National Research Council, Washington, D.C., 1987, pp. 27–34.

Convergence Properties of Some Iterative Traffic Assignment Algorithms

ALAN J. HOROWITZ

This paper examines the convergence properties of four popular traffic assignment algorithms: Frank-Wolfe decomposition for fixed-demand equilibrium assignment, an ad hoc variation of the Evans algorithm for elastic-demand equilibrium assignment, fixed-demand incremental assignment, and elastic-demand incremental assignment. The algorithms were evaluated according to errors associated with insufficient iterations, arbitrary selection of starting point, inexact theory, and small variations in data. Each of the four algorithms reached its intended solution, but did so very slowly. Elastic-demand incremental assignment emerged as the preferred technique, principally because of its more accurate response to small variations in data and its adaptability to various models of travel demand.

The most popular traffic assignment algorithms may be thought of as logical extensions to traditional iterative capacity restraint. That is, the algorithms consist of a series of all-or-nothing assignments interspersed with computations to improve estimates of link impedances and, perhaps, link volume. Some of these algorithms, such as the Frank-Wolfe decomposition method for fixed-demand assignment (*1*) or Evans's method for elastic-demand assignment (*2*), have a strong theoretical basis. Other algorithms are ad hoc. In spite of the large body of theoretical work on traffic assignment, transportation planners have had little guidance about the algorithm that yields the best performance within the usual limits on resources. In addition, there is little accurate information on how to employ an algorithm most effectively once a choice has been made. Many common rules-of-thumb are seriously misleading.

REVIEW OF THE ALGORITHMS

The purpose of this paper is to reevaluate a few existing algorithms rather than to break new theoretical ground. The following are brief descriptions of the algorithms considered:

- Iterative capacity restraint: Iterative capacity restraint is still popular, despite its terrible convergence characteristics. This algorithm is included in this comparison because it has aptly served as a "straw man" in studies by other researchers.
- Equilibrium: This fixed-demand, equilibrium assignment technique, available in most major planning packages, is an implementation of Frank-Wolfe decomposition.
- Modified Evans: Modified Evans is an ad hoc variation of the Frank-Wolfe decomposition algorithm that recalculates demand at each iteration. It resembles the Evans algorithm in both purpose and performance.
- Fixed-demand incremental: Incremental traffic assignment loads a fraction of the trip table at each iteration using all-or-nothing assignment. This technique can be implemented as a slight variation of equilibrium assignment.
- Elastic-demand incremental: At each iteration, the trip table is recalculated and a portion of it is loaded to the network. This algorithm can be implemented as a slight variation of the modified Evans algorithm.

Occasional reference will be made to Evans's precise algorithm for elastic-demand equilibrium assignment. Although the Evans algorithm is not explicitly evaluated, it is possible to determine the extent to which the other algorithms differ from the results of a true elastic-demand equilibrium assignment—the intended product of the Evans algorithm. The Evans algorithm was dropped from consideration because of its comparatively large computational requirements on multipurpose networks.

The algorithms were tested on two networks. The first was the five-zone UTOWN network, developed for testing the equilibrium assignment in the Urban Transportation Planning System (UTPS). The second was the a.m. peak-hour network for East Brunswick, N.J. The East Brunswick network contained 129 zones, and it gives a good indication of how algorithms would perform in actual practice. Both networks had five trip purposes. The tests were performed with an experimental version of QRS II running on a Zenith Z-248 (IBM PC-AT compatible).

Convergence Error

Generally, assignment error is the difference between assigned and actual volumes. Unfortunately, we can never measure actual volume with sufficient accuracy to use it as a criterion in evaluating the differences between assignment algorithms because the results are far too similar.

Total error is quite large. Various studies have shown (*3*, *4*) that root mean square (RMS) errors can regularly exceed 50 percent. Established guidelines for error (*5*) take into consideration the better performance on high-volume links, but a 20 percent error is still considered acceptable.

Convergence error, a component of total error, can be measured by comparing the results of two assignments, assuming that one of the assignments is essentially perfect. For

Center for Urban Transportation Studies, University of Wisconsin—Milwaukee, P.O. Box 784, Milwaukee, Wis. 53201.

example, a network can be run through a huge number of iterations of equilibrium assignment to obtain a nearly perfect solution to the fixed-demand problem. This solution becomes a standard for comparison. Because the primary purpose of an assignment algorithm is to forecast volumes on links, it makes sense to measure convergence error as the RMS difference in volume between the test algorithm and the standard algorithm. The RMS difference is analogous to standard error and is in units of vehicles, so it is easily interpreted.

Other researchers have attempted to measure convergence error by monitoring the objective function of the equilibrium assignment algorithm:

$$U = \sum_{\substack{\text{all} \\ \text{links} \\ i}} \int_0^{V_i} t_i(v)\, dv \qquad (1)$$

where $t_i\ (v)$ is the functional relationship between travel time and volume on link i, and V_i is the assigned volume. Since the equilibrium solution is achieved when U is minimized, an experienced individual can roughly judge the progress of an algorithm by comparing U at successive iterations. However, this objective function is deceptive. Surprisingly large changes in volume can be associated with very small changes in U. It is known (at least for the fixed-demand problem) that smaller values of U are better, but it is difficult to determine how much better or how fast the solution is improving.

A related criticism applies to monitoring the RMS change in volume between successive iterations [see paper by Sheffi and Powell (*6*) for an example]. The algorithms, as a group, converge slowly. It is not possible to determine the ultimate amount of change in volume by the change from a single iteration.

A given level of convergence error can be either important or unimportant, depending on the purpose of the forecast. To understand the role of convergence error in forecast validity, it is first necessary to list various forms it can take.

1. Insufficient iterations: Solutions generally improve at each new iteration. There can be significant convergence error associated with terminating an algorithm prematurely.
2. Resolution: An algorithm should be able to reach the same solution to a given problem each time that it is run. Since an algorithm is trying to replicate real-world processes, we would also expect it to produce similar solutions to similar problems. If it cannot do this, the algorithm is flawed.
3. Starting point: An algorithm should arrive at the same solution regardless of how it is started. Practically speaking, the solutions produced by all algorithms are affected by the choice of starting point. Insensitivity of an algorithm to its starting point is an important characteristic.
4. Ad hoc algorithm: An ad hoc algorithm could fail to converge or it could converge to a solution that is inconsistent with assignment theory. The justifications for choosing an ad hoc algorithm are potentially less error due to insufficient iterations and potentially better resolution.

It is important to keep these errors in perspective. Assignment algorithms are highly imperfect models of travel behavior. Much more significant errors stem from our poor understanding of route choice behavior, limited knowledge of impedance functions, problems in collecting demographic and network data, and our inability to show the network as it actually exists. Imperfections in theory and data are much more serious than imperfections in algorithms to implement the theory.

Test Conditions

To the best of their ability, tests were representative of planning practice. Neither the UTOWN nor East Brunswick network was modified in any way. With the exceptions of the assignment algorithm and the number of iterations of the trip distribution model, all parameters were set to the defaults for QRS II.

A doubly constrained entropy-maximizing model was used for trip distribution. For the UTOWN network, the attraction-end constraints were satisfied by 10 iterations of the trip distribution model. Trip distribution on the East Brunswick network was iterated only three times.

A Fibonacci search was used to find the averaging weights for the equilibrium and modified Evans algorithms, which minimize U. The Fibonacci search was permitted to run for 21 iterations, assuring four significant digits in the weights.

Only links that would normally carry traffic were compared for error. Centroid connectors and other artificial network elements were ignored. Also ignored were links that received no volume in any of the assignments.

Relationship Between Equilibrium and Incremental Assignments

Each iteration of equilibrium assignment consists of (a) an all-or-nothing assignment, (b) an averaging of volumes, and (c) a recalculation of link travel times given the averaged volumes. The averaging step consists of finding a weighted average between the all-or-nothing assignment and the results of the previous iteration such that U is minimized. Each iteration has a different weight, and it is impossible to know ahead of time what those weights will be.

It is easy to give the algorithm a predetermined series of weights. Although it will not necessarily converge to the equilibrium solution, the algorithm runs faster, behaves more predictably, and is easier to explain to those outside the field. One particular sequence of weights yields an incremental assignment:

$$W = 1/(i + 1) \qquad (2)$$

where W is the weight given to the all-or-nothing assignment that is calculated at iteration i. Regardless of the number of iterations, each all-or-nothing assignment (including the one from the 0th iteration that starts the algorithm) is weighted equally in the final average. Running the equilibrium algorithm with this particular fixed series of weights is a form of incremental assignment, in which the link travel times for the next increment are calculated from extrapolations of the partial volumes that have already been assigned (*7*).

Incremental assignment, as described in this paper, is a case of the method of successive averages (MSA) (*6, 8*). As a group, algorithms based on MSA are not as precise as purer optimization methods but have a greater range of applicability.

The close relationship between equilibrium and incremental assignment suggests that their solutions would be similar. It

TABLE 1 COMPARISON OF ITERATIVE CAPACITY RESTRAINT WITH MODIFIED EVANS ALGORITHM (UTOWN NETWORK)

Iteration	% RMS Difference in Link Volumes	% of Optimal Objective Function
0	84	413
1	156	407
2	134	559
3	104	1110
4	133	540
5	119	631
10	135	561
20	144	777

Modified Evans' algorithm was run for 200 iterations.

is expected that equilibrium assignment would converge faster when measured by iterations, but equilibrium assignment might well be slower when measured by total computer time.

Straw Man: Iterative Capacity Restraint

It is important to understand that an ad hoc algorithm can be either good or bad, depending on its design. A popular ad hoc algorithm is iterative capacity restraint. As implemented in QRS II, each iteration consists of (a) calculation of a trip table with travel times from the previous iteration, (b) an all-or-nothing assignment, and (c) a recalculation of link travel times. Thus, the algorithm can be considered "elastic demand"; it attempts to find a trip table that is consistent with link loads. Travel times are recalculated with the Bureau of Public Roads (BPR) speed-volume function. To provide some stability to the algorithm, link travel times were damped. That is, link travel times were taken as a weighted average of the results of the BPR function (25 percent) and the link travel times from the previous iteration (75 percent).

The UTOWN network was run for the 7:00 to 8:00 a.m. peak hour through 20 iterations of iterative capacity restraint. These results were compared with those of the modified Evans algorithm. The modified Evans algorithm is also ad hoc, but (as will be seen later) converges nicely. The comparison volumes were taken from the 200th iteration.

As expected, Table 1 shows that iterative capacity restraint performs poorly. Link volumes oscillate wildly. RMS error never becomes better than 84 percent; the value of the Frank-Wolfe objective function, *U*, never falls below its starting value.

The weaknesses of iterative capacity restraint are well documented, so these results are not totally unexpected. The especially poor performance seen in Table 1 illustrates that the UTOWN network can be hostile to ad hoc algorithms.

Ad Hoc Error of the Modified Evans Algorithm

Evans's algorithm correctly solves an elastic-demand assignment problem. It produces a solution consisting of (a) link volumes that are consistent with both link travel times and the trip table and (b) a trip table that is consistent with path travel times. In practice, the Evans algorithm looks like a variation of equilibrium assignment. Each iteration consists of computation of a trip table, an all-or-nothing assignment, an averaging step, and a recalculation of link travel time from the averaged volumes. The major obstacle to implementation of Evans's algorithm is the objective function of its averaging step. It requires far more computation and memory than regular equilibrium assignment, especially on large, multipurpose networks.

The elastic-demand equilibrium algorithm in QRS II replaces Evans's objective function with Equation 1. Consequently, QRS II is ensured of converging to a slightly wrong solution. It is possible to estimate the size of the error by the following procedure.

1. Run the modified Evans algorithm through enough iterations that link volumes are no longer changing. The selected number of iterations for the UTOWN network was 1,000. The assignment for the East Brunswick network was terminated at 100 iterations.
2. Save the trip table at the final iteration.
3. Run a fixed-demand equilibrium assignment for the same large number of iterations on this same network using the saved trip table.
4. Compare the volumes from the two assignments.

To control computation errors in the trip table, the trip distribution model was iterated 20 times (for each assignment iteration) on the UTOWN network and 10 times (for each assignment iteration) on the East Brunswick network.

The comparison is not a tautology. The modified Evans algorithm does not converge to the exact solution because the averaging weights disregard information about trip distribution. As the algorithm progresses, an inconsistency develops between the averaged volumes and the trip table, which is recomputed at each iteration. If this inconsistency is small, then final path travel times and, thus, the final iteration trip table are at the equilibrium solution. However, the final assigned volumes partially come from trip tables that were not at the equilibrium solution. The inconsistency can be measured by locking the trip table at its known equilibrium solution and running an exact, fixed-demand equilibrium assignment.

With UTOWN the link volumes differed (RMS) by 1.1 percent. With East Brunswick, the link volumes differed by

TABLE 2 PERCENT RMS ERROR FROM INSUFFICIENT ITERATIONS (UTOWN NETWORK)

Iteration	Modified Evans'	Equilibrium	Fixed-Demand Incremental	Elastic-Demand Incremental
1	64.4	46.8	70.4	84.9
2	55.3	35.6	48.0	53.6
3	34.2	32.5	39.7	41.3
4	28.1	24.9	29.5	34.3
5	25.1	21.8	25.2	27.6
10	15.9	14.8	15.6	18.7
20	9.1	9.3	10.7	11.3
50	3.7	4.0	4.5	4.3
100	1.4	1.3	2.0	1.5

TABLE 3 EQUILIBRIUM OBJECTIVE FUNCTION BY ITERATION (UTOWN NETWORK)

Iteration	Modified Evans'	Equilibrium	Fixed-Demand Incremental	Elastic-Demand Incremental
1	21.533	14.466	178.018	208.939
2	13.398	11.291	32.121	36.710
3	11.188	10.553	15.356	16.801
4	10.679	10.147	11.732	12.947
5	10.484	9.985	10.652	11.226
10	9.940	9.745	9.817	9.965
20	9.708	9.616	9.632	9.710
50	9.578	9.525	9.505	9.563
100	9.546	9.485	9.466	9.535
200	9.532	9.464	9.447	9.525

Units are 100,000 vehicle-minutes. Fixed-demand trip tables were taken from the 20th iteration of modified Evans'.

0.4 percent. Some of this error may be due to rounding. The small differences in assigned volumes indicate that the ad hoc error of the modified Evans algorithm, when used with a doubly constrained trip distribution model, is unimportant.

These comparisons were repeated using elastic-demand incremental assignment. For the UTOWN network the RMS difference in link volumes was 0.8 percent. When the East Brunswick network was subjected to the same comparison, the RMS difference in link volumes was 0.7 percent. The ad hoc error of elastic-demand incremental assignment is similar to that of the modified Evans algorithm.

Convergence Rates of Iterative Algorithms

An important attribute of an algorithm is its speed of convergence—often measured as the number of iterations necessary to reach a convergence criterion. Convergence speed was tested on four algorithms: equilibrium, modified Evans, fixed-demand incremental, and elastic-demand incremental. The first tests concerned performance on the UTOWN network. The volumes from various iterations of each algorithm were compared with volumes from 200 iterations of the same algorithm. The RMS differences are summarized in Table 2.

The convergence rates of all the algorithms were remarkably slow. Regardless of the algorithm, it took approximately 20 iterations before the convergence error fell below 10 percent. A convergence error of less than 5 percent required nearly 50 iterations. Interestingly, the variable-weight algorithms (equilibrium or modified Evans) did not significantly outperform either incremental assignment algorithm. Some of the slow convergence can be attributed to the hostility of the UTOWN network.

Table 3 gives the values of the equilibrium objective function, U, at each iteration. Note that fixed-demand and elastic-demand assignments approach slightly different values of the objective function, as expected. Table 3 illustrates the deceptive nature of the objective function. By the fifth iteration, U is changing only by about 1 percent per iteration, but the link volumes are nowhere near their equilibrium values.

The incremental algorithms did surprisingly well; after 20 iterations their objective functions were lower than their variable-weight counterparts (equilibrium and modified Evans). Furthermore, the incremental assignments required considerably less time to reach the same number of iterations. For example, 10 iterations of the modified Evans algorithm took 406 sec of elapsed time; 10 iterations of elastic-demand incremental assignment took just 225 sec.

Similar tests were performed on the East Brunswick network. The comparison assignments were obtained from the 50th iteration of each algorithm. These results are shown in Table 4. Convergence rates, as measured by percent RMS difference in link volumes, were twice as fast as with the UTOWN network. Nonetheless, it took approximately 10 iterations to achieve a 10 percent error. Usually a 10 percent computational error is considered unacceptable.

TABLE 4 PERCENT RMS ERROR FROM INSUFFICIENT ITERATIONS (EAST BRUNSWICK NETWORK)

Iteration	Modified Evans'	Equilibrium	Fixed-Demand .Incremental	Elastic-Demand Incremental
1	34.8	36.3	34.8	33.6
2	22.9	23.2	29.3	32.4
3	16.6	17.4	25.4	22.4
4	15.7	15.4	17.5	16.2
5	13.1	11.9	13.7	13.1
10	8.7	7.6	7.1	6.4

TABLE 5 EQUILIBRIUM OBJECTIVE FUNCTION BY ITERATION FOR FIXED-DEMAND ASSIGNMENTS (EAST BRUNSWICK NETWORK)

Iteration	Equilibrium	Fixed-Demand Incremental
1	2.373	3.013
2	2.448	2.496
3	2.359	2.465
4	2.286	2.325
5	2.247	2.282
10	2.161	2.178
50	2.057	2.078

Units are 100,000 vehicle-minutes.

This research did not evaluate methods of accelerating equilibrium assignment (*9*), so these tests may somewhat understate its potential. Similar acceleration techniques would also apply to the original Evans algorithm; however, one would guess that the amount of acceleration is insufficient to overcome the algorithm's large computational requirements on meaningfully complex networks.

Ad Hoc Error of Incremental Assignment

The previous results show that the two incremental algorithms run at about the same rate (measured by iterations) as equilibrium assignment. As a further comparison, the East Brunswick network was run with fixed-demand incremental assignment for a total of 50 iterations. These results were compared with 50 iterations of equilibrium. The RMS difference in link volumes was 1.7 percent. Table 5 shows that the values of U for the two assignments were also close after the third iteration.

A similar comparison has already been seen in Table 3. The last line shows that at 200 iterations on the UTOWN network, incremental assignment actually outperformed equilibrium assignment. Incremental assignment was slightly closer to the equilibrium solution. The RMS difference in link volumes was 1.0 percent. The superior performance of incremental assignment on this network should be considered unusual.

Resolution Error

In many planning situations, a serious concern is the ability of an algorithm to produce similar results from similar networks. For example, a small change in a single zone's trip production should have just a small effect on volume. Table 6 shows the behavior of the several assignment algorithms when 1,000 dwelling units are added to a single zone of the UTOWN network. Each line in the table compares the volumes obtained from the base network with the volumes from the modified network when run through the same number of iterations of the same algorithm.

The first line in Table 6 should be considered the correct answer. It compares the two networks after 200 iterations of the modified Evans algorithm. It is seen that the addition of 1,000 dwelling units causes a 4.2 percent RMS change in assigned volumes.

The other algorithms, if they are working properly, should always show a smaller RMS change than all-or-nothing assignment. The other algorithms are inherently multipath, so the additional trips are split among a greater number of links. As expected, the comparison using all-or-nothing assignment (line 4) is larger than that obtained with 200 iterations of the modified Evans algorithm.

The remaining lines in Table 6 show that the other algorithms are not working properly. They all overestimate the amount of change. The most accurate was the modified Evans algorithm at 20 iterations (overestimating the change by 1.3 percent of average link volume); the least accurate was elastic-demand incremental at 10 iterations (overestimating the change by 6.9 percent of average link volume).

Given these disturbing results, a more elaborate series of tests was run on the East Brunswick network; the results are summarized in Table 7. As with the tests of the UTOWN network, each cell in the table represents a comparison of two slightly different networks, which were run on exactly the same algorithm. Each pair of networks differed by the addition of 84 dwelling units to a single zone of one network. Five separate zones were arbitrarily chosen for investigation. The iterative assignment algorithms were run for just 10 iterations.

The RMS difference using all-or-nothing assignment gives a slight overestimate of the expected change. At most, the addition of 84 dwelling units to Zone C resulted in an (RMS) impact of 2.6 percent. Three of the five zones had impacts of less than 1 percent.

All of the iterative assignment techniques estimated the impact badly. For example, we know from the all-or-nothing assignments that the correct impact for Zone A is less than 0.5 percent. However, the iterative assignment algorithms yielded impacts between 1.6 and 7.2 percent. The elastic-demand incremental algorithm behaved best for every zone.

It appears that resolution error is largely a consequence of error due to insufficient iterations. This convergence error has both random and systematic components. The systematic

TABLE 6 PERCENT RMS DIFFERENCE IN VOLUME AFTER 1,000-DWELLING UNIT INCREASE IN ONE ZONE (UTOWN NETWORK)

Algorithm	Iterations	Percent RMS Difference
A. Modified Evans'	200	4.2
B. Modified Evans'	10	7.9
C. Modified Evans'	20	5.5
D. All-or-Nothing	0	6.1
E. Equilibrium	10	6.5
F. Equilibrium	20	7.5
G. Elastic-Demand Incremental	10	11.1
H. Elastic-Demand Incremental	20	7.6

TABLE 7 PERCENT RMS DIFFERENCE IN VOLUME AFTER 84-DWELLING UNIT INCREASE IN SINGLE ZONE (EAST BRUNSWICK NETWORK)

Zone	All-or-Nothing	Modified Evans'	Equilibrium	Elastic-Demand Incremental
A	0.5	7.2	2.0	1.6
B	0.6	7.0	2.6	2.4
C	2.6	7.1	3.3	2.1
D	2.5	7.9	4.3	3.0
E	0.7	9.4	2.2	1.6

TABLE 8 PERCENT RMS DIFFERENCE IN VOLUME FROM VARIOUS STARTING POINTS (UTOWN NETWORK)

Total Iterations	Modified Evans'	Equilibrium	Elastic-Demand Incremental
10	10.6	15.7	10.8
20	6.3	12.1	6.2

component vanishes in the comparison; the random component does not. As seen here, large amounts of random error can mask the actual impact. Comparing the errors in Table 4 with those in Table 7 shows that the convergence error in the modified Evans algorithm is almost entirely random, whereas the convergence error in elastic-demand incremental assignment has a large systematic component.

The distinction between random convergence error and systematic convergence error is critical to the selection of an assignment algorithm. The nature of transportation planning is to compare alternatives. During such comparisons the only important errors are random. Random convergence error can be attenuated only by running additional iterations.

Starting Point Error

All iterative assignment algorithms require an initial estimate of link travel times. In practice, the results of assignment algorithms depend on this estimate.

Table 8 shows the effect of the starting point on the UTOWN network. Each cell in Table 8 compares two assignments for the identical network on an identical algorithm. The two assignments differ only by the method of estimating the initial link travel times. One assignment uses free travel time; the other assignment uses travel times estimated from volumes resulting from an all-or-nothing assignment.

Starting point errors are almost as large as errors due to insufficient iterations. Interestingly, the two ad hoc algorithms (modified Evans and elastic-demand incremental) were shown to be far less sensitive to the starting point than equilibrium assignment.

There exists a rule of thumb that a good initial estimate of link travel times will produce a better assignment than an inaccurate initial guess. Although partially correct, this rule of thumb is not very helpful. Table 9 shows the effect of an optimal set of initial travel times on the objective function (U) of the modified Evans algorithm. The optimal link travel times were taken from the 200th iteration of the same algorithm. A comparison of Table 9 with the first column of Table 3 shows that optimal link travel times were essentially useless. Any early advantage was erased by the 20th iteration. Similar results were obtained with the other algorithms.

CONCLUSIONS

All algorithms tested, with the exception of iterative capacity restraint, are derived from Frank-Wolfe decomposition. For practical purposes, they all converge to their intended solu-

TABLE 9 EQUILIBRIUM OBJECTIVE FUNCTION FOR OPTIMAL STARTING POINT OF MODIFIED EVANS ALGORITHM (UTOWN NETWORK)

Iteration	Optimal Start
1	11.582
2	10.610
3	10.314
4	10.188
5	10.082
10	9.911
20	9.736

Units are 100,000 vehicle-minutes.

tions at about the same rate, as measured by iterations. However, this convergence rate is unexpectedly slow. An unacceptable 10 percent convergence error remains after 20 iterations on the UTOWN network and after 10 iterations on the East Brunswick network. A more reasonable error of 5 percent is reached after about 50 iterations on the UTOWN network. Given these slow convergence rates, it is more appropriate to refer to "near-equilibrium" solutions, that is, solutions within some acceptable error limit.

The most disturbing aspect of convergence error is its random component. Even a small amount of random error can completely invalidate comparisons of close alternatives; the only proven method of reducing random error is to run more iterations. Incremental assignment algorithms appear to have much smaller random components in their convergence errors, suggesting that fewer iterations are required.

The existence of convergence error should force planners to adopt innovative methods of assignment. For example, it is sometimes possible to forecast only the increment of traffic due to site development. Such a forecast will have more validity if the random error can be confined to the increment, while treating any errors in background volumes as entirely systematic.

Ad hoc algorithms are not necessarily bad. It is possible for an ad hoc algorithm to greatly outperform a rigorously derived algorithm, given the same computer budget. Because ad hoc algorithms do not come with a pedigree, confidence in an ad hoc algorithm must be established through extensive testing.

If the results of a simulation are to be readily accepted, its algorithms must be lucid. Given the choice, planners should pick an assignment algorithm that can be easily explained to decision makers. The elastic-demand incremental algorithm is conceptually simple; Evans's algorithm is conceptually complex. Both algorithms produce essentially the same answer.

The existence of several algorithms that can consistently produce near-equilibrium solutions to a given traffic model should enhance prospects of improving the model. Model developers should concentrate on incorporating better traffic theory and not be overly concerned with finding an algorithm that delivers the intended solution. The algorithm appearing to adapt most easily to different traffic models is elastic-demand incremental assignment.

Overall, the tests indicate that elastic-demand incremental assignment produces the best solutions. The method is easy to implement, it can be quickly modified to handle a variety of demand models, and it converges reasonably well. Its speed of convergence is no worse than that of more precise algorithms; its ad hoc error is insignificant; it is relatively insensitive to the starting point; it has the best resolution among the tested algorithms; and it is easy to understand. The relative success of elastic-demand incremental assignment contributes evidence of the resiliency of incremental (or successive average) methods.

REFERENCES

1. L. Leblanc, E. Morlok, and W. Pierskella. An Efficient Approach to Solving the Road Network Equilibrium Traffic Assignment Problem. *Transportation Research*, Vol. 9, 1975, pp. 309–318.
2. S. P. Evans. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research*, Vol. 10, 1976, pp. 37–57.
3. G. R. M. Jansen and P. H. L. Bovy. The Effect of Zone Size and Network Detail on All-or-Nothing and Equilibrium Assignment Outcomes. *Traffic Engineering and Control*, Vol. 23, 1982, pp. 311–317.
4. B. R. Wildermuth, D. J. Delaney, and K. E. Thompson. Effect of Zone Size on Traffic Assignment and Trip Distribution. In *Highway Research Record 392*, HRB, National Research Council, Washington, D.C., 1972, pp. 58–75.
5. N. J. Pedersen and D. R. Samdahl. *NCHRP Report 255: Highway Traffic Data for Urbanized Area Project Planning and Design.* HRB, National Research Council, Washington, D.C., 1982.
6. Y. Sheffi and W. Powell. A Comparison of Stochastic and Deterministic Traffic Assignment over Congested Networks. *Transportation Research*, Vol. 15B, 1981, pp. 53–64.
7. C. Fisk. Some Developments in Equilibrium Traffic Assignment. *Transportation Research*, Vol. 14B, 1980, pp. 243–255.
8. W. B. Powell and Y. Sheffi. The Convergence of Equilibrium Algorithms and Predetermined Step Sizes. *Transportation Science*, Vol. 16, 1982, pp. 45–55.
9. L. J. LeBlanc, R. V. Helgason, and D. E. Boyce. Improved Efficiency of the Frank-Wolfe Algorithm for Convex Network Programs. *Transportation Science*, Vol. 19, 1985, pp. 445–462.

Dynamic Assignment in Three-Dimensional Time Space

RUDI HAMERSLAG

In traditional assignment models, cars are assigned to a route and are therefore present on all links on that route simultaneously. Calculations from this type of model give few positive results. If the assignment is done in space with time as a third dimension, this problem can be overcome. The first part of the paper gives a simple example of the equilibrium assignment model showing that, in some parts of the network, congestion is unrealistically calculated as a consequence of bottlenecks upstream. The second part of the paper gives a description of the three-dimensional assignment models. The proposed algorithm conforms with existing two-dimensional assignment models, although details in the algorithm are different. The effect of improving the capacity of bottlenecks on congestion downstream is shown. A computer model of the assignment model works under MS-DOS on a microcomputer.

In traditional assignment models two-dimensional (2-D) origin–destination (O-D) matrices are assigned to two-dimensional networks. Cars between each O-D pair are assigned to the links belonging to a certain route. Because these links do not have a time dimension, the implicit assumption is made that cars are present on all links at the same time. So cars that in reality are caught in a particular bottleneck can also be considered in the calculation the cause of congestion downstream. To improve the assignment process, a time dimension is added to the traditional 2-D assigned space. A three-dimensional (3-D) O-D matrix is assigned to a three-dimensional network.

The following are discussed in this paper:

1. The problems with the 2-D assignment,
2. The principles of dynamic assignment in 3-D time space,
3. The principles of the algorithm used,
4. The increase in capacity downstream from the bottleneck, and
5. A few remarks about computing.

TRADITIONAL ASSIGNMENT MODELS

In traditional 2-D assignment models [e.g., that of van Vliet (*1*)], networks are defined by links. These links connect two models (e.g., j and k). Each node j (1, 2, 3, . . .) and k (1, 2, 3, . . .) has coordinates $x_j;y_j$ and $x_k;y_k$. Each link has a certain length (z_{jk}) with a distance, time, or generalized time dimension. In this paper, time will be used as a dimension. The shortest routes are calculated between each O-D pair. In the all-or-nothing assignment program, all cars between each O-D pair are assigned to the shortest route.

The equilibrium method (*2,3*) can be used if there are overloaded links in a network. The time on every link jk (z_{jk}) is calculated by using a delay function:

$$z_{jk} = F(q_{jk}, C_{jk}, z_{jk0}) \tag{1}$$

where

q_{jk} = the traffic flow on link jk,
C_{jk} = the capacity of link jk,
z_{jk0} = the time of a link jk in an unloaded network, and
z_{jk} = the time of link jk in a loaded network.

See Brandston's overview (*4*).

The value of q_{jk} is calculated by an iterative process. Equilibrium will be reached when the flow on all routes in use is equal and when there are no more unused links (Wardrop's principle). To reach equilibrium, the linear approximation method can be used (*3*). The flow in iteration i (q^i_{jk}) is calculated as a linear combination of q^{i-1}_{jk} and q^+_{jk}. The value q^+_{jk} is the assigned traffic to the shortest routes in the network with $z^{i-1}_{jk} = F[q^{(i-1)}_{jk}, C_{jk}]$.

The next example was inspired by the traffic system southwest of Rotterdam where a bridge limits the traffic crossing the river. The O-D matrix in Table 1 was assigned to the network in Figure 1. The traffic flows run from right to left. Figure 1 shows an all-or-nothing assignment and an equilibrium assignment. The equilibrium model shows that part of the cars are assigned to routes 4-10-9 and 2-1-8. This assignment is made because of congestion on links 3-2 and 5-8. In reality this congestion does not appear because the cars are held in bottleneck 7-6. The equilibrium assignment model gives fewer satisfactory results in this example.

ASSIGNMENT IN TIME SPACE

The main problem of the assignment models is that traffic is assigned to a network without a time dimension. To improve these methods, a time dimension has been added. Links are defined by nodes jk and period p.

Instead of time on a link jk, the time on link jk is introduced during period p. A period capacity is used instead of an hour capacity. The traffic flows are also defined by nodes jk and period p. The routes are calculated on the surface and in space, so a 3-D time space is used.

If a link is overloaded, then the path (a) will switch to a route along other nodes as in the 2-D space, (b) will switch to a route in a later period, or (c) both.

The delay on the links is also determined in time space and

Departments of Civil Engineering and of Technical Mathematics and Informatics, Delft Technical University, Stevinweg 1, 2600 GN Delft, The Netherlands.

TABLE 1 ORIGIN-DESTINATION MATRIX

From To	4	7	sum
8	1250	3750	5000
9	1250	3750	5000
sum	2500	7500	

Flows in overloaded links

link	Flow	capacity
7 - 6	7500	4000
6 - 5	7500	4000
3 - 2	5000	4000
2 - 9	5000	4000
8 - 5	5000	4000

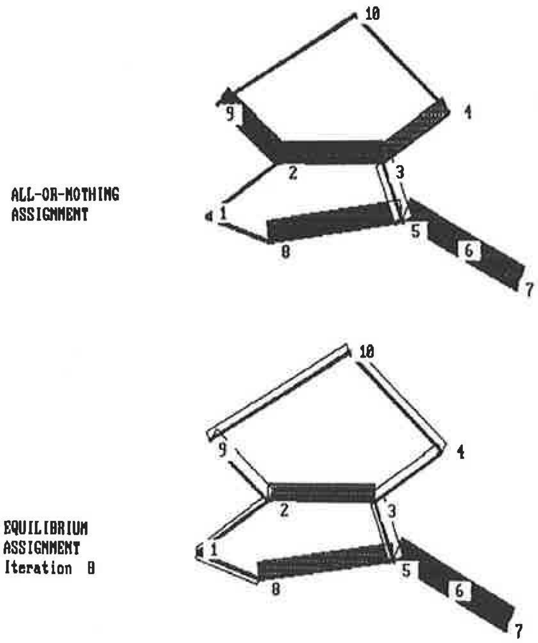

FIGURE 1 Assignment of O-D matrix in Table 1 to a network with the all-or-nothing method (*top*) and the equilibrium method (*bottom*). (In the following figures, links loaded between 85 and 95 percent of capacity are lightly shaded, and links that are loaded more than 95 percent of capacity are shaded more darkly.)

may be different from period to period. At the end a 3-D O-D matrix is assigned to a 3-D network. This method, "Dynamic Assignment in the Three-Dimensional Timespace," was first published in 1987 (*11*).

An Example

The example in Figure 1 is now represented in 3-D time space. Figure 2 gives the flows during the successive eight periods. Traffic is held in the upstream bottleneck, links 7-6 and 6-5. There is no congestion in links 3-2 and 5-8 downstream, as in 2-D space. The less logical routes of the equilibrium assignment in 2-D space do not appear in 3-D time space; consequently, the difficulties with the 2-D equilibrium assignment model are solved in 3-D space.

Other Methods

The essential differences between the present method and the well-known CONTRAM and SATURN methods are the following:

- In the SATURN method (*5,6*), trip-dependent O-D matrices are assigned to independent networks for various periods.
- In the CONTRAM assignment method (*7,8*), a limited number of cars are sequentially assigned to independent networks. Overloaded links lead to an overflow to the links in another period.

Kroes et al. (*9*) mentioned a method called "equilibrium assignment in the timespace," which was also developed in the Netherlands. In addition to the road network, a network with shadow links was made that represented the alternative of driving at other than peak-period times. With a 2-D equilibrium model, part of the traffic is assigned to this network. Although the name is similar, this method is different from the method presented in this paper.

In the method proposed by Ben-Akiva et al. (*10*), an equilibrium method is used to change the departure times and link times. The method can be used only for a very small hypothetical network. The 3-D assignment method in this paper uses 3-D O-D matrices and networks with departure times that are not affected by congestion. A study has been started to integrate our method with those of Kroes et al. (*9*) and Ben-Akiva et al. (*10*).

THE ALGORITHM

Three-dimensional assignment can be formulated as a 3-D equilibrium model. The algorithm consists of the following steps:

1. Read a 2-D network.
2. Determine the 3-D O-D matrix.
3. Determine the period capacity of the links.
4. Calculate the delay in the links.
5. Calculate the shortest routes in 3-D space.
6. Assign the 3-D O-D matrix to the shortest routes.

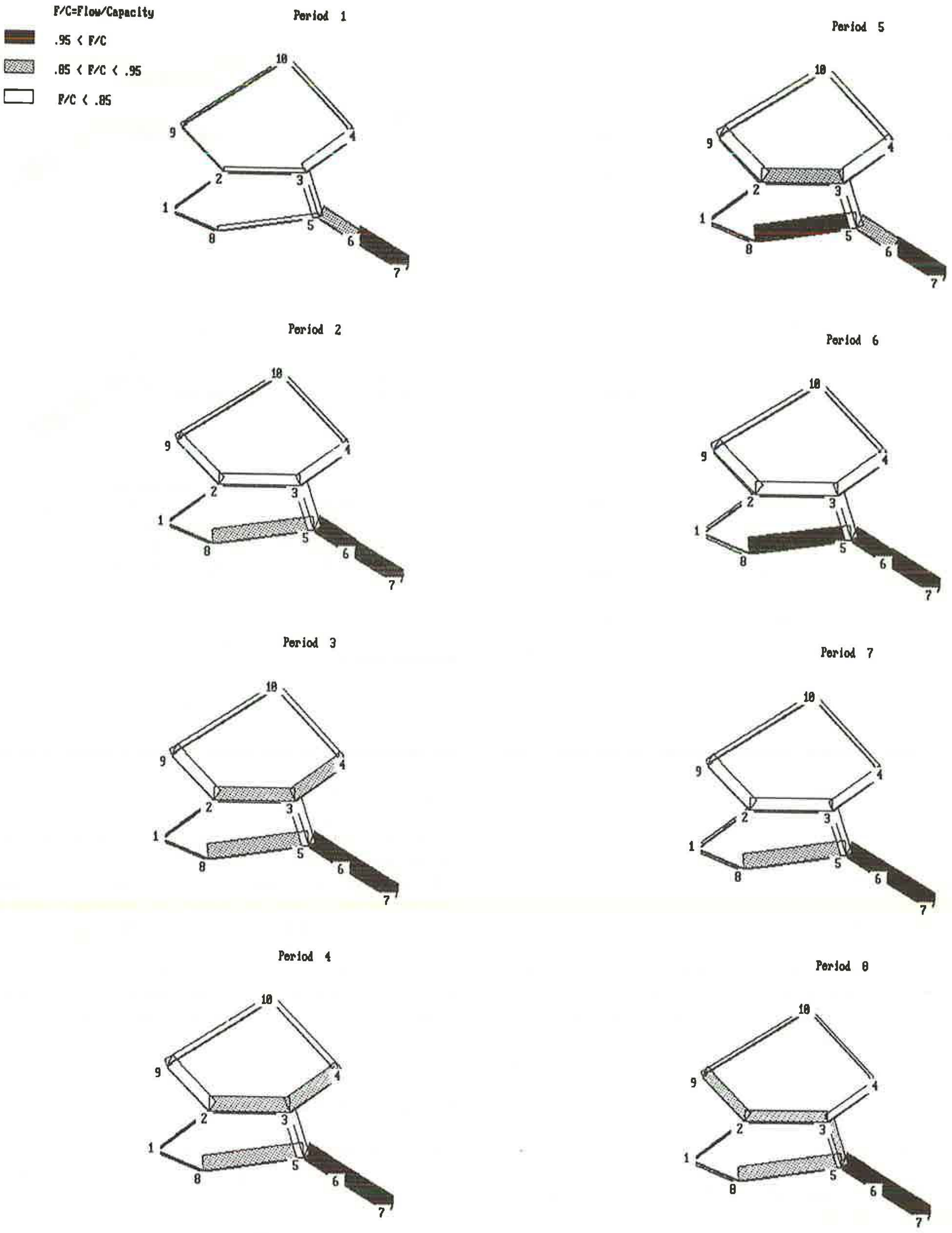

FIGURE 2 Example of an assignment in time space of the O-D matrix of Table 1.

7. Load the network.
8. If the stop criterion has not been reached, return to Step 4.

Although the 3-D algorithm, generally speaking, is similar to the algorithm in 2-D space, there are some important differences on a more detailed level:

1. Read the 2-D O-D matrix and the 2-D network. Existing 2-D networks can be used in the 3-D calculations; consequently there is no need for conversion or extra input of data. This factor is a practical advantage of the method.
2. Determine the 3-D O-D matrix. The 3-D matrix is determined by splitting up the 2-D O-D matrix into periods defined by the departure time fractions. This system is a good way to approximate the peak periods. For longer periods (e.g., holiday traffic), more complicated methods should be used to determine the 3-D matrix.
3. Determine the period capacity. The period capacity of the link can be determined as a fraction of the hourly capacity. The capacity is multiplied by the ratio of period length and 60 min. It is also possible to reduce some of the period capacities to account for delays caused by highway construction.
4. Determine the link delay. A 3-D delay function is used to determine the delay in 3-D links. This function is similar to that in 2-D space.

$$z_{jkp} = F(q_{jkp}, Q_{jkp}, C_{jkp}, z'_{jkp}) \tag{2}$$

where

q_{jkp} = the number of cars on link *kj* during period *p*,
Q_{jkp} = the number of waiting cars from previous periods,
C_{jkp} = the capacity (cars per unit) during period *p*,
z'_{jkp} = the time on the unloaded link *jk* during period *p*, and
z_{jkp} = the time on the loaded link *jk* during period *p*.

In general, z_{jkp} will have different values for the various periods. The overloaded links in *previous* periods influence the delay in the *later* periods.

5. Determine the shortest routes in 3-D space. Figure 3 shows a 3-D network with a string of links. Link 2-3 has a lower capacity than the other links. The *Y*-axis is the time scale. The links of the successive time periods are shown. The dashed lines are the 3-D paths of the first and last cars in each period. The last car in the first period is the same as the first car in the second period, and so on.

The departure time of the first car in the first period equals zero. This car uses links in period 1. The departure time of the last car in the first period equals 10. This car uses the links in the second period. Some of the cars that depart between the first and last cars are using links during the first and the second periods. The departure time of the first car of period 3 is 20 min. Because of a delay in node 2-3 caused by congestion, the car arrives in node 3 more than 10 min later. This car also uses links in different periods.

So the departure times are established in increments of 10 (0, 10, 20, 30, etc.), instead of zero as in 2-D assignment. The point at which the nodes are passed depends on delays, which may be different from period to period. In addition, the cars use links in different periods.

In 3-D space, the determination of routes from the origin is similar to that in 2-D space. It is possible to use the Moore or the Dijkstra algorithms. The difference is that this route determination is done for all the periods, rather than just one period, as in 2-D space. Another important difference is that the routes in 3-D space are found by comparing space paths and time paths simultaneously, which enables a comparison of routes between origin and destination.

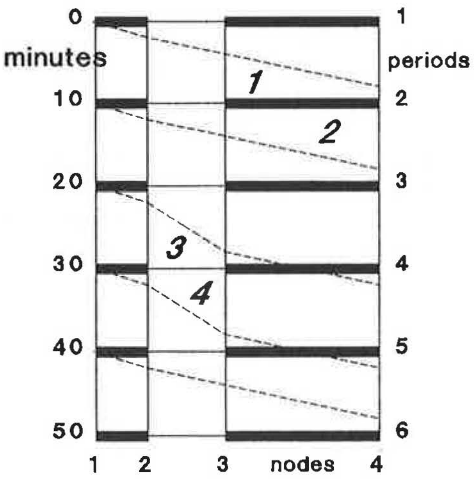

FIGURE 3 Example of the 3-D assignment in time space of a network.

6. Assign the 3-D O-D matrix to the shortest 3-D routes. The important difference is that in the 2-D space all car trips of an O-D pair are assigned to all links along the shortest route. Because in the 3-D space links are used during different periods, the cars must be assigned to different periods. The cars that depart in the first period use link 3-4 in the first and second periods. The ratio of the car trips that are assigned to link 3-4 in the first and second periods is proportionate to the areas marked *1* and *2* (Figure 3). The car trips of the third period are partly assigned to link 2-3 in the third and fourth periods. The ratio is proportionate to the areas marked *3* and *4*.
7. Load the network. Loading the network is done by part of the all-or-nothing assignment flows just calculated and with flows from the previous iteration. As in the 2-D space, it can be done in various ways.

Two methods have been tested. The first method is similar to the linear approximation method of the equilibrium method. The first experience with this method was not very successful, as was reported at the UTSG conference in London (*12*) in 1988. However, the method is being improved, so linear approximation may be useful after all. Some research is still required to make this suitable for publication.

The second algorithm uses the equation

$$q^{i}_{jkp} = q^{+}_{jkp} \cdot g^{i} + q^{i-1}_{jkp} \tag{3}$$

The value of g^i depends on the number of iterations (i) and will also be chosen in such a way that there are no overloaded links.

$$g^{i} = \min[1/(i + 1), (q^{i-1}_{jkp} - C_{jkp})/q^{+}_{jkp}] \tag{4}$$

EXAMPLE OF UPSTREAM CONGESTION

It is possible to gain insight into the problem of new congestion that appears after improvements upstream. The example in

Figure 2 is used to demonstrate the effect of increased capacity of links 7-6 and 6-5. Although the congestion on these links disappears, new congestion arises on links 2-3 and 5-8 downstream. Some traffic uses routes 3-4-10-9-2 and 2-1-8 (in period 3, 4, 6, 7) to avoid the congestion.

The calculation shows the influence of upstream bottlenecks on downstream congestion. It seems possible that downstream congestion can be prevented by delaying the traffic feeding onto links upstream. The 3-D assignment technique can give better insight into the ability of this method to improve the working of the traffic system.

COMPUTING

The system runs as part of the TFTP workbench on OLIVETTI M21 and PC-AT and PC-386 with EGA cards for small networks (*13*). The assignment of large networks is also possible. However, 3-D calculation needs more computer time than 2-D assignment. The calculation time is the product of

- The number of iterations,
- The number of time periods, and
- The time necessary for the calculation of an all-or-nothing assignment.

The calculation time necessary will be about 100 times a 2-D all-or-nothing assignment or 10 times an equilibrium assignment.

To improve the calculation speed, a special processor is being developed so the system can be used for very large networks. The first prototype of this processor is about 200 times faster than a Microvax. An even faster execution is possible (*14*). Because of these improvements it is expected that a longer calculation time will not be required for very large networks.

CONCLUDING REMARKS

Since the traditional 2-D assignment methods have some shortcomings, a time dimension has been introduced to improve this method. The algorithm, generally speaking, is similar to the 2-D variant. However, on a detailed level there are some differences that cannot be neglected: the method can be used for large networks; the existing 2-D networks can be used as input for the calculations; the calculation time is longer; and the development of computer hardware makes the method suitable for very large networks.

In conclusion, the dynamic assignment in 3-D time space can be used for the following purposes:

- A more realistic assignment of traffic on congested networks;
- Acquisition of new insights into new downstream congestion after improving capacity upstream;
- Ability to calculate, based on downstream congestion, the influence of decreases in capacity caused by such factors as road construction, road maintenance, and accidents;
- Ability to calculate the areawide effect of feeding cars into a network system on certain strategic chosen links; and
- Ability to use the program as part of a delay warning system during road congestion.

REFERENCES

1. D. van Vliet. Road Assignment. *Transportation Research*, Vol. 10, 1976, pp. 137–157.
2. L. J. Leblanc and E. K. Morlok. An Analysis and Comparison of Behavioral Assumption in Traffic Assignment. In *Proceedings of International Symposium on Equilibrium Methods*, Montreal, Quebec, Canada, 1974.
3. E. R. Ruiter. Implementation of Operational Network Equilibrium Procedures. In *Transportation Research Record 491*, TRB, National Research Council, Washington, D.C., 1974, pp. 40–51.
4. D. Brandston. Link Capacity Functions: A Review. *Transportation Research*, Vol. 10, 1976, pp. 223–236.
5. M. D. Hall, D. van Vliet, and L. G. Willumsen. SATURN—A Simulation Assignment Model for the Evaluation of Traffic Management Schemes. *Traffic Engineering and Control*, Vol. 21, 1980, pp. 168–176.
6. D. van Vliet. SATURN—A Modern Assignment Model. *Traffic Engineering and Control*, Vol. 23, Dec. 1982, pp. 578–581.
7. D. R. Leonard, J. B. Though, and P. C. Baguley. CONTRAM: A Traffic Assignment Model for Predicting Flows and Queues During Peak Periods. TRRL Laboratory Report 841. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1978.
8. D. I. Robertson, D. R. Leonard, and J. B. Though. CONTRAM, een Methode voor het Testen van Verkeerscirculatie Projecten. *Verkeerskunde*, No. 5, 1980.
9. E. P. Kroes, R. W. Antonisse, and S. Bexelius. Return to Peak. *Proc., PTRC Summer Annual Meeting*, PTRC Education and Research Services, Ltd., London, 1987.
10. M. Ben-Akiva, M. Cyna, and A. de Palma. Dynamic Models of Peak Congestion. *Transportation Research*, Vol. 18B, No. 4/5, 1984.
11. R. Hamerslag and P. C. H. Opstal. Tijddynamische Toedeling in de Tijdruimte. *Colloquium Vervoersplanologisch Speurwerk*, 1987, Delft, pp. 433–454.
12. R. Hamerslag. Dynamic Assignment in the Three-Dimensional Timespace. Presented at UTSG Annual Meeting, London, England, 1988.
13. R. Hamerslag. Teacher Friendly Transportation Programs. *Micro Computers in Civil Engineering*, Vol. 3, 1988, pp. 81–89.
14. H. J. M. van Grol. *An Algorithm Oriented Processor for Traffic Routing*. Department of Applied Physics, Subsection of Computational Physics, Delft Technical University, Delft, The Netherlands.

Balancing Link Counts at Nodes Using a Variety of Criteria: An Application in Local Area Traffic Assignment

REFAT BARBOUR AND JON D. FRICKER

A study of the impact of a major change in a campus street network began with the collection of link flows before the change. Despite the care with which the link counts were made, conservation of flow at each node was not satisfied. Therefore, the flows through the nodes had to be "balanced." This paper discusses the variety of techniques developed to balance the network. The techniques fell into two categories: algorithms and mathematical programming formulations. A comparison was made between these procedures and the maximum-likelihood method advocated in the literature. It became evident that the node-balancing solution depends on the criteria chosen to evaluate the solution, which in turn can offer guidance as to the specific method to choose or develop.

Recently, the street network in the northeast portion of the Purdue University campus underwent a major change. The main entrance to the university was permanently closed to permit construction of a new academic building. Before this change took place, link flows in this portion of campus and on the urban streets immediately adjacent to it (a study area hereafter referred to as "Campus NE") were observed and recorded during the afternoon peak hour. The intent was to provide the basis for a forecast of the link flows after the network change and thereby identify potential traffic bottlenecks. Therefore, all street facilities used by vehicles in the area of interest were represented by links in the network abstraction of Campus NE. Because of this level of detail, and because the study area was less than 1 mi^2 in size, we use the phrase "local area traffic assignment" to distinguish our activity from that of the traditional city- or regionwide travel demand modeling process. In fact, our work could be considered a type of site impact analysis, although our initial emphasis was on route choice behavior, with signal timing confined to a subsequent phase of the project.

NODE-BALANCING ALGORITHMS

Despite the data collectors' best efforts to be accurate in recording the link flows and turning movements they observed, when the information was put into the link-node model of the network, it was clear that the recorded flows at most of the intersections violated the conservation of flow requirement, which is that the sum of flows in equals the sum of flows out. In other words, these intersection nodes were "unbalanced" with respect to their recorded flows.

Since the origin–destination (O-D) table estimation and traffic assignment models available to us required balanced link counts, we had to improve the traffic count data to restore conservation of flow at all nodes (*1, 2*). We found the node-balancing (NB) method presented previously (*1*) to be the principal alternative to a manual trial-and-error adjustment of link flows. This method assumes that observed flows are Poisson distributed and then employs a maximum-likelihood method (MLM) to find the most likely set of link flows from the many possible solutions. Although we accepted the idea behind the MLM, we believed that if we were going to write any computer code, we would be more comfortable trying to apply some familiar network algorithms to adjust the unbalanced link flows than trying to convert the ideas described elsewhere (*1*) into FORTRAN. We were also curious about the impact of various objectives or solution criteria on the solution itself. In the next section we report on the evolution of the NB algorithms we developed. In later sections, we present a set of optimization procedures and some comparative evaluations.

Method NB1: Automated Trial and Error

Method NB1 automates a form of the trial-and-error method that we might have used manually. It was encoded to provide a basis of comparison against what we anticipated would be more sophisticated methods. In the steps below, V(in) and V(out) are the inflow and outflow rates at an unbalanced node u, and $I(u)$ is the amount of the imbalance at a node u, $I(u) = V(\text{in}) - V(\text{out})$. At any iteration $k > 1$, node u is considered to be approximately balanced if the absolute value of $I(u)$ is either (a) less than or equal to 1 or (b) within 1 percent of $0.5 * [V(\text{in}) + V(\text{out})]$.

Method NB1

Step 0. $k = 0$.

Step 1. $k = k + 1$. Identify all nodes j in the network that are not origin or destination centroids but have unbalanced flows and place them in the set of unbalanced nodes (U) in

R. Barbour, Strand Associates, Inc., 910 West Wingra Drive, Madison, Wis. 53715. J. D. Fricker, School of Civil Engineering, Purdue University, West Lafayette, Ind. 47907.

order of their original node numbers. If all nodes are balanced, go to Step 4.

Step 2. If set U is empty, go to Step 1. Otherwise remove the first node in set U and call it the "u-node."

Step 3. (a) If $I(u) > 0$, decrease each inflow by $p(i,u) * 0.5 * I(u)$, where $p(i,u)$ is the proportion of all inflows that enter node u via link (i,u). For example, if $I(14) = +36$, and node 14 receives 25 percent of its inflow from link (4,14), then $V(4,14)$ will be reduced by $0.25 * 0.5 * 36 = 4.5$ vehicles. Likewise, each outflow will be increased by $p(u,i) * 0.5 * I(u)$, where $p(u,i)$ is the proportion of outflows that depart node u via link (u,i). (b) If $I(u) < 0$, add $-p(i,u) * 0.5 * I(u)$ to each inflow link (i,u) and subtract $-p(u,i) * 0.5 * I(u)$ from each outflow link (u,i). [Note: the minus sign before "p" is necessary because $I(u) < 0$.] (c) If any of these flow adjustments would cause a link flow to become negative, leave that link's flow unchanged and redefine $p(u,i)$ or $p(i,u)$ among the remaining links involved. (d) Go to Step 2.

Step 4. (a) Identify those noncentroid nodes u that are approximately (but not exactly) balanced. (b) If $I(u) > 0$, find the centroid Z nearest u. If Z is an origin centroid, subtract $I(u)$ from each link on the shortest path between Z and u. If Z is a destination centroid, add $I(u)$ to each link between u and Z. (c) If $I(u) < 0$, find the centroid Z nearest u. If Z is an origin centroid, add $-I(u)$ to each link between Z and u. If Z is a destination centroid, subtract $-I(u)$ from each link between u and Z. (d) Stop.

Discussion of Method NB1

Method NB1 is admittedly crude, but it is fairly easy to program. The relaxed definition of "balanced" in Step 1 after iteration $k = 1$ is a recognition that exact balance for all nodes may never result from this method. Our experience indicates that, after about 20 iterations, all nodes are in approximate balance and further iterations are of little value. Therefore, Step 4—a housecleaning step—is used to avoid the creation of minicentroids by making very minor changes to the $T(i)$ and $T(j)$ values we have collected at the parking facilities and the study area boundaries.

Method NB2: Minimum-Weight Paths

For the minimum-weight path method, we introduce link weights defined in terms of the difference between the original observed link flows V_o and the improved flows V_b that exist on a link during the NB process. For each link (i,j),

$$d = \frac{|V_o - V_b|}{V_o} + E_o \qquad (1)$$

where E_o is a very small number, such as 1×10^{-6}. This second term in the expression is necessary to prevent $d = 0$ on all links at the start of the process and on any link not yet adjusted. As link flow adjustments take place, the first term begins to dominate the second.

Method NB2

Step 1. Identify all nodes j in the network that are not origin or destination centroids but have unbalanced flows and place them in the set of unbalanced nodes (U) in order of their original node numbers.

TABLE 1 SIGN OF UNIT FLOW CHANGE FOR NB2 STEP 6

$I(u)$	Z	Flow Change Along Path	Link Flow Change[a]
>0	P	1 less from Z to u	−1[b]
	A	1 more from u to Z	+1[b]
<0	P	1 more from Z to u	+1[b]
	A	1 less from u to Z	−1[b]

[a] Change in flow on each link along path (Z,u) or (u,Z).
[b] If direction of link is opposite that of path flow change, sign of link flow change should be reversed.

Step 2. If set U is empty, stop. Otherwise, remove the first node in set U and call it the "u-node."

Step 3. Calculate the link weights $d(i,j)$ for each link in the network using Equation 1.

Step 4. Using an appropriate shortest-path algorithm (*3*) and the link weights $d(i,j)$, find the minimum-weight paths from the current u-node to all centroids, treating all links as two-way links, regardless of their actual orientation.

Step 5. Identify the centroid Z having the smallest path weight from the u-node. This path from u to Z has the least accumulated differences along it.

Step 6. Send one unit of flow along the path (u,Z). This unit of flow will be positive or negative, depending on the sign of $I(u)$, the orientation of each link along the path, and whether the centroid Z is an origin (P) or a destination (A) (see Table 1). Update $I(u)$ such that $|I(u)| = |I(u)| - 1$.

Step 7. If $I(u) = 0$, go to Step 2. Otherwise, go to Step 3.

An example implementation of Step 6 may be helpful at this point. Figure 1 shows the minimum-weight path from u to the centroid Z identified in Step 5. Let us say that $I(u) = +1$ and the centroid Z is an origin (P) node. This path contains 2 two-way links, $(u,9)$ and $(7,Z)$, and 2 one-way links, (8,7) and (8,9). We do not know which of the four links incident to node u has the faulty counts that caused $I(u)$ to be nonzero, so we will transfer this flow imbalance $I(u)$ to the nearest (in terms of link d-weights) centroid. Since Z is a P-node in this illustration, Table 1 indicates that one unit of flow must be deducted from all links on this path from Z to u, unless this direction violates a link orientation. Such a violation occurs for link (8,7), so flow on this link is increased by one unit in its only permitted direction.

In accordance with Table 1, we make the following adjustments to the link flows along the minimum-weight path in Figure 1: $V(Z,7) = V(Z,7) - 1$; $V(8,9) = V(8,9) - 1$; $V(9,u) = V(9,u) - 1$; but $V(8,7) = V(8,7) + 1$. At node 9, one less unit of flow is received from node 8, but one less flow unit is sent on to node u, so the previous value of $I(9)$ is preserved. At node 7, which is one end of the "backwards" link (8,7), one less flow unit is received from node Z, but one more unit is received from node 8, thereby preserving $I(7)$. Likewise, node 8 is preserved by sending one more unit to node 7, but one unit less to node 9. If Z were a destination (A) centroid, the direction of flow adjustment would be reversed (see row 2, Table 1) and link (8,9) would be the backwards link. It would have its flow reduced by 1, whereas the "forward" links along the path from u to Z would have a flow change of +1. The reader is invited to verify that, in this case and for the cases of rows 3 and 4 of Table 1, the link flow changes along the minimum-weight path produce the desired

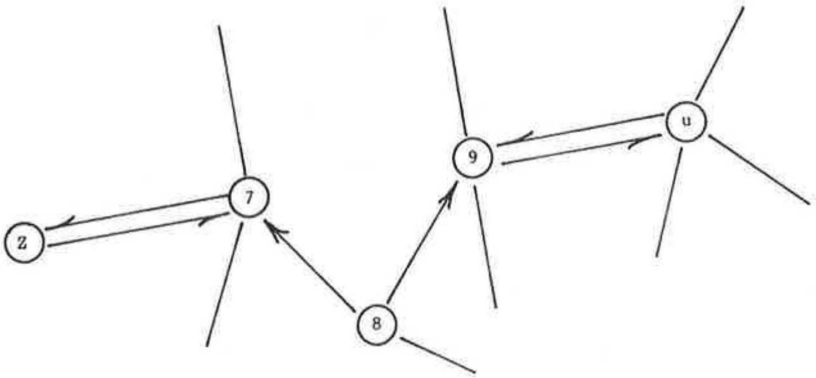

FIGURE 1 Example of NB2 Step 6.

results: $I(u)$ moves toward zero and $I(i)$ for all other nodes i is unchanged.

Discussion of Method NB2

We found several features of NB2 appealing. Once a u-node is balanced, it stays balanced. By sending flows through the network from the current u-node to a centroid in accordance with Step 6 (Table 1), every intervening node i—some of which may have already been balanced—has its $I(i)$ value unchanged. (See the case of node 9 in Figure 1 for the first example presented above.) Unlike NB1, where convergence is not guaranteed and the choice of the number of iterations can affect the outcome, NB2 "visits" each u-node only once, balances it, and moves on to the next u-node. We have devised an effective system for adjusting link volumes in a way that disturbs as few link counts as possible—and primarily those links with their V_b values still close to their V_o values. This would seem to bias the flow changes in a favorable way—toward smaller eventual network-wide error (goodness-of-fit) measure values.

A possible inefficiency in NB2 is its use of a unit flow adjustment. In cases where $I(u)$ could approach 100—any greater imbalance in our network would probably be due to a data collection or processing error—it might be wiser to use a larger flow adjustment. An adjustment of perhaps 0.5 * $I(u)$ could be used in at least the first several applications of Step 6. However, we did not adopt this procedure for two reasons:

1. A belief that too many vehicles might be sent along the smallest minimum-weight path, thereby distorting the link d-weights for the remainder of NB2 and precluding the best fit of V_b versus V_o. Until an adequate investigation of the best fraction of $I(u)$ to send—and to what extent it may change from network to network—is carried out, we prefer the unit flow adjustment.
2. A desire to build from simplicity. The unit flow adjustment may be somewhat inefficient, but unless this potential flaw becomes a detriment in real applications, its current form appears suitable for comparison with other methods.

Method NB3: Minimax Variation

The minimax variation method is designed to pay closer attention to the d-weights of certain individual links on the minimum-weight paths from the u-node to the various centroids. The idea behind NB3 is to change flows on links that have been changed relatively little earlier in the balancing process and avoid those links that have had relatively large changes. Method NB3 is a minor variation of NB2; only Step 5 is different.

Method NB3

Steps 1–4. Same as those in Method NB2.

Step 5. (a) Find the link with the largest d-weight on each of the minimum-weight paths found in Step 4. Call these links "maxilinks." (b) Select the minimum-weight path having the maxilink with the smallest d-weight and call its associated centroid Z.

Steps 6 and 7. Same as those in Method NB2.

Discussion of Method NB3

The selection rule in Step 5 can be best explained using an example. The minimum-weight paths from the u-node to three centroids are shown in Figure 2. Method NB2 would choose path $(u,z1)$ at its Step 5, because that path's total weight (0.08) is smaller than 0.12 for path $(u,z2)$ and 0.10 for path $(u,z3)$. However, the maxilinks on these three paths have d-weights 0.08, 0.04, and 0.06, respectively. Thus, the second path, $(u,z2)$, has the minimum maxilink and would be chosen by

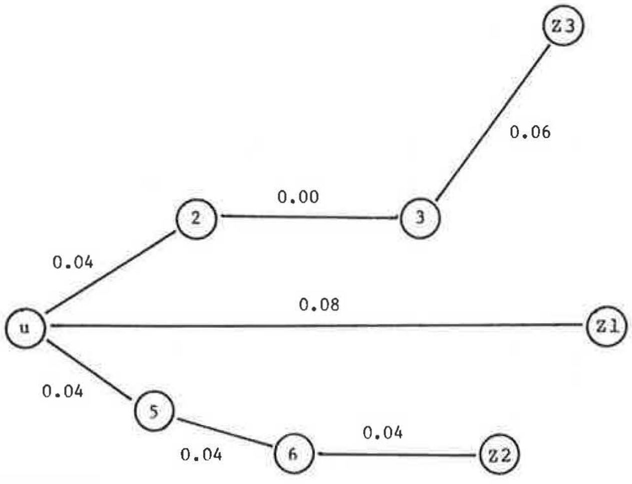

FIGURE 2 NB3 Step 5.

NB3 at its Step 5. Whereas Method NB2 would make changes to one link with *d*-weight 0.08, Method NB3 would cause changes on three links, each with *d*-weight 0.04 in this example. Until tests are conducted, it is not clear which method would be generally superior.

Method NB4: Minimum *d*-Weight Links

The minimum *d*-weight link method extends the evolution begun with NB2 and continued with NB3. As in NB3, the links on the minimum *d*-weight paths found in Step 2 are modified to find a path between the current *u*-node and each centroid that has links with *d*-weights as small as possible. Whereas NB3 identifies a maxilink on each minimum *d*-weight path, NB4 is allowed to modify the minimum *d*-weight path itself if a link on that path can be replaced by a link having a smaller *d*-weight. This is accomplished through an adaptation of the "triple operation" used in Floyd's shortest-path algorithm (*3*). Floyd replaces the subpath (i,k) with the subpath (i,j,k) if $w(i,j) + w(j,k) < w(i,k)$, where $w(i,j)$ is the current smallest cost or distance from node i to node j. Method NB4 replaces link (i,k) on the minimum *d*-weight path with the subpath (i,j,k) if $\text{Max}\{d(i,j),d(j,k)\} < d(i,k)$. This procedure modifies a good (namely, minimum *d*-weight) path to produce a path that is better with respect to a specific objective: to avoid involving links that already have large $|$percent V_b − percent $V_o|$ values in the balancing process.

Steps 1–4. Same as those in Methods NB2 and NB3.

Step 5. Along each minimum *d*-weight path from the current *u*-node to all centroids, replace a link (i,k) with the subpath (i,j,k) if $d(i,k) > \text{Max}\{d(i,j), d(j,k)\}$.

Steps 6 and 7. Same as those in Methods NB2 and NB3.

COMPARATIVE ANALYSIS

As Methods NB1 to NB4 were being developed, a discussion appeared (*4*) of a computer program written in BASIC to implement the NB scheme outlined elsewhere (*1,5*). We welcomed the opportunity to test our four methods against something besides each other. We started with the example problems cited previously (*4,6*) and compared the results of our methods with those just mentioned.

Measures of Merit

On what basis should the methods be compared? Clearly, the only frame of reference we have is the original link counts V_o. Although we need to balance the nodes to prepare the data for our traffic assignment model, we should do so by revising the V_o-values as little as necessary. Therefore, we considered several measures of merit for our comparative analysis.

1. Root Mean Square Error (RMSE). The RMSE standard measure for comparing a generated value with its target takes the form of

$$\text{RMSE} = \left|\frac{1}{n}\sum_{i,j}(V_b - V_o)^2\right|^{1/2}$$

This measure, of course, assesses disproportiontely larger penalties to greater differences between V_b and V_o.

2. Mean of the absolute percent difference between V_b and V_o (mean abs % diff). For each link, the percent difference (PD) between V_b and V_o is calculated as

$$PD = \frac{V_o - V_b}{V_o} \times 100$$

which leads to

$$\text{mean abs \% diff} = \frac{1}{n}\sum_{i,j} |\, PD(i,j)\,|$$

This is a more intuitive measure of the differences occurring between V_b and V_o than is RMSE.

3. Maximum percent absolute difference (max % diff). This measure identifies the worst single match of V_b to V_o at the end of the NB procedure, based on the expression for PD above.

4. Number of links with absolute difference greater than X percent (links > $X\%$). An NB method may produce one or two poor matches of V_b to V_o for measure 3 but for most links provides a good match. This measure offers a more specific description of this kind of behavior than the others. For the size of the networks we tested, we thought it best to set X at approximately one-half the mean value of "max % diff" observed for the methods being tested.

5. Mean difference (mean diff). One might expect that little or no change in total link volumes would result from an NB procedure, but that is not the case. The flow adjustment process may have to add or deduct flows from links to restore conservation of flow at each node. Thus, the mean value of $(V_o - V_b)$ with the sign of this difference for each link retained is a rough indication of the change in vehicle miles of travel that accompanies the NB procedure.

6. Worst-case computational complexity. The MLM method was analyzed as having a worst-case complexity of $O(n^3)$. For NB1, it was $O(k{*}n)$, where k equals the number of iterations needed to reach convergence. For NB2 and NB3, it was $O(W * n^2)$, where $W = \Sigma_u|I(u)|$. Finally, NB4 was evaluated at $O(W * n^3)$.

These measures have been introduced in order of their importance to us in assessing the performance of the various methods. The parenthetical abbreviations for the measures correspond to the column headings in Tables 2 and 3, which summarize our tests on the two sample problems.

Discussion

In both problems, Methods NB3 and NB4 accomplish what they were designed to do—minimize the worst case (max % diff). However, the actual computer time for NB4 was considerably longer than that for the other methods, making it unlikely that NB4 would be practical on a microcomputer. Considering all the measures in Tables 2 and 3, Methods NB2 and NB3 performed moderately well on the smaller problem and not so well on the larger problem. What was surprising to us was the behavior of Method NB1, Automated Trial and Error, which did poorly on the small network but quite well

TABLE 2 COMPARATIVE RESULTS FOR SAMPLE PROBLEM 1

	Measure of Merit				
Method	(1)RMSE	(2)Mean Abs % Diff	(3)Max % Diff	(4)Links >9.5%	(5)Mean Diff
MLM	38.5	10.7	20.0	8	+ 4.6
NB1	54.8	14.1	−30.2	10	−14.7
NB2	46.8	12.1	−16.8	7	+ 5.9
NB3	43.2	12.1	−14.2	10	+ 4.9
NB4	44.6	12.3	−14.2	11	+ 7.0

NOTE: Six nodes (3 centroids), 12 links, mean V_o = 337.8, W = 631. See paper by Beagan (*4*).

TABLE 3 COMPARATIVE RESULTS FOR SAMPLE PROBLEM 2

	Measure of Merit				
Method	(1)RMSE	(2)Mean Abs % Diff	(3)Max % Diff	(4)Links >9.5%	(5)Mean Diff
MLM	169.4	3.8	− 8.8	10	+ 5.9
NB1	169.7	3.7	− 9.2	9	+ 0.03
NB2	168.8	4.5	−10.1	12	−16.6
NB3	178.8	4.6	− 7.2	18	+27.6
NB4	186.6	4.3	− 7.2	17	+34.7

NOTE: Twenty-three nodes (5 centroids), 32 links, mean V_o = 3303.7, W = 2486. See presentation by Beagan (*6*).

TABLE 4 COMPARATIVE RESULTS FOR CAMPUS NE NETWORK

	Measure of Merit						
Method	(1)RMSE	(2)Mean Abs % Diff	(3)Max % Diff	(4)Links >10%	(5)Mean Diff	(7)Nodes with $I(u) \geq 1$	(8)Max $I(u)$
MLM	37.3	8.2	64.0	24	+ 3.8	11	+1, −1
NB1(10)	38.1	7.1	− 59.1	20	+ 4.4	22	−14
NB1(100)	68.1	13.2	−143.1	36	+27.4	24	− 6
NB2	48.7	13.8	− 63.8	51	+ 1.7	0	0
NB3	45.0	17.7	75.0	67	+ 5.2	0	0

NOTE: Fifty-four nodes (21 centroids), 128 links, mean V_o = 291.69.

on the larger one. In fact, Method NB1 outperformed the MLM procedure (Table 3) in three of the first five measures.

Already some hypotheses can be formulated. These pertain to the influence of network size (number of nodes, links, and centroids), initial magnitude of $W = \Sigma_{\text{all } u} |I(u)|$, and mean link flow values (V_o) on the performance of an NB method. We applied the MLM method and Methods NB1 through NB3 to the network that required node balancing in the first place, Campus NE. The results are summarized in Table 4.

The format of Table 4 is a variation of that of Tables 2 and 3, brought about by the behavior of Method NB1 on Campus NE. As the number of iterations increases, the nodes become more nearly balanced. In the process, however, the difference between the improved link flows V_b and the observed flows V_o tends to increase. Thus, we list two versions of Method NB1: NB1(10) has gone through 10 iterations and NB1(100) has been carried through 100 iterations. Also, the measures of merit used in Tables 2 and 3 are augmented by two that reflect the true objective of the methods. These are

7. Number of nodes still unbalanced (nodes with $I(u) > 1$). We define "unbalanced" here as having an imbalance $I(u)$ of more than one vehicle.

8. Maximum absolute imbalance (max $|I(u)|$). This is thought to be an indication of the balancing work left undone and perhaps an early indication of cases in which convergence is impossible.

Comparing NB1(10) and NB1(100) without these new measures would lead us to conclude that 10 iterations are better than 100, but measures 7 and 8 indicate otherwise. These new measures also point out the superiority of methods NB2 and NB3. In one well-organized iteration, they produced a set of link flows in which each node had $I(u) = 0$. Furthermore, their first five measures are competitive with methods NB1(10) and NB1(100).

MATHEMATICAL PROGRAMMING APPROACHES

Having acquired some experience with the criteria (measures of merit) by which NB methods might be evaluated, we began to think of each of these criteria as the basis for an NB method. The result was six mathematical programming (MP) formulations, identified as NB5 through NB10. Each MP formu-

TABLE 5 SUMMARY OF MP RESULTS FOR EXAMPLE 1

	Measure of Merit					
Method	8	3	5	9	10	2
MLM	58.0	− 20.0	4.6	0	35.1	10.8
NB3	78.0	− 14.2	4.9	0	40.9	12.1
NB5	− 47.7	19.5	13.3	2.9	45.1	14.2
NB6	78.1	14.0	14.8	2.9	41.7	12.2
NB7	314.0	100.0	0	−0.4	87.7	31.5
NB8	−404.5	−159.3	26.3	0	156.2	49.2
NB9	−184.0	− 73.6	−5.3	−3.1	25.4	9.2
NB10	184.0	63.2	25.4	8.3	25.4	8.3

NOTE: Example 1 has 6 nodes, 12 links.

lation included the conservation-of-flow constraints. In each formulation, the objective function was written simply as min z. What distinguished the six formulations were the rest of the constraints, which also defined z in the objective function. In the following, the objective function for each MP formulation is written in words, followed by the constraint equations that define z.

NB5: Minimize the largest absolute link volume change.

$$z \geqslant |V_o(i,j) - V_b(i,j)| \qquad \text{for each link } i,j$$

(MP formulation NB5 has its basis in measure of merit 8.)

NB6: Minimize the largest percent absolute link volume change.

$$z \geqslant \frac{|V_o(i,j) - V_b(i,j)|}{V_o(i,j)} \qquad \text{for each link } i,j$$

(See also measure of merit 3.)

NB7: Minimize the average link volume change.

$$c(i,j) = \frac{V_o(i,j) - V_b(i,j)}{V_o(i,j)} \qquad \text{for each link } i,j$$

$$z = \sum_{i,j} c(i,j)$$

(See also measure of merit 5.)

NB8: Minimize the average percent link volume change.

$$c(i,j) = \frac{V_o(i,j) - V_b(i,j)}{V_o(i,j)} \qquad \text{for each link } i,j$$

$$z = \sum_{i,j} c(i,j)$$

This is a variation of measure of merit 5. Let "average percent link volume change" be measure of merit 9.

NB9: Minimize the average absolute link volume change.

$$c(i,j) = |V_o(i,j) - V_b(i,j)| \qquad \text{for each link } i,j$$

$$z \geq \sum_{i,j} c(i,j)$$

Another variation of measure of merit 5, "average absolute link volume change" becomes measure of merit 10.

NB10: Minimize the average percent absolute link volume change.

$$c(i,j) = \frac{|V_o(i,j) - V_b(i,j)|}{V_o(i,j)} \qquad \text{for each link } i,j$$

$$z \geq \sum_{i,j} c(i,j)$$

(MP method NB10 is based on measure of merit 2.)

These six formulations were applied to Example 1, with the results summarized in Table 5. That method NB5 was designed to optimize the NB solution with respect to measure of merit 8 is borne out by the best entry (underlined) in column 2, row 3. Likewise, NB6 performs best with respect to measure of merit 3. The results for the MLM and NB3 methods used earlier in this paper are included here for comparison. This is because MLM is the "literature standard," but we prefer the objective built into NB3.

Although space limitations prevent a link-by-link listing of each solution, we can report that only NB5 and NB6 had sets of balanced link flows that were similar. There were considerable differences in $V_b(i,j)$ values produced by the various methods. Table 5 indicates that NB7 and NB8 perform poorly for any measure of merit other than their own, and this was confirmed in other tests. Unless measure of merit 5 or 9 is the *only* important one, these methods should not be used and are omitted from the remainder of this paper.

The results of the surviving methods in Example 2 are tabulated in Table 6. Again, a method performed best where it was designed to; methods NB7 and NB8 were clearly and consistently inferior elsewhere and have been omitted from Table 6, and the earlier methods (MLM and NB3) are competitive. Also, NB5 does not seem to perform well in this example, which is a simplification of a real highway corridor.

For our final example, we return to Campus NE, an exact representation of a network with 54 nodes and 128 links. Table 7 shows that NB6 performs well for its own measure of merit (3), whereas NB5 is among the two worst surviving methods for four of the five criteria for which it was *not* designed.

CONCLUSIONS

We have sought to investigate how the NB solution for a network is affected by choice of method, which by implication also means choice of criterion. We have found that it is possible to find an optimal solution with respect to one criterion, but that the solution may be unacceptable according to other reasonable alternative criteria. On the basis of our tests, mathematical programming methods NB5 through NB8 lack versatility in this respect, whereas NB9 and NB10 do fairly well. Also doing well for most criteria are the standard MLM method and our NB3 algorithm.

It is interesting to note that NB3 and NB6 are designed to pursue the same objective: minimize percent link volume change. As an optimization routine, NB6 is always superior to NB3 for criterion 2, usually by a small margin. For most

TABLE 6 SUMMARY OF MP RESULTS FOR EXAMPLE 2

	Measure of Merit					
Method	8	3	5	9	10	2
MLM	418.0	−8.8	5.9	0.1	116.2	3.5
NB3	482.0	−7.2	27.6	0.7	130.0	4.6
NB5	257.5	100.0	57.2	9.4	175.4	17.1
NB6	489.0	−7.1	7.5	3.0	135.8	5.2
NB9	−515.0	64.3	4.3	0.4	103.0	8.4
NB10	−614.0	−12.5	24.7	0.2	113.5	2.2

NOTE: Example 2 has 23 nodes, 32 links.

TABLE 7 SUMMARY OF MP RESULTS FOR CAMPUS NE

	Measure of Merit					
Method	8	3	5	9	10	2
MLM	228.0	64.0	3.8	1.3	19.4	8.2
NB3	215.0	75.0	5.2	5.4	32.5	17.7
NB5	172.0	−550.0	−1.9	−5.8	66.5	63.8
NB6	724.1	60.5	179.8	55.8	182.2	58.2
NB9	−324.0	−241.4	−0.3	−0.2	14.0	8.2
NB10	356.0	100.0	0	0	16.9	3.5

NOTE: Campus NE has 54 nodes, 128 links.

TABLE 8 SUGGESTED RANKINGS BY CRITERION

	Method by Criterion					
Rank	1	2	3	4	5	6
Best	MLM	NB3	NB9	MLM	NB9	NB10
Second	NB3	MLM	MLM	NB10	NB10	MLM
Third	NB9	NB10	NB10	NB9	MLM	NB9
Fourth	NB10	NB9	NB3	NB3	NB3	NB3

of the other criteria, NB3 is usually superior and usually by a large margin.

Any of the four surviving methods (MLM, NB3, NB9, and NB10) has exhibited adequate versatility in our tests, but in the event that one or more of our six criteria take on special importance, we rank the four methods for each criterion in Table 8. Of course, the MP-based method formulated for a specific criterion will provide the optimal solution if the size of the problem does not exceed time or storage constraints. Another finding is that a simple algorithmic method such as NB3 seems to provide NB solutions that are of high enough quality to be comparable with the MLM. We are convinced that any competent (not necessarily advanced) computer programmer can convert the steps described for method NB3 in this paper into a usable computer code in a short time.

ACKNOWLEDGMENT

The original draft of this paper was prepared by J.D.F. during his sabbatical leave at the Institut fuer Verkehrswesen, Universitaet (TH) Karlsruhe, West Germany, May to December 1987. He wishes to thank W. Leutzbach, Director of the Institute, and others at the Institute who provided him with the resources and support that made the completion of that draft possible.

REFERENCES

1. H. J. van Zuylen and L. G. Willumsen. The Most Likely Trip Matrix Estimated From Traffic Counts. *Transportation Research B*, Vol. 14B, 1980, pp. 281–293.
2. L. G. Willumsen. Simplified Transport Models Based on Traffic Counts. *Transportation*, Vol. 10, 1981, pp. 257–278.
3. R. W. Floyd. Algorithm 97, Shortest Path. *Communications of the Association of Computing Machinery*, Vol. 5, 1962, p. 345.
4. D. F. Beagan. Balancing Traffic Counts on a Network. *McTrans*, Vol. 1, No. 2, Fall 1986, pp. 8–9.
5. H. J. van Zuylen and D. M. Branston. Consistent Link Flow Estimation From Counts. *Transportation Research*, Vol. 16B, No. 6, 1982, pp. 473–476.
6. D. F. Beagan. Maximizing the Information From Traffic Counts. Presented at the National Conference on Transportation Planning Applications, Orlando, Fla., 1987.

Trip Generation Models for Infrequent Trips

JOSE MONZON, KONSTADINOS GOULIAS, AND RYUICHI KITAMURA

The adequacy of conventional linear regression models in trip generation analysis is examined in this study. Simulation experiments are conducted to determine whether model coefficients can be accurately estimated by least-squares estimation when the dependent variable is a nonnegative integer. Following this, nonlinear, two-stage model systems are estimated by using an empirical data set to examine whether more elaborate representation of the decision process underlying trip generation will lead to improved prediction. The results of this study indicate that linear regression models of trip generation offer consistent coefficient estimates and accurate predictions, and improved performance may not be obtained by adopting more complex model systems.

The most frequently used statistical methods for trip generation analysis are the least-squares estimation of linear regression models and trip rate analysis based on cross-classification of households on the basis of a few grouping variables. Both methods draw on principles that are relatively easy to understand, and models can be estimated using commonly available statistical software packages. However, these methods involve certain assumptions and limitations that need to be well understood for valid formulation of trip generation models.

Sample size requirements usually limit the number of grouping variables that can be used in cross-classification analysis, leaving least-squares regression the standard method used whenever an adequate data set and statistical package are available. Linear regression analysis is based on the assumption that the dependent variable (number of trips) is an untruncated continuous variable. Also important is the typical assumption that one model structure explains the entire range of trip generation behavior.

These assumptions may not be entirely satisfied in typical trip generation analyses. The dependent variable in this case is a nonnegative discrete variable, not an untruncated continuous variable. Trip generation behavior may result from a two-stage decision process in which a decision to make trips on a given day is made first; then, given that trips will be made at all, the number of trips is determined. This can be most typically seen in trip generation by purpose (e.g., the number of shopping trips on a given day) or by mode (e.g., the number of transit trips).

The question that naturally arises is whether linear least-squares regression can be used successfully in trip generation analysis when its assumptions are not satisfied. The latter is especially the case when models are formulated for infrequent trips whose observed frequencies are zero for many behavioral units. The dependent variable will then be heavily truncated and the underlying decision process may involve more than one stage that cannot be adequately represented by a single model.

The objective of this paper is to shed light on the following two questions.

1. Is the linear regression method suited for trip generation analysis in which the dependent variable (number of trips) is a nonnegative integer rather than an untruncated continuous variable?
2. Can a single regression model capture trip generation behavior that may involve multistage decision processes?

The first question is examined through simulation experiments in which the number of trips made by an individual has a Poisson distribution. In the simulation, discrete numbers are generated from Poisson distributions as the number of trips generated, the parameters of the distributions are estimated by least-squares regression, and the accuracy of the parameter estimates is examined against the true values used in the simulation to generate the data. Timmermans (*1*) offers a comprehensive discussion of trip generation analysis by examining a set of alternative trip generation model formulations, including Poisson regression models, and testing their goodness-of-fit empirically. The emphasis in this study is on the extent of estimation errors that result from the application of linear regression models to Poisson data (linear regression models are misspecified in this case).

The second question is examined by estimating two models on empirical data and comparing their relative fits. The first model is a regular linear regression model. The second is based on the assumption that the trip generation process consists of two stages: in the first stage the decision is made whether to make trips of a given type; then in the second stage the number of trips is determined.

The rest of this paper is organized as follows. Trip generation models used in this study are briefly described. The results of simulation experiments are presented after that together with the discussion on whether linear regression models can be successfully used with discrete and truncated dependent variables. The next two sections offer a description of the data used to address the second question, the results of the empirical analysis, and a comparison of the two models. The final section summarizes the study.

J. Monzon, Department of Civil Engineering, University of California at Davis, Davis, Calif. 95616. Current affiliation: Planning Department, General Highway Directorate, Guatemala City, Guatemala. K. Goulias and R. Kitamura, Department of Civil Engineering, University of California at Davis, Davis, Calif. 95616.

TRIP GENERATION METHODS

Cross-classification analysis of trip generation is based on the premise that each group of households, defined in terms of a set of grouping variables, has an average trip rate that remains stable over time. The grouping variables are categorical, and groups of households can be defined by combinations of their categories. An important advantage of this straightforward method is its capability to represent the interaction effect of the classification variables, that is, systematic variation in trip rates that is uniquely associated with a particular combination of categories.

For example, let S and T be the grouping variables, and let s and t be categories of S and T, respectively. In the case of household trip generation, variable S may represent household size and T, the number of cars available to the household. Let the set of values S be assumed to be {1, 2, 3, 4, 5, 6 or more} and that for T be {0, 1, 2, 3 or more}. Let the mean trip rate, $Y(st)$, of the group of households with $S = s$ and $T = t$ be

$$Y(st) = \mu + V(s) + W(t) + Z(st)$$

where

μ = grand mean,
$V(s)$ and $W(t)$ = the effects of category s of variable S and category t of variable T, respectively, and
$Z(st)$ = interaction effect of category s and category t.

$V(s)$ and $W(t)$ represent the effects that are attributable to S and T, respectively, whereas effect $Z(st)$ is the contribution of the particular combination of categories. The statistical significance of these models can be tested by analysis of variance (ANOVA), available in most statistical packages.

In a linear regression model, the expected number of trips made by household i is represented as

$$Y_i = \beta_0 + \beta_1 H_i + \beta_2 A_i$$

where

β = model coefficient,
H_i = number of persons in household i, and
A_i = number of cars available to household i.

In this formulation, β_1 represents the average number of trips generated per household member, and β_2 the average number of trips per automobile. The number of trips is linearly related to the explanatory variables, and no interaction effect is assumed in this formulation.

Interaction effects can be represented in a linear regression model by introducing terms representing combinations of categories. Possible nonlinear effects of an explanatory variable can also be included in a linear model by using nonlinear transformation of the variable (including a step function represented by a set of dummy variables). Although it is limited to the case in which the model is linear in terms of its coefficients, the least-squares method can be used in a variety of cases involving nonlinear relations or interaction effects.

A critical limitation of the least-squares approach to trip generation may stem from the assumption that the random variation in the dependent variable can be represented by a random error term that has a continuous, untruncated distribution. The dependent variable of trip generation analysis, the number of trips, is a nonnegative integer. Ideally this variable can be modeled by using a discrete distribution, such as a Poisson or negative binomial distribution (*1,2*). Application of the least-squares method, therefore, assumes that this discrete distribution can be replaced by a continuous distribution.

Problems arise when the expected number of trips is close to 0. For example, suppose the expected value is 0.2 trip. Then possible values that the error term may assume are −0.2, 0.8, 1.8, 2.8, and so on. The error distribution is truncated at −0.2 with the probability mass associated with this error value equaling the probability that no trip will be made. If the number of trips has a Poisson distribution with a mean of 0.2, this probability will be 0.819 and the distribution of the error term will be heavily skewed.

The validity and usefulness of the least-squares estimation and resulting trip generation models may be severely limited when there are many zeroes in the observation. This situation arises when models are formulated by purpose or by mode. Another example is the case in which models are specified at the person level rather than at the household level. In these instances, the probability is much higher that no trip of a given type will be generated by a given behavioral unit. The effect of error truncation may become significant, and the quality of estimated model coefficients and test statistics may deteriorate. This problem exists in addition to the more obvious problem of producing negative values as predicted numbers of trips. The possible extent of this problem is discussed later by using simulation examples.

As a second example, it may not be possible to properly capture travel behavior with one linear model. Trip generation behavior may be a result of a two-stage decision process in which a decision is first made to make, or not to make, trips of a given type at all on a given day; then, given trips will be made, and the number of trips is chosen in the second stage. If this is the case, it is probable that the decision to make trips at all is governed by a different causal mechanism than is the choice of the number of trips. For example, consider the case in which a transit trip generation model is developed at the household level. The primary determinants of the first-stage decision may include household car ownership and the number of nondrivers, whereas the second-stage decision may be described as a function of the number of household members and number of workers.

In the analysis of this study, this possible two-stage decision mechanism is represented by a system of two models: a binary probit model that represents the decision to make a trip of a given type at all, and a linear regression model applied to the number of trips, given that trips are made. Formally, the model system can be presented as

$$A_i = \alpha' X_i + u_i$$

$$Y_i = 0 \quad \text{if } A_i \leq 0$$

$$Y_i = \beta' Z_i + v_i \quad \text{if } A_i > 0$$

where

A_i = latent variable underlying the binary choice,
Y_i = number of trips made,
α' and β' = coefficient vectors,
X_i and Z_i = vectors of explanatory variables,
u_i and v_i = normal random error terms, and u_i is assumed to have a unit variance.

The two vectors of explanatory variables, X_i and Z_i, may contain the same variables. The binary choice probability is given as

$$\Pr[Y_i > 0] = \Phi(-\alpha' X_i)$$

where the lefthand side is the probability that trips will be made at all, and Φ on the righthand side is the standard cumulative normal distribution function.

The number of trips, Y_i, is defined to be 0 if $A_i \leq 0$. Given that trips are made at all ($A_i > 0$), the expected number of trips is $\beta' Z_i$. The unconditional expected number of trips can be obtained as

$$E[Y_i] = E[Y_i \mid Y_i = 0]\,Pr[Y_i = 0] + E[Y_i \mid Y_i > 0]\,Pr[Y_i > 0]$$
$$= E[Y_i \mid Y_i > 0]\,\Pr[Y_i > 0]$$
$$= \beta' Z_i \Phi(-\alpha' X_i).$$

The model system can be estimated simultaneously using the maximum likelihood method. Use of this method, however, requires the development of a computer code to estimate the coefficients. Alternatively, the model system can be estimated equation by equation using easily available binary probit and linear regression codes. A problem arises when the error term of the probit trip choice model (u_i) and that of the linear trip generation model (v_i) are correlated. Possible biases in coefficient estimates are avoided in this study by introducing a correction term into the linear regression model. Further discussions of this method are given elsewhere (*3–7*).

LINEAR REGRESSION ON SIMULATED POISSON TRIP DATA

The question of whether the least-squares regression approach will produce adequate model coefficients and test statistics is addressed in this section. Simulated data sets are generated assuming that trip generation is a Poisson process, the values of the parameters used to generate the data are then estimated by least-squares regression, and the quality of the parameter estimates is examined.

Trip generation is simulated as follows. For each case simulated, the expected number of trips is assumed to be

$$m_i = \beta_0 + \beta_{1i} X_{1i} + \beta_{2i} X_{2i}$$

where

m_i = expected number of trips for case i,
X_{1i} and X_{2i} = independent variables, and
β_0, β_1, and β_2 = model parameters to be estimated later by least-squares regression.

In the simulation, X_{1i} and X_{2i} are assumed to be 0–1 binary variables. Therefore, each case has one of the following four possible expected values: β_0, $\beta_0 + \beta_1$, $\beta_0 + \beta_2$, $\beta_0 + \beta_1 + \beta_2$. The number of trips, Y_i, is simulated using the following Poisson probability:

$$\Pr[Y_i = n] = \exp(-m_i) m_i^n / n! \quad n = 0, 1, 2, \ldots$$

where m_i is the expected number of trips for case i as defined above.

In each simulation run, cases are evenly divided into four groups, each having fixed values of the X's (and therefore one of the above four expected values). Consequently, dependent variable values in the data set come from four Poisson distributions. One hundred cases are generated for each group, and least-squares regression is applied to the resulting 400 cases in each simulation run.

TABLE 1 ORDINARY LEAST-SQUARES ESTIMATES OF THE PARAMETERS OF SIMULATED POISSON TRIP GENERATION

	Theoretical Values	Simulation Results*	Estimated S.E.*	True S.E.
Mean Y	.550	.547		
Constant	.100	.109	.121	.040
β_1	.300	.285	.139	.083
β_2	.600	.593	.139	.056
R^2	.170	.165		
Mean Y	1.000	0.980		
Constant	.100	.126	.084	.062
β_1	.600	.553	.097	.136
β_2	1.200	1.160	.097	.091
R^2	.310	.310		
Mean Y	1.900	1.889		
Constant	.100	.115	.083	.072
β_1	1.200	1.204	.097	.165
β_2	2.400	2.385	.097	.133
R^2	.486	.498		
Mean Y	2.800	2.771		
Constant	.100	.098	.140	.064
β_1	1.800	1.801	.162	.188
β_2	3.600	3.547	.162	.126
R^2	.591	.604		
Mean Y	3.450	3.337		
Constant	3.000	3.217	.168	.115
β_1	.300	.052	.194	.122
β_2	.600	.179	.194	.242
R^2	.032	.007		
Mean Y	3.900	3.847		
Constant	3.000	2.956	.166	.154
β_1	.600	.481	.192	.282
β_2	1.200	1.303	.192	.389
R^2	.103	.123		
Mean Y	4.800	4.797		
Constant	3.000	3.016	.188	.169
β_1	1.200	1.266	.217	.163
β_2	2.400	2.297	.217	.151
R^2	.273	.270		
Mean Y	5.700	5.580		
Constant	3.000	3.183	.194	.088
β_1	1:800	1.260	.225	.113
β_2	3.600	3.541	.225	.184
R^2	.415	.414		

*Average of 10 Simulation runs.

Note: The parameter values (constant, β1, β2) used to simulate data are shown under "Theoretical Values", and the ordinary least squares estimates of the parameters are shown under "Simulation Results".

A total of 10 simulation runs are performed for each combination of parameter values. The results of the simulation experiments with ordinary least-squares estimation are summarized in Table 1 for the eight sets of parameter values examined in this study.

The simulation experiment offers evidence that least-squares regression yields adequate estimates of trip generation parameters even when the actual generation process is not compatible with its assumptions. The regression method performs very well when the theoretical R^2 (defined as the ratio of the systematic variation of the mean number of trips to the total theoretical variance) is higher than 0.10. The only exception is the case with a large constant (3.0) combined with small slope coefficients ($\beta_1 = 0.3$, $\beta_2 = 0.6$), for which the theoretical R^2 is only 0.032. Other than this exceptional case, the least-squares estimates adequately account for variations in trip generation as indicated by the R^2 values that are close to the theoretically expected values and the parameter estimates whose averages accurately replicate the true values used to generate the simulated data.

The estimated standard errors of the coefficient estimates, however, do not accurately represent the true standard errors obtained by evaluating the standard deviations of coefficient estimates from 10 repeated simulation runs. To examine whether this is due to heteroscedasticity (a variation in the variance of random errors across cases), weighted least-squares estimation was performed using as the weight the inverse of the square root of the predicted number of trips obtained by ordinary least-squares estimation. This weight was theoretically derived from the fact that the variance of a Poisson-distributed random variable equals its expectation.

Weighted least-squares estimation offered some improvement in estimated standard errors, although this improvement was at the cost of significantly diminished accuracy of coefficient estimates. The divergence between the estimated and true coefficient values was so large that it was only appropriate to conclude that the weighted least-squares procedure was not suitable for trip generation analysis when the underlying processes are composite Poisson processes with relatively small means (ranging from 0.1 to 7 trips). Although the reason for the poor performance is still undetermined, the parameters of trip generation processes may be accurately estimated by ordinary least-squares regression when the systematic variation in the data is reasonably high (with an R^2 of, say, 0.1 or higher).

TWO-STAGE TRIP GENERATION MODELS

Data Set

In the remainder of this paper the adequacy of linear trip generation models is examined by applying an alternative model formulation to empirical data. The conventional linear regression models and two-stage models described earlier are estimated and their relative performance is studied. The intent of the effort is to infer the validity of conventional linear models and the value of more elaborate models. Note that, unlike the simulation analysis above, the true behavioral mechanism is not known in this empirical analysis. The validity of the alternative models is therefore evaluated in the study on the basis of its statistical fit.

The results of the 1980 Southeastern Michigan Transportation Authority survey are used in the estimation of two-stage trip generation models. This standard home interview survey file contains demographic and socioeconomic attributes of the household and its members and records of all trips made by each household member (5 years old and over) on the survey day, including trips made by nonmechanized modes. The person, rather than the household, is used as the unit of analysis in this study. All individuals at least 16 years of age are included in the study sample. This particular cutoff age is selected because individuals can qualify to be licensed to drive at this age and become active users of the automotive transportation system.

A wide range of variables is considered in the model development to best capture trip generation behavior using the two types of models. These variables include age, sex, occupation, car availability, household composition, life cycle stage, income, residence county, residence area type, and day of the week (Table 2). The age and sex of an individual are known to influence trip generation significantly (*8, 9*) and therefore are included in this analysis. In addition, detailed occupation categories are used in the model development with the anticipation that variations in lifestyles can be captured by them.

Past studies also indicate that household structure influences trip generation behavior even when the model is formulated at the individual level. For example, a study shows that various measures of individual mobility vary significantly and meaningfully across subgroups defined by life cycle stages (*10*). Household structure is represented in this study by the number of household members by age and sex and by a set of five life cycle stages as defined in the data file.

Because the models are formulated at the individual level, car availability, rather than car ownership, is used to explain trip generation. The following four levels of car availability are defined according to the license-holding status of the individual and car ownership of the household:

Always: the individual holds a driver's license, and the number of cars available to the household equals or exceeds the number of adults in the household;

Usually: the individual holds a driver's license and at least one car is available to the household, but the number of cars available is less than the number of adults in the household;

Sometimes: the individual does not hold a driver's license but at least one car is available to the household; and

Never: no car is available to the household.

Combined household income is classified in the data file into 11 categories. In the analysis these categories are combined into four income classes as shown in Table 2.

The land use type and density variables are introduced to account for the possibility that trip generation is influenced by the availability of opportunities around the home base. The residence county variables are introduced in the belief that differences in lifestyles that are not reflected in the household and person-attribute variables in the data file can be captured by these variables. Note, however, that the notion that trip generation depends on residence area contradicts the commonly held belief that trip generation of a household or individual of given characteristics is invariant across areas.

Estimation Results

The final model forms and estimation results are summarized in Table 3. The dependent variable is the total number of person trips generated by an adult household member. All regression models are estimated using weighted least squares with the weight defined as $\theta(|Y|)^{\tau}$ where θ and τ are estimated

TABLE 2 VARIABLES USED IN MODEL FORMULATION

VARIABLE	DEFINITION
Age and Sex	
AGE:16-30	1 if the age is between 16 and 30; 0 otherwise
AGE:31-50	1 if the age is between 31 and 50
AGE:51-64	1 if the age is between 51 and 64
AGE:65+	1 if the age is 65 or over
MALE	1 if male
FEMALE	1 if female
Occupation	
PRO/TECH	1 if professional or technical
FARM	1 if farmer, farm manager, farm laborer, or farm foreman
LABORER	1 if non-farm laborer
MANAGER	1 if manager, official, or owner of a business
CLERICAL	1 if clerical and similar worker
SALES	1 if sales worker
CRAFTSMAN	1 if craftsman, foreman, and similar worker
OPERATOR	1 if equipment operator or motor vehicle operator
HHLDWORKER	1 if private household worker, maid, butler, etc.
SERVICE	1 if service worker
MILITARY	1 if in military
OTHER	1 if other worker
Car Availability	
ALWAYS	1 if the individual has a driver's license and the number of cars is no less than the number of adults in the household
USUALLY	1 if the individual has a driver's license and the number of cars is less than the number of adults in the household
SOMETIMES	1 if the individual does not have a driver's license and the household has at least one car available
NEVER	1 if no car is available to the household
Household Structure	
NADULTS	Number of adults (≥ 18 years old) in the household
NCHILD:0-4	Number of children of 0 to 4 years old
NCHILD:5-15	Number of children of 5 to 15 years old
NCHILD:16-18	Number of children of 16 to 18 years old
NMALES	Number of males in the household
NFEMALES	Number of females in the household
Household Lifecycle Stage	
NOCHLD-YNG	1 if head of household less than 35 years of age, and no children in the household less than 18 years of age
NOCHLD-MID	1 if head of household 35 years of age or older, but less than 65 years of age, no children in the household
NOCHLD-OLD	1 if head of household 65 years of age or older, no children in the household less than 18 years of age
PRESCHOOL	1 if the youngest child in the household is less than 6 years of age, for head of household of any age
SCHOOLAGE	1 if the youngest child in the household is 6 years of age or older, for head of household of any age
Household Income	
LOW	1 if household annual income is less than $10,000
MID-LOW	1 if household annual income is between $10,000 and $20,999
MID-HIGH	1 if household annual income is between $21,000 and $34,999
HIGH	1 if household annual income is $35,000 or more
Residence County	
DETROIT	1 if residence zone is in Detroit
WAYNE	1 if residence zone is in Wayne County
OAKLAND	1 if residence zone is in Oakland County
MACOMB	1 if residence zone is in Macomb County
WASHTENAW	1 if residence zone is in Washtenaw County
MONROE	1 if residence zone is in Monroe County
STCLAIR	1 if residence zone is in St. Clair County
LIVINGSTON	1 if residence zone is in Livingston County
Residence Area Type	
COMMERCIAL	1 if 10 or more employees per acre of usable land
HIDENSITY	1 if less than 10 employees and more than 5 dwelling units per acre of usable land
MIDDENSITY	1 if less than 10 employees and from 0.5 to 5.0 dwelling units per acre of usable land
LOWDENSITY	1 if less than 10 employees and less than 0.5 dwelling units per acre of usable land
Day of Week	
MONDAY	1 if Monday
TUESDAY	1 if Tuesday
WEDNESDAY	1 if Wednesday
THURSDAY	1 if Thursday
FRIDAY	1 if Friday

TABLE 3 TWO MODELS OF TOTAL PERSON-TRIP GENERATION

	Conventional Linear Model		Two-Stage Model System			
			Probit Trip Choice		Conditional Trip Generatn	
	(WLS)		(ML)		(WLS)	
	ß	t	ß	t	ß	t
AGE:31-50	-.113	-1.33	-.302	-5.23		
AGE:51-64	-.608	-6.51	-.608	-10.24		
AGE:65+	-.820	-7.34	-.753	-10.68		
MALE			.326	6.80	-.411	-5.30
PRO/TECH	.661	5.78	.290	3.36		
LABORER					-.398	-2.68
MANAGER	.927	5.38	.453	3.29	.305	1.79
CLERICAL	.476	3.37	.391	3.71	-.232	-1.68
SALES	.381	2.25	.122	1.03		
CRAFTSMAN			.238	1.91	-.388	-2.66
SERVICE	.693	3.68	.469	3.34		
OTHER					-.385	-2.39
ALWAYS	.912	9.26	.206	3.92	.368	4.14
USUALLY	.439	4.58				
NEVER					-.392	-3.46
NADULTS			-.058	-1.84	.109	2.98
NCHILD:0-4			-.039	-.44		
NCHILD:5-15	.133	3.09			.151	3.82
NCHILD:16-18			.205	3.17	-.236	-2.75
NMALES			-.117	-2.67		
NFEMALES	.076	1.55				
NOCHLD-YNG			-.299	-5.15	.284	2.81
SCHOOLAGE	.332	3.77	.237	4.06		
LOW	-.248	-2.78	-.436	-6.63		
MID-LOW			-.117	-2.01		
HIGH			.124	1.65		
WAYNE	-.158	-1.10	-.215	-4.18		
OAKLAND	.349	2.49			.291	3.01
MACOMB	.203	1.28			.195	1.66
WASHTENAW	.601	3.14			.842	4.95
HIDENSITY	-.188	-1.97			-.352	-3.86
MIDDENSITY	.182	2.18			.114	1.37
MONDAY	-.203	-2.25			-.155	-1.62
TUESDAY			.166	3.09	-.154	-1.74
WEDNESDAY	-.148	-1.58				
FRIDAY	.098	1.00	.108	1.77	.053	.51
Correction Term*					.485	3.97
Constant	2.231		1.172		2.834	
R^2	.138				.079	
F (df)	34.07 (24,5109)				15.85 (21,3884)	
-2[L(ß)-L(0)] (df)			2594.2 (24)			
-2[L(ß)-L(C)] (df)			1040.0 (23)			
N	5134		5077		3906	

*Introduced to correct for possible biases due to the correlation between the error term of the probit choice model and that of the conditional trip generation model.

WLS: Weighted least squares regression
ML: Maximum likelihood estimation
df: degrees of freedom
L(0): Log-likelihood with all coefficients constrained to 0
L(c): Log-likelihood with the constant term alone
L(ß): Log-likelihood with no constraints
N: Sample size

-2[L(ß)-L(0)] and -2[L(ß)-L(0)] have chi-square distributions with indicated degrees of freedom, respectively. The former can be used to test the collective significance of all model coefficients, and the latter to test the significance of the model coefficients excluding the constant term.

by regressing the squared residual on the predicted number of trips (unlike the simulation analysis above, practically no differences emerged in this case between the ordinary least-squares and weighted least-squares estimation results).

Columns 1 and 2 of Table 3 present the estimated model coefficients and *t*-statistics of the conventional linear regression model. The coefficients of the age variables show the well-established relationship that trip generation declines as age increases. The results also suggest that white-collar workers tend to make more trips, that the presence of school-age children increases the adult members' trip generation, and that low-income families make fewer trips. The car availability variables are highly significant, indicating that trip generation increases with car availability.

The two sets of variables that are not normally included in trip generation models, residence county and day of the week, are both significant. The day-of-the-week variables suggest that trip generation is suppressed on Mondays. The coefficient for Friday trip generation is positive, although insignificant. This finding is consistent with earlier results that trip generation increases toward the end of the week (*11*), but statistically is not as conclusive.

The set of residence county variables suggests that residents of suburban counties tend to make more trips. This area-specific effect is in addition to those represented by the income variables or by the land-use type variables, the latter of which indicate that residents in the area with 0.5 to 5 dwelling units per acre make more trips. Although it is not possible to pinpoint the reasons for the significance of the county variables, it is conceivable that these variables act as proxies for unobserved and geographically correlated factors such as ethnic backgrounds.

The probit trip choice model includes a set of variables that is similar to that of the conventional model. However, the effect of income variables is more pronounced, whereas car availability variables are less dominant in the probit choice model. The sex variable is significant in the probit model and indicates that a man makes trips on any given day more frequently than does an equally situated woman. Importantly, no land use type variable is present and only one residence county variable is included in the model. This result suggests that the choice of whether to make trips at all does not vary substantially by geographical area.

Columns 7 and 8 of Table 3 show the conditional trip generation model that accompanies the probit trip choice model. Quite notable is the result that the age variables and income variables, both significant in the probit choice model, are excluded from the conditional trip generation model because of their insignificance. On the other hand, the land use type variables, which are not included in the probit model, are included in the conditional trip generation model.

It is also notable that when a variable is included in both models, its coefficient values tend to contradict each other. For example, the sex variable (MALE) is positive and significant in the probit model, but negative and significant in the conditional trip generation model. These values imply that women have a higher probability of not making any trips on a given day, but given that they make trips at all, they tend to make more trips than men.

The two-stage model system thus offers indications that the choice of making trips at all and the determination of the number of trips are influenced by overlapping but different sets of factors. New behavioral insights are offered by the model system. However, the indications are not so clear-cut as to reject the conventional linear model as an inferior formulation. In fact, the coefficients of the linear model are consistent with those of the two equations in the two-stage model system when viewed collectively. The same conclusions have been obtained from a similar analysis of shopping trips.

TABLE 4 PREDICTION RESULTS

Model	Total Trips	Shopping Trips
Linear	0.1297	0.0276
Two-Stage	0.1253	0.0268

The explanatory powers of the two alternative model formulations are evaluated by examining the correlation between the observation and prediction. The correlation is estimated by regressing the observed number of trips on the predicted number of trips. The results, summarized in Table 4 in terms of R^2, indicate that the two formulations have virtually identical fits to the observation, with the two-stage models showing slightly inferior fits for both total person trips and shopping trips. The conventional linear regression models are capable of accounting for as many variations in trip generation as are the more elaborate two-stage model systems.

CONCLUSION

The adequacy of conventional linear regression models in trip generation analysis has been the subject of this study. The following two issues have been addressed as possible factors that may invalidate linear regression analysis: (a) the incompatibility between the continuous, untruncated error term of a linear regression model and the discrete and nonnegative number of trips generated by a household or individual and (b) the possibility of a two-stage decision mechanism in which the choice of making trips at all is first made, and then the number of trips is determined given that trips are made.

Simulation experiments were conducted to address the first issue. In the simulation, trips were generated assuming Poisson distributions. Although the resulting error distributions were heavily truncated, the analysis indicated that model parameters can be consistently estimated and the expected number of trips can be forecast accurately by using the linear model and ordinary least-squares estimation method. The estimated standard errors of model coefficients were biased. The analysis indicated that weighted least-squares could not be applied to the simulated data to solve this problem because of the inaccurate coefficient estimates that the method produced. Further research is needed to identify the reason for the poor performance of weighted least-squares regression.

Two-stage model systems were estimated by using an empirical data set and then compared with linear regression models. The results indicated that the choice of making trips at all and the determination of the number of trips to make are influenced by overlapping, but different, sets of factors. However, the linear regression models offered essentially the same characterization of trip generation behavior as the two-stage models. Furthermore, the explanatory powers of the two alternative model formulations were found to be identical. The two-stage models provided some additional behavioral insights, but failed to show any improvement in fit despite their complex model structure, which involves an increased number of parameters and an elaborate estimation procedure.

The results of this study have indicated that linear regression models of trip generation offer consistent coefficient estimates and produce as accurate predictions as a more complex two-stage model system. The ordinary least-squares estimation is appropriate for generation models of infrequent trips for which the assumptions underlying the estimation method are unlikely to hold. Improvement in trip generation analysis may not be obtained by adopting more complex model systems.

REFERENCES

1. G. V. F. Timmermans. *A Comparative Analysis of Disaggregate Trip Generation Models*. Master's thesis. Northwestern University, Evanston, Ill., 1982.
2. S. R. Lerman and S. L. Gonzalez. Poisson Regression Analysis Under Alternate Sampling Strategies. *Transportation Science*, Vol. 14, 1980, pp. 346–364.
3. J. J. Heckman. Sample Selection Bias as a Specification Error. *Econometrica*, Vol. 47, 1979, pp. 153–161.
4. G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, Cambridge, England, 1983.
5. R. Kitamura and P. H. L. Bovy. Analysis of Attrition Biases and Trip Reporting Errors for Panel Data. *Transportation Research A*, Vol. 21A, 1987, pp. 287–302.
6. D. A. Hensher, P. O. Bernard, N. C. Smith, and F. W. Milthrope. Modelling the Dynamics of Car Ownership and Use: A Methodological and Empirical Synthesis. Presented at the 5th International Conference on Travel Behavior, Aix-en-Provence, France, 1987.
7. F. L. Mannering. Selectivity Bias in Models of Discrete and Continuous Choice: An Empirical Analysis. *Transportation Research Record 1085*, TRB, National Research Council, Washington, D.C., 1985, pp. 58–62.
8. H. S. Levinson. Urban Travel Characteristics. In *Transportation and Traffic Engineering Handbook* (J. E. Baerwald, ed.), Prentice Hall, Englewood Cliffs, N.J., pp. 138–206.
9. R. Kitamura. Lifestyle and Travel Demand. In *Special Report 220: A Look Ahead: Year 2020*, TRB, National Research Council, Washington, D.C., 1988, pp. 149–189.
10. L. P. Kostyniuk and R. Kitamura. Household Lifecycle: Predictor of Travel Expenditure. In *Behavioural Research for Transport Policy*, VNU Science Press, Utrecht, The Netherlands, 1986, pp. 343–362.
11. R. Kitamura and T. van der Hoorn. Regularity and Irreversibility of Weekly Travel Behavior. *Transportation*, Vol. 14, 1987, pp. 227–251.

Interregional Stability of Household Trip Generation Rates from the 1986 New Jersey Home Interview Survey

W. Thomas Walker and Olayinka A. Olanipekun

The geographic stability of trip generation rates is a major factor in determining data collection strategies. Expensive home interview trip diaries need be collected only in specific geographic areas if the resulting trip rates will be different from those of other areas already surveyed. In the fall of 1986, the New Jersey Department of Transportation, through a consultant, conducted a statewide small sample telephone home interview survey. This survey was divided into two independent parts, northern New Jersey and southern New Jersey, each consisting of about 1,400 household interviews. Differences in home-based trip generation rates tabulated for the areas studied, including the urban and rural portions of the southern study area, provide valuable insight into the geographic stability of trip generation rates, because these areas differ significantly in character. In a summary of the results of a comparative trip generation rate analysis for the New Jersey surveys, stratification schemes are tabulated and analyzed to determine the most appropriate basis for making disaggregate trip rate comparisons between regions. Trip rates are also tabulated for the Delaware Valley Regional Planning Commission counties and the remainder of southern New Jersey to facilitate comparisons of trip-making characteristics between these geographic areas. Finally, comparisons between the trip-making characteristics of southern and northern New Jersey residents are made.

The almost universal availability of the automobile has done much to standardize aggregate trip generation rates, although considerable variation in individual household rates still exists. Widely used traffic analysis methods such as the Institute of Transportation Engineers (ITE) *Trip Generation Manual* implicitly assume the interregional transferability of trip generation rates because individual observations of trip making made throughout the United States are averaged and analyzed in cross section (1).

The geographic stability of trip generation rates is a major factor in determining data collection strategies. Expensive home interview trip diaries need be collected only in specific geographic areas if the resulting trip rates will be different from those of other areas already surveyed. Of course, trip generation rates are not the only factor influencing the need for transportation data collection. Transit usage and modal split factors may also vary significantly between study areas, especially if the density and scale development and the type and amount of public transit service also differ. However, the adequate estimation of trip generation rates is a major factor influencing data collection decisions, particularly in regions in which public transportation ridership is relatively insignificant.

In the fall of 1986, the New Jersey Department of Transportation (NJ DOT), through a consultant, conducted a statewide small sample telephone home interview survey. This survey was divided into two independent parts—northern New Jersey and southern New Jersey—each consisting of about 1,400 household interviews (see Figure 1). Because the NJ DOT contract with the consultant did not include tabulation or analysis of the results of the southern survey, Delaware Valley Regional Planning Commission (DVRPC) staff was requested to undertake the activities, for both the New Jersey counties within the DVRPC region and the remaining southern New Jersey counties.

The three tabulations of these rates provide valuable insight into the geographic stability of trip generation rates, because these areas differ significantly in character. Northern New Jersey is part of the New York metropolitan region, with large areas of intensive commercial and residential development. The cities of Newark, Jersey City, and New Brunswick and their suburbs are prime examples of this development. The DVRPC counties are also urban and suburban in character, centered on the cities of Camden and Trenton but with less intensive development patterns than in the north. The remainder of South Jersey is mostly rural in character with smaller cities such as Atlantic City and Vineland.

This paper summarizes the results of the comparative trip generation rate analysis for the New Jersey surveys. Stratification schemes based on family size, income, automobile ownership, and area type (DVRPC region only) are tabulated and analyzed to determine the impact of these input variables on trip making and to identify the most appropriate basis for making disaggregate trip rate comparisons between regions. Trip rates are tabulated for the DVRPC counties and the remainder of southern New Jersey to facilitate comparisons of trip-making characteristics between these geographic areas. Finally, comparisons between the trip-making characteristics found in the southern and northern New Jersey surveys are made.

SURVEY DATA AND STATISTICAL ANALYSIS METHODS

The southern New Jersey survey consisted of 1,413 telephone household interviews taken on Monday through Friday from

Delaware Valley Regional Planning Commission, The Bourse Building, 21 South 5th Street, Philadelphia, Pa. 19106.

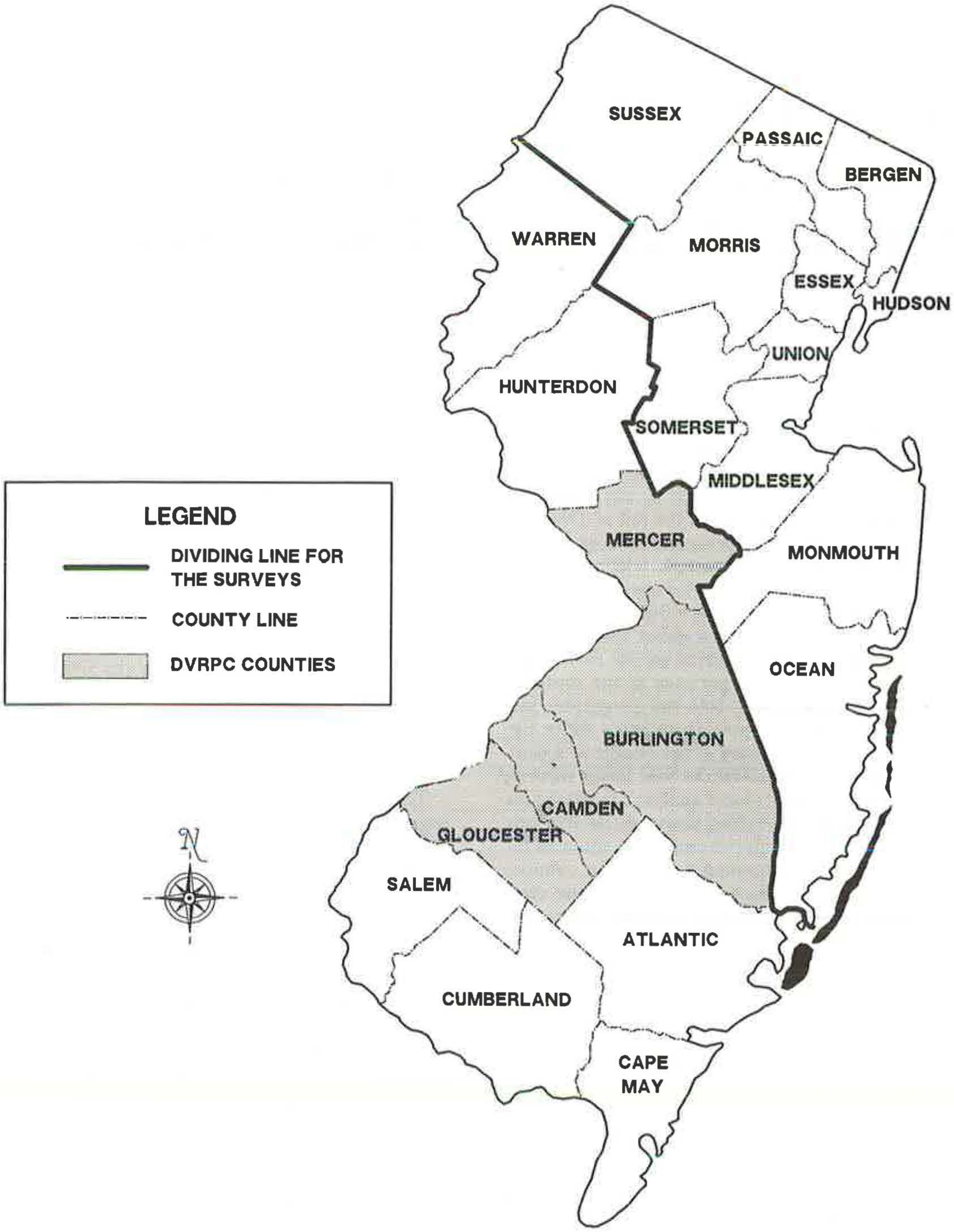

FIGURE 1 Counties included in the southern and northern New Jersey home interview surveys.

October 17 to November 25, 1987. These 1,413 households generated 11,087 weekday trips by all modes, for an average of 7.83 trips per household. Of these 1,413 households, 159 (11 percent) refused to answer the household income question and had to be dropped from the income tabulations. This reduced sample resulted in an overall rate of 8.03 trips per household, about 2.6 percent higher than that for the entire sample. This difference is not statistically significant, however. Rates stratified by automobile ownership and households are based on the 1,413 household sample, and rates for income strata are based on the smaller sample.

The tabulations of trip rates for the northern and southern New Jersey home interview surveys were accomplished by processing the trip data contained in the survey household files. These files contain household-level totals of trip production by purpose as well as the socioeconomic indicators used to allocate the household to a given cell in the cross-classification matrix. In all of the tabulations, rates are reflective of total travel (internal/internal + internal/external). These total trip generation rates are useful in travel simulation and project analyses.

For purposes of trip rate analyses and comparisons, three groupings of the southern New Jersey data were prepared:

1. the entire southern New Jersey study area;
2. the DVRPC region; and
3. the remainder of southern New Jersey.

The first grouping is useful for overall rate tabulations and for comparisons with rates calculated for northern New Jersey. The tabulations for the DVRPC region and the remainder of southern New Jersey are used to compare trip generation rates within the DVRPC counties with those outside the region. Because area type is defined at the census-tract level, tabulations involving area type are confined to the DVRPC region where DVRPC staff have geocoded the trip end addresses to tracts. Outside the DVRPC region only consultant-supplied Minor Civil Division codes are available.

The methodology of trip-generation-rate analysis implicit in the southern New Jersey home interview survey is usually termed the "cross-classification" method. This method is similar to the widely used multiple regression technique in that changes in trip rates are measured when changes in two or more dependent variables are accounted for. In this case, however, an *n*-dimensional matrix of mean or average trip rates is calculated in which each variable (trip purpose, automobile ownership, income, etc.) has at least two subcategories defined by contiguous ranges of the appropriate variable. Cross-classification analysis is disaggregate in that rates are tabulated directly from household data rather than relying on zonal averages of trip rates or independent variables. The use of this technique makes the results comparable with the northern New Jersey trip generation tabulations prepared by the consultant (*2*).

Three statistical indicators are calculated for each trip rate cell in the cross-classification matrix: the mean or average trip rate for households within that stratum; the number of observations; and the cell standard deviation. The primary output is the cell mean trip rate. The number of observations in the cell and its standard deviation provide statistical measures of the accuracy of the rate (via confidence interval) and facilitate hypothesis tests regarding the difference between rates in selected strata or geographic areas. The confidence interval about the mean trip rate is as follows:

$$\overline{X} \pm \frac{s}{(n)^{1/2}} \cdot t_{\alpha}/2,n - 1$$

where

$\overline{X}$ = mean trip rate for cell,
s = cell standard deviation,
n = number of observations, and
$t_{\alpha}/2,n - 1$ = *t*-test statistic (1.960 for 31 or more observations; 12.706 to 2.045 for 2 to 30 observations).

This formula clearly indicates that rate estimation becomes more accurate as the number of observations in the cell increases and decreases as the cell standard deviation grows larger. The *t*-test regarding the statistical significance of differences between two mean trip rates is based on the idea that the hypothesis that trip rates differ must be rejected if their confidence intervals overlap. This leads to the two major statistical objectives in evaluating alternative cross-classification schemes: minimize the standard deviation and maximize the number of observations per cell (at least 30 for practical purposes).

Although more nebulous and difficult to define, rate differences may also be categorized in terms of planning significance. Planning significance is related to the magnitude of the difference more than its statistical significance. For this reason, selected tables also contain estimates of difference and percent difference so that any logical patterns of these differences may be identified. A difference of 1 percent may be statistically significant if the sample is large enough and the mean is tightly constrained by the cell standard deviation. This difference is of little planning significance, however.

On the other hand, a difference of 30 percent is of great planning significance even if not statistically significant, provided that the overall rate patterns are logical and on that basis accepted into the trip generation model. A travel difference of 30 percent may significantly change the design of a proposed facility or even its functional class. We somewhat arbitrarily define a difference of 10 percent or more as being of planning significance, particularly if this difference is part of a logical overall pattern of trip rate variation.

ANALYSIS OF SOUTHERN NEW JERSEY STUDY AREA TRIP PRODUCTION RATES

In general, different demographic distributions of households within alternate geographically defined survey areas may cause average overall trip rates to differ. For this reason, it is desirable to make disaggregate comparisons of trip rates based on demographic variables known to be associated with differences in household trip making. On the basis of past experience in travel forecasting at the regional level and the work of other researchers (*3–6*), the following variables were analyzed as candidate bases for detailed comparison of trip rates:

1. Household size (persons per household),
2. Automobile availability,
3. Household income, and
4. Trip purpose.

Household size is defined as the number of persons occupying a housing unit regardless of the relationship to the householder. Automobile availability is defined as the number of passenger cars available at home for the use of the members of the household. The term "automobile" includes station wagons, vans, and pickups but excludes larger trucks. Income is defined as money received from wages and salaries; nonfarm self-employment; interest, dividends, and net rental; Social Security; public assistance; and all other sources. Trip purpose defines the principal reason for making the trip.

Table 1 presents a percentage breakdown of travel by trip purpose for the southern New Jersey survey. Home-based travel (home-based work, home-based nonwork, and home-based school) together account for 79 percent of total travel generated by residential land uses. As in the northern New Jersey tabulations, home-based nonwork travel excludes school

TABLE 1 SOUTHERN NEW JERSEY SURVEY: PERCENT OF TOTAL TRIPS BY PURPOSE

Trip Purpose	Percent of Total Travel
Home Based Work	27.3%
Home Based Non-Work (excluding school)	40.6%
Home Based Shopping	12.1%
Home Based Social Recreational	8.6%
Home Based Personal Business	10.5%
Home Based Eat Meal	3.6%
Home Based Other	5.8%
Home Based School	11.1%
Non-Home Based	21.0%

	100.0%

TABLE 2 SOUTHERN NEW JERSEY SURVEY: TRIP PRODUCTION RATES BY TRIP PURPOSE AND HOUSEHOLD SIZE

			Household Size				All
Purpose	1	2	3	4	5	6	Households
TOTAL	3.13	6.19	8.07	10.74	12.59	17.26	8.03
T H B	2.23	4.55	6.51	8.78	10.45	13.86	6.34
HBWRK	0.75	1.76	2.58	2.79	3.13	3.79	2.18
HBNWRK	1.41	2.67	3.24	4.40	4.68	7.21	3.29
HBSCH	0.06	0.12	0.69	1.59	2.64	2.86	0.87
N H B	0.90	1.65	1.56	1.96	2.14	3.40	1.69

Abbreviations:

TOTAL:	Total Productions
THB:	Total Home Based Productions
HBWRK:	Home Based Work Productions
HBNWRK:	Home Based Productions Excluding Work and School Productions
HBSCH:	Home Based School Productions
NHB:	Non-Home Based Productions

trips. Commuting travel to and from work accounts for more than 27 percent of travel. Of the home-based nonwork subpurposes, shopping contributes the most trip making (12.1 percent) followed by personal business (10.5 percent), social-recreational (8.6 percent), and eating meals (3.6 percent). All other home-based nonwork nonschool travel accounts for 5.8 percent of total travel. School travel constitutes 11.1 percent of trips generated, and non-home-based travel generates the remaining 21 percent measured in the home interview survey. Overall, these proportions of travel appear to be reasonable.

Average trip rates stratified by purpose and family size are shown in Table 2. Trips per household for all of southern New Jersey, when stratified by family size, range from 3.13 for a household with one person to 17.26 average weekday trips for households with six or more persons. Household trip rates for individual purposes also increase smoothly with household size. Home-based nonwork trip rates sustain the largest absolute increase, with household size increasing by almost six trips per household. Home-based school has the fastest rate of increase, increasing almost 50-fold over the

TABLE 3 SOUTHERN NEW JERSEY SURVEY: TRIP PRODUCTION RATES BY TRIP PURPOSE AND INCOME

	Household Income Code									All
Purpose	0	1	2	3	4	5	6	7	8	Households
TOTAL	3.20	5.21	5.39	7.58	8.66	9.63	10.48	10.75	10.23	8.03
T H B	2.53	4.15	4.25	6.26	6.94	7.57	7.85	8.22	7.88	6.34
HBWRK	0.29	0.96	1.39	2.04	2.48	2.70	3.24	2.97	2.92	2.18
HBNWRK	1.89	2.66	2.21	3.27	3.48	4.04	3.72	3.93	3.81	3.29
HBSCH	0.35	0.53	0.65	0.95	0.98	0.84	0.89	1.32	1.15	0.87
N H B	0.67	1.07	1.14	1.33	1.71	2.06	2.63	2.54	2.35	1.69

Definition of Income Ranges:

Income Code	Definition	Income Code	Definition	Income Code	Definition	Income Code	Definition
0	Under $10,000						
1	$10,000 - $14,999	3	$20,000 - $29,999	5	$40,000 - $49,999	7	$60,000 - $69,999
2	$15,000 - $19,999	4	$30,000 - $39,999	6	$50,000 - $59,999	8	$70,000 and over

TABLE 4 SOUTHERN NEW JERSEY SURVEY: TRIP PRODUCTION RATES BY TRIP PURPOSE AND AUTOMOBILE AVAILABILITY

	Autos Available Per Household						All
Purpose	0	1	2	3	4	5+	Households
TOTAL	3.23	5.34	8.74	10.16	11.12	15.04	7.81
T H B	2.87	4.16	6.82	8.13	9.35	12.22	6.17
HBWRK	0.68	1.27	2.28	3.18	4.30	5.04	2.14
HBNWRK	1.66	2.38	3.50	3.94	3.84	5.91	3.18
HBSCH	0.49	0.52	1.04	1.01	1.21	1.26	0.85
N H B	0.41	1.18	1.92	2.03	1.77	2.83	1.64

range of household sizes. Home-based work and home-based nonwork trip rates also have strong tendencies to increase with household size, with work trips increasing slightly faster than non-home-based trips. Clearly, there is a strong tendency for travel of all types to increase with household size.

The trip rates resulting from preparing similar tabulations stratified by income code are shown in Table 3. Overall, daily trip rates range from 3.20 trips per household for income code 0 (under $10,000) to 10.23 for income code 8 ($70,000 and over). Like the stratification by family size discussed above, trip rates tend to increase in a regular fashion with income, increasing both in total and by trip purpose. However, Table 3 clearly indicates a slower rate of growth in the trip rates as income increases, because incomes vary significantly among individuals. A one-person household and a four-person household may both have an income of $50,000, but the four-person household makes more trips.

Similar tabulations of trip rates stratified by automobile ownership and trip purpose are given in Table 4. Household trip rates increase by automobile ownership as well. The highest total rate of 15.04 (for 5+ car households) is about 4.5 times the rate for 0-car households (3.23). This places the substratum variation in rates for automobile availability between those observed for household size and for income.

STATISTICAL SIGNIFICANCE OF RATE DIFFERENCES BETWEEN NEIGHBORING CROSS-CLASSIFICATION CELLS

Another method for determining the significance of trip rate differences between substrata is to analyze *t*-statistics based on the rate differences between neighboring cells. Table 5 presents the results of this analysis by major trip purpose for

TABLE 5 *t*-TEST FOR STATISTICAL SIGNIFICANCE OF RATE DIFFERENCES BETWEEN NEIGHBORING CELLS

	Persons Per Household				
Trip Purpose	1 v.s 2	2 v.s 3	3 v.s 4	4 v.s 5	5 v.s 6-9
Total Trips	12.38	5.21	5.56	3.02	2.84
Home Based Work	9.73	6.17	1.27	1.51	1.35
Home Based Non-work	7.24	2.32	3.53	0.69	2.33
Home Based School	1.72	7.77	6.84	4.54	0.48
Non-Home Based	5.19	0.52	1.96	0.63	2.02

	Household Income (thousands of dollars)							
Trip Purpose	1-10 v.s 10-15	10-15 v.s 15-20	15-20 v.s 20-30	20-30 v.s 30-40	30-40 v.s 40-50	40-50 v.s 50-60	50-60 v.s 60-70	60-70 v.s 70 & Over
Total Trips	3.85	0.21	4.61	1.71	1.28	0.47	0.53	0.49
Home Based Work	4.01	2.09	3.83	2.98	1.25	2.20	0.88	0.16
Home Based Non-Work	2.13	1.14	3.12	0.67	1.47	0.69	0.37	0.21
Home Based School	0.94	0.57	1.73	0.20	0.88	0.28	1.63	0.60
Non Home Based	1.93	0.31	0.96	2.03	1.41	1.66	0.23	0.51

	Auto Availability Per Household				
Trip Purpose	0 Car v.s 1 Car	1 Car v.s 2 Car	2 Car v.s 3 Car	3 Car v.s 4 Car	4 Car v.s 5 Car
Total Trips	4.21	10.94	2.95	1.50	1.86
Home Based Work	3.15	11.04	6.13	2.99	1.01
Home Based Non-Work	2.62	5.80	1.42	0.11	1.53
Home Based School	0.16	5.78	0.37	0.69	0.18
Non-Home Based	5.51	5.74	0.53	0.61	1.42

household size, income, and automobile ownership. For instance, in Table 5 the *t*-statistic associated with the rate difference in total travel between one- and two-person households is 12.38, and the corresponding value between two- and three-person households is 5.21. Since all cells in Table 5 are based on more than 30 degrees of freedom, the threshold value at 95 percent confidence is 1.96. On this basis, all household size categories have a significantly different rate for total travel. This rate does not apply for work travel, however, where households with three or more persons do not have rates that are statistically significantly different from those in the previous strata. On this basis, three categories of work-trip-related household sizes may be defined as one person, two persons, and three or more persons. However, home-based nonwork and home-based school travel tend to increase significantly in trip rates throughout the range of household sizes. Non-home-based travel tends to separate into three ranges: one- and two-person, three- and four-person, and 5+-person households.

Of the three variables considered, the stratification by income presents the most opportunities to collapse strata, because there is a strong tendency for growth in trip rates to level off and lack statistical significance among the higher income strata (see Table 5). Except for work travel, no income stratum above $40,000 has a trip rate significantly different from that in the next lower income stratum. These higher-income groups tend to be a composite of a wide range of family sizes whose aggregate trip rates may change in response to changes in income levels and the workforce participation rate among women and children. Cross-classification by income and family size will reduce the impact of these factors on rate instability but result in an increase in the number of trip rates to be considered.

Automobile availability categories 3, 4, and 5+ may be

TABLE 6 COMPARISON OF DELAWARE VALLEY TRIP RATES WITH THOSE FOR REMAINDER OF SOUTHERN NEW JERSEY AREA

Study	Persons Per Household						All
Area	1	2	3	4	5	6-9	Households
Total Trips Per Household							
DVRPC Region	3.24	6.30	8.03	10.80	12.31	17.17	8.11
Rest of S. Jersey	2.96	6.02	8.12	10.59	12.95	17.50	7.91
Difference	0.28	0.28	-0.09	0.21	-0.64	-0.33	0.20
Percent Difference	8.6%	4.4%	-1.1%	1.9%	-5.2%	-1.9%	2.5%
"t" Test Statistic	0.77	0.74	-0.14	0.27	-0.65	-0.09	0.59
Home Based Work Trips per Household							
DVRPC Region	0.82	1.89	2.712	2.76	3.25	3.70	2.27
Rest of S. Jersey	0.64	1.54	2.37	2.83	2.98	4.00	2.03
Difference	0.18	0.35	0.34	-0.07	0.27	-0.30	0.24
Percent Differ.	22.0%	18.5%	12.5%	-2.5%	8.3%	-8.1%	10.6%
"t" Test Statistic	1.23	2.25*	1.51	-0.26	0.71	-0.30	2.17*
Home Based Non-Work Trip per Household (Non-School)							
DVRPC Region	1.40	2.66	3.06	4.67	4.65	7.10	3.33
Rest of S. Jersey	1.43	2.68	3.52	3.85	4.72	7.50	3.23
Difference	-0.03	-0.02	-0.46	0.82	-0.07	-0.40	0.10
Percent Differ.	-2.1%	-0.75%	-15.0%	17.6%	-1.51%	-5.63%	3.00%
"t" Test Statistic	-0.12	-0.08	-1.06	1.53	0.11	-0.17	0.49
Home Based School Trips per Household							
DVRPC Region	0.11	0.13	0.65	1.48	2.81	2.83	0.86
Rest of S. Jersey	0.00	0.11	0.76	1.82	2.42	2.92	0.87
Difference	-0.04	-0.02	-0.11	-0.34	0.39	-0.09	0.01
Percent Differ.	100.0%	15.4%	-16.9%	-23.0%	13.9%	-3.2%	-1.2%
"t" Test Statistic	2.07*	0.41	-0.77	-1.43	0.96	-0.10	-0.11
Non-Home Based Trips per Household							
DVRPC Region	0.90	1.62	1.61	1.90	1.59	3.53	1.64
Rest of S. Jersey	0.89	1.69	1.47	2.09	2.83	3.08	1.77
Difference	0.01	-0.07	0.14	-0.19	-1.24	0.45	-0.13
Percent Differ.	1.11%	-4.3%	8.7%	-10.0%	-78.0%	12.8%	-7.9%
"t" Test Statistic	0.05	-0.32	0.50	-0.59	-2.63*	0.35	-0.98

* Indicates Statistically Significant Difference

aggregated into a single 3+ category because the rate difference between these categories, although logical, is not statistically significant except for work travel. Futhermore, the 4 and 5+ automobile categories of households together constitute only 5.7 percent of the sample.

For purposes of comparing trip rates by geographic area, trip purpose and household size were selected as the most appropiate bases for comparison. Trip purpose was selected because of the theoretical importance placed on this variable in most trip generation analyses. Furthermore, trip rates vary significantly by purpose. Household size was selected over income and automobile ownership as the second basis for comparison because trip rates vary significantly across the range of household sizes for all trip purposes except for work and because this variable had the most uniform distribution of home interview surveys to individual subcategories.

COMPARISON OF TRIP RATES BY GEOGRAPHIC AREA

The 1986 home interview survey was unique in that it covered the entire state of New Jersey in two separate surveys. For this reason, it is possible to test the statistical significance of differences between trip rates calculated for the DVRPC counties and the remainder of the southern New Jersey study area and between northern and southern New Jersey.

Southern New Jersey and DVRPC Counties

Table 6 presents a comparison of trip rates stratified by trip purpose and family size between the DVRPC counties (Mercer, Burlington, Camden, and Gloucester) and the remaining counties in the southern New Jersey study area (Atlantic, Cape May, Cumberland, Hunterdon, Salem, and Warren). Table 6 contains the trip rate for each geographical area together with the *t*-test statistic that measures the degree of statistical significance to be attached to the difference and percent difference, also shown. Since all cells in Table 6 have 30 or more degrees of freedom (40 to 408), the *t*-statistic must have a value of 1.96 or greater for the difference between the trip rates to be statistically significant with a 95 percent confidence level.

The *t*-test statistics clearly show that trip rates for the DVRPC counties and the rest of southern New Jersey are for the most part statistically equivalent. Only 4 of the 35 trip rates in this table have significant differences, and these differences tend to be scattered throughout the table and do not appear indicative of a clear pattern.

No trip rate for total travel was found to be statistically different either when stratified by household or in total. The highest *t*-value for the total trip purpose was 0.77, which does not even approach the value needed for statistical significance (1.96). Two rates for home-based-work travel were statistically different: the rate for two-person households and for total work travel. The work trip rates for the DVRPC region were 18 percent higher for two-person households and 10 percent higher for all households (about 0.2 trip per household per day). The rate for school trips for one-person households in the DVRPC region was significantly higher because of an obvious deficiency in the data for the remainder of southern New Jersey (no observed travel). The rate differences for two-, three-, four-, and five-person households, although large in absolute terms (14 to 25 percent), lack the rational pattern required for planning importance and are of no statistical significance. Similarly, the non-home-based rate for the DVRPC region was significantly lower for five-person households but irrational in pattern because the five-person rate was lower than the four-person rate.

In summary, the principal difference between trip rates for the DVRPC counties and the rest of southern New Jersey is in work trips. DVRPC counties have a higher rate for four out of six household strata. These rates probably resulted from the higher labor participation rate in the DVRPC counties—1.25 employed residents per household versus 1.16 employed residents per household in the remainder of the study area. Table 7 contains a comparison of home-based-work rates based on employed residents per household for the DVRPC counties versus the rest of southern New Jersey. When stratified in this manner, the statistical significance of the rate difference disappears. It is interesting to note that the corresponding value from the 1980 census was about 1.54. The reasons for the difference may involve seasonal variations in second jobs (Christmas-related in the New Jersey survey) and the fact that external-local work trips were excluded from the census tabulation. In addition, 1.54 represents a weighted average for the entire Delaware Valley Region. The relatively small sample from urban areas in the New Jersey survey may have underrepresented these areas in the average.

TABLE 7 HOME-BASED WORK TRIP RATE COMPARISONS BY EMPLOYED RESIDENTS PER HOUSEHOLD

	Employed Residents/Household			
Study Area	1	2	3	4
DVRPC Region	1.79	3.59	5.27	7.07
Rest of S. Jersey	1.74	3.47	4.97	7.80
Difference	0.05	0.12	0.30	-0.73
Percent Difference	2.8%	3.3%	5.7%	-10.3%
"t" Test Statistic	0.74	1.17	1.00	-0.90

TABLE 8 COMPARISON OF SOUTHERN AND NORTHERN NEW JERSEY HOME INTERVIEW SURVEY TRIP RATES

Study	Persons Per Household						All
Area	1	2	3	4	5	6-9	Households
			Total Trips per Household				
Southern Jersey	3.13	6.19	8.07	10.74	12.59	17.26	8.03
Northern Jersey	3.17	6.33	7.91	10.62	12.19	17.47	7.85
Difference	-0.04	-0.14	0.16	0.12	0.40	-0.21	0.18
% Difference	-1.3%	-2.3%	2.0%	1.1%	3.2%	-1.2%	2.2%
"t" Test Statistic	-.17	-.51	.39	.23	.51	-.10	.78
			Home Based Work Trips Per Household				
Southern Jersey	0.75	1.76	2.58	2.79	3.13	3.79	2.18
Northern Jersey	0.76	1.80	2.66	3.13	3.10	4.59	2.27
Difference	-0.01	-0.04	-0.08	-0.34	0.03	-0.80	0.09
% Difference	-1.3%	-2.3%	-3.1%	-12.2%	1.0%	-21.1%	4.1%
"t" Test Statistic	-.01	-.37	-.54	-1.98*	.11	-1.38	1.20
			Home Based Non-Work Trips per Household (non-school)				
Southern Jersey	1.41	2.67	3.24	4.40	4.68	7.21	3.29
Northern Jersey	1.37	2.68	2.91	3.69	4.78	6.51	3.01
Difference	0.04	-0.01	0.33	0.71	-0.10	0.70	0.28
% Difference	2.8%	-0.4%	10.2%	16.1%	-2.1%	9.7%	8.5%
"t" Test Statistic	0.25	0.05	1.22	2.13*	0.20	0.56	2.12*
			Home Based School Trips per Household				
Southern Jersey	0.06	0.12	0.69	1.59	2.64	2.86	0.87
Northern Jersey	0.03	0.13	0.55	1.52	2.24	2.88	0.74
Difference	0.03	-0.01	0.14	.07	0.40	-0.02	0.13
% Difference	50.0%	-8.3%	20.3%	4.4%	15.2%	-0.7%	14.9%
"t" Test Statistic	1.03	-.29	1.48	.45	1.35	-.04	2.13*
			Non-Home Based Trips Per Household				
Southern Jersey	0.90	1.65	1.56	1.96	2.14	3.40	1.69
Northern Jersey	1.00	1.72	1.80	2.27	2.06	3.49	1.82
Difference	-0.10	-0.07	-0.24	-0.31	-0.49	-0.09	-0.13
% Difference	-11.1%	-4.2%	-15.4%	-15.8%	-22.9%	-2.6%	-7.7%
"t" Test Statistic	-.75	-.45	-1.24	-1.38	-1.34	-.12	-1.41

* Indicates Statistically Significant Difference

TABLE 9 VARIATION OF HOUSEHOLD TRIP RATES BY TRIP PURPOSE AND AREA TYPE FOR DVRPC REGION

Trip	Area Type				All
Purpose	Urban	Suburban	Rural	Open Rural	Households
HBW	2.17	2.25	2.30	2.30	2.25
HBNW	2.72	3.24	3.13	3.70	3.17
HBSCH	0.57	0.76	1.26	0.78	0.85
NHB	1.66	1.56	1.55	1.83	1.57
TOTAL	7.13	7.82	8.24	8.61	7.84

Southern Versus Northern New Jersey

Trip rates similar to those described above were tabulated for the northern New Jersey study area by the consultant and presented in a report (*2*). The northern New Jersey data were collected by the same consultant using identical sampling and telephone interview techniques. Table 8 compares trip rates by purpose and family size from the southern New Jersey survey with the corresponding value from the northern New Jersey home interview survey. The cutoff value for statistical significance here is 1.96 as well, because the degrees of freedom in the *t*-statistic range from 99 to 808.

Overall, these comparisons (Table 8) show very little difference in trip rates between the surveys. Only four cell values were significantly different. Total trip rates were virtually identical. In aggregate, work trip rates were 4 percent higher in the northern survey, particularly for 4- and 6+-person households (12 and 21 percent, respectively). Home-based nonwork and home-based school production rates were generally higher (8.5 percent and 14.9 percent, respectively) in the southern survey, and non-home-based rates were 7.7 percent higher in the northern survey. However, these comparisons are generally lacking in statistical significance. Only the four-person households for home-based work and nonwork and overall rates for home-based nonwork and school trips were statistically different.

Although these statistical comparisons indicate that few significant differences in trip rates occur between the northern and southern study areas and between the DVRPC counties and the remainder of southern New Jersey, geographic variations in trip rates occur within each study area. The study areas considered are large diverse heterogeneous mixtures of land uses, including numerous urban, suburban, and rural areas. Because of this, the trip rates analyzed are averaged over diverse land uses and area types, which may mask significant geographical variations in trip-making patterns within each study area.

For instance, area type was available as a basis for stratification within the DVRPC counties, and the resulting trip rates are shown in Table 9. Although lacking in statistical significance because of small sample sizes in urban areas, this pattern of trip rates clearly shows a logical increase in trip rate as the density of development declines for home-based work, nonwork, and total travel. This is thought to occur because walk travel is omitted from the trip diaries except for work trips. The long distances associated with rural travel make walking less feasible. There is no consistent pattern for home-based school or non-home-based trips by area type. Other researchers have observed significant differences between urban and rural trip rates (*6–8*).

CONCLUSIONS

Little difference in terms of trip rates was found between the DVRPC counties and the rest of southern New Jersey or between the northern and southern New Jersey study areas. Size-based rates were statistically significantly different between the DVRPC counties and the rest of southern New Jersey in only 4 of 35 households. The principal difference from a planning perspective is related to work trips, which usually had a higher rate per household in the DVRPC region (10 percent higher in total). This difference, which may have resulted from the higher labor participation rate in the DVRPC counties, disappears when work trip rates are stratified by employed residents per household.

The principal differences in trip rates between the northern and southern New Jersey surveys were in home-based nonwork, school, and non-home-based travel. Residents in the southern study area more frequently made home-based nonwork (8 percent) and school trips (15 percent), whereas northern New Jersey residents made 8 percent more non-home-based trips. In total, both work and nonwork trip rates showed no statistically significant variation between the northern and southern study areas.

However, this result should be qualified by the small sample sizes and the large heterogeneous nature of the study areas considered. There is some evidence within the DVRPC counties that trip rates vary significantly by area type, with urban rates being lower than suburban and rural rates. This variation results from the higher tendency to make walk trips in large urban areas. The small sample associated with urban land uses made it difficult to draw strong statistical conclusions in this regard, however.

REFERENCES

1. Institute of Transportation Engineers. *Trip Generation Manual*, 4th Ed., Prentice-Hall, Englewood Cliffs, N.J., 1987.
2. Barton-Aschman Associates. *Preliminary Analysis for the Trip Production Model*. New Jersey Department of Transportation, Trenton, 1987.

3. P. R. Stopher and K. G. McDonald. Trip Generation by Cross-Classification: An Alternative Methodology. In *Transportation Research Record 944*, TRB, National Research Council, Washington, D.C., 1983, pp. 84–91.
4. K. G. McDonald and P. R. Stopher. Some Contrary Indications for the Use of Household Structure in Trip-Generation Analysis. In *Transportation Research Record 944*, TRB, National Research Council, Washington, D.C., 1983, pp. 92–100.
5. *Trip Generation Analysis*. FHWA, U.S. Department of Transportation, 1975.
6. P. M. Allaman, T. J. Tardiff, and F. C. Dunbar. *NCHRP Report 250: New Approaches to Understanding Travel Behavior*. TRB, National Research Council, Washington, D.C., 1982, 147 pp.
7. D. R. Martinson. *A Practical Approach to Trip Generation Analysis for a Multi-County Region*. Master's thesis. Marquette University, Milwaukee, Wis., 1974.
8. L. R. Goode and C. L. Heimbach. Evaluation of the Transferability of Trip Generation Models from One Urban Area to Another. In *Transportation Research Record 931*, TRB, National Research Council, Washington, D.C., 1983, pp. 120–125.

Regional Travel Forecasting Model System for the San Francisco Bay Area

HANNA P. H. KOLLO* AND CHARLES L. PURVIS

A regional travel forecasting model system update using a 1980 data base is reported. Use of the 1981 Bay Area travel survey and the 1980 census Urban Transportation Planning Package is described in terms of providing a data base for model estimation and validation. Historical model development efforts in the Bay Area are compared with current efforts. The demand model development process is characterized as a six-step process involving development of component models and the subsequent packaging into an aggregate forecasting system. The MTCFCAST-80/81 forecasting system involved reestimation of all model components. Simplifications to the original MTCFCAST system were introduced where warranted; the structure of the mobility and work trip models was tampered with the least. In contrast, the work-trip mode choice model was expanded to distinguish between two-occupant and three-plus-occupant carpools, in support of travel forecasting for high-occupancy-vehicle lane projects. Continuity is seen as the key to maintaining and updating regional travel demand model systems.

The use of travel demand models in transportation systems analysis has found widespread acceptance among metropolitan planning organizations (MPOs) across the United States. Typically, the focus of model development activities has been on the estimation and validation of individual model components, particularly the work-trip mode choice model. Less attention is generally given to the "packaging," or combination of individual travel demand models into a comprehensive regional travel forecasting model system.

This paper summarizes the modeling system developed by Metropolitan Transportation Commission (MTC) staff to describe base-year behavior and to be used for travel forecasting in the San Francisco nine-county Bay Area. The model system is part of the 1980–1981 model update to best represent recent survey, census, and networks. The system is designed by building on previous modeling efforts in the Bay Area.

The model system described here is called MTCFCAST-80/81. The "80/81" label distinguishes it from the previous version, MTCFCAST, developed from the 1965 data base. It includes a set of worker/nonworker models, two-automobile ownership models, a full sequence of work-trip demand models, and three sets of nonwork demand models, and relies on UMTA's Urban Transportation Planning System (UTPS) for network and trip assignment models. The demand models are implemented in a system written in FORTRAN for mainframe computers.

HISTORY OF MODEL DEVELOPMENT IN THE BAY AREA

This section provides the background for understanding the regional model system and its individual components, which are traced from 1965 to 1980. The earlier models are described briefly to provide a context for the present model system.

Bay Area Transportation Study Commission

Model development in the San Francisco Bay Area dates back to the 1960s when the Bay Area Transportation Study Commission (BATSC) was created by the California legislature to conduct comprehensive transportation studies, prepare a master regional plan, and provide for an ongoing planning program. One of the major undertakings of BATSC was the 1965 Home-Interview Survey. Some 30,000 households were surveyed for their socioeconomic characteristics and their travel diaries. This survey became the backbone of model development and travel forecasting through the 1970s.

The BATSC models, developed in house, were mainly of the traditional aggregate type, characteristic of MPO efforts of that era. The exception was the trip generation research into disaggregate household trip production models (*1*). Eight trip purposes were carried through trip distribution and three into mode split. The models used in forecasting trip generation productions were a mix of zonal linear regression and household trip rates stratified by income and housing structure type. Trip attraction models were of the zonal linear regression type. Both production and attraction models were stratified by land use type. The trip distribution models were of the gravity type with fitted friction factors and balanced attractions through iteration. The mode split model was of a diversion type with transit-to-automobile-travel-time ratios stratified by three residential density ranges at the production end and by central business district (CBD) versus non-CBD at the attraction end (*2*). Networks and assignments were done in the TRANPLAN software with all-or-nothing loadings.

Regional Transit Travel Projections Project

The second generation of Bay Area travel models was developed by a consultant for Bay Area Rapid Transit and MTC in 1973 as part of the Regional Transit Travel Projections Project. The purpose of these models was to produce forecasts for five transit corridor-planning projects. The models were based on the 1965 data base and can best be described

*Deceased.

Metropolitan Transportation Commission, MetroCenter, 101 8th Street, Oakland, Calif. 94607.

as "aggregate stratified." Several stratification levels by household types were carried through mode split. Employment type and density levels were also used at the attraction end. The trip generation production was a household cross-classification model. Trip attraction used trip rates by employment type. Trip distribution was a gravity model. Mode split was a modified logit in which the parameters were estimated by trial and error to fit the aggregate data rather than by statistical estimation. FORTRAN software was written for the demand models and the TRANPLAN package was used for the networks and assignments.

Travel Model Development Project

An evaluation of the MTC modeling needs was undertaken by MTC management in 1974. Other regional agencies and transit operators were sympathetic toward a quantum jump in the state of the art of travel forecasting. It was decided to put the region in the forefront and have a commitment to a continued effort in model development. The Travel Model Development Project (TMDP) was initiated, and a consultant team was selected to carry out a two-phase study. Phase 1 was devoted to review of data bases, a comparison of model systems, and the preparation of a work program for Phase 2. Thus, the third cycle of model development in the Bay Area was under way in 1975. The 1965 data base was revised and reexpanded, networks were converted to the UTPS, and extensive use of disaggregate logit models was made. The demand models included 21 components covering four trip purposes that were packaged in a system, written in FORTRAN, and fully compatible with UTPS. The forecasting version of the model system, known as MTCFCAST, used market segmentation by three-income or three-automobile-ownership groups. The models were complemented by UTPS network and trip assignment procedures. The models are documented in a three-volume final report (*3*). Summary reviews of the original MTCFCAST travel forecasting model system were conducted by Ruiter and Ben-Akiva (*4*) and Ben-Akiva et al. (*5*). Transportation planning textbooks by Manheim (*6*), Meyer and Miller (*7*), and Ben-Akiva and Lerman (*8*) provide highlights of the Bay Area forecasting system.

Several versions of MTCFCAST have been used in the Bay Area for specific studies. These include the Santa Clara Valley Corridor Evaluation, the Air Quality Plan Update, and the Guadalupe Corridor Alternatives Analysis. Each version incorporates some type of refinement of the MTCFCAST system. Some refinements were the reestimation or replacement of the work-trip mode choice model, recalibration of the trip distribution model, and/or the aggregate validation of the models to a 1975 data base. Several versions of the models have been applied to average values of zonal variables without market segmentation by income, automobile ownership, or any other stratification. This was done in conjunction with work-trip, person-trip tables derived from the traditional gravity or FRATAR trip distribution models. Two examples of this are the model application in 1977 to generate travel forecasts for the Air Quality Management Plan by MTC staff, and the Guadalupe Corridor Alternatives Analysis mode choice model application in 1984 to generate travel forecasts for the Fremont-South Bay Phase I Corridor Study, by a consultant.

PREPARATION OF 1980–1981 DATA BASE

Base-year data are an important component in any travel model update. A great deal of effort was expended in securing the best 1980 data possible. The effort included the acquisition of 1980 census and Association of Bay Area Governments (ABAG) demographic, land use, and employment data. The key to a meaningful model update was the collection, preparation, and use of a special travel pattern data base.

By the end of the 1970s, the 1965, 1970, and 1975 data bases had been exhausted. In particular, the age of the 1965 travel survey had called into question its reflection of present travel behavior in light of major changes to transit service in the region. On the other hand, fiscal constraints against large-scale surveys dampened the desire for travel data updates. Thus, the concept of a small sample survey became appealing, especially when the new breed of disaggregate models was thought to require fewer data for their development. Experience with the Bay Area disaggregate models developed from the 1965 data indicated that a rich aggregate data base was necessary for base-year validation in addition to a small survey for the estimation of model coefficients. All these factors prompted MTC to embark on a new survey to coincide as closely as possible with the 1980 census journey-to-work questionnaire for compatible disaggregate and aggregate travel data sets.

1981 Household Travel Survey

The 1981 household survey was conducted in the spring of 1981 by telephone with a sample of about 6,200 households and their trip diaries (*9*). The sample was of a stratified type selected disproportionately throughout the region. About one-half of the surveyed households were residents of San Francisco County, at a sampling rate of 1.0 percent. The other eight counties had a sampling rate of 0.2 percent. Beyond this sample control total, households were selected by using telephone directory-based random digit dialing in such a way that unlisted households could be selected.

Extensive preparation and analysis of the survey data were undertaken by MTC staff. This included data cleanup, trip linking, household weighting, trip expansion, and reporting of key data (*10–12*). The survey was assumed to represent 1980 travel behavior and was therefore expanded to total 1980 households in the region. Because of the disproportionate nature of the sample, this expansion was necessary to weight the survey observations. It was also necessary for the development of aggregate nonwork models. Master files of household and trip characteristics were prepared to provide a common and easy-to-access data base for the development of individual component models.

1980 Census Urban Transportation Planning Package

The 1980 census provided valuable aggregate data, at the census-tract level, of household characteristics derived from the weighted sample or from the 100 percent counts. In addition to the standard files and reports, the Urban Transportation Planning Package (UTPP) for the nine-county region was purchased from the Bureau of the Census to be the basis for tract-level work-trip locations. Responses to the journey-to-work questionnaire were collected from a sample of 1-in-6 and coded to a geography ranging from tract to county for a reduced sample of 1-in-12 by the Bureau of the Census. The main data files included the number of workers by place of residence reporting their mode of travel to work destinations. MTC staff processed the data and converted the information to aggregate home-based work-trip tables as follows:

1. For unallocated place-of-work data, the census-reported geography of "place" (mainly for Sonoma and Napa counties) and "county" were allocated to the 550-regional-zone system. This entailed detailed analysis of ABAG's employment data to form the basis for judgments about the allocation of workers to zone of work.
2. The tract-zone-county commuter data were aggregated to 550 zone matrices by drive alone, shared-ride passengers in two-occupant vehicles, shared-ride passengers in three-or-more-occupant vehicles, and transit passenger modes.
3. The 1981 survey data were analyzed to produce county and superdistrict estimates of home-based work trips per employed person.

The factors from item 3 were used to convert the census commuter matrices to home-based work trips by mode. The results are called the 1980 "observed" trip tables and form the most reliable aggregate data base available in this region. These tables were the main source for base-year model system validation.

DEMAND MODEL DEVELOPMENT PROCESS

With the 1980–1981 data base on hand, the task of demand model updating was undertaken in house. The objective was to develop a bank of model components that could be packaged in various combinations for various uses. Past experience with model development and application indicated that the update should build on the disaggregate model structure of the earlier Travel Model Development Project. The disaggregate models are considered to be the most advanced and to have more behavioral content than other model types. Although the main framework of the earlier effort was used, many changes were introduced to improve the models and to simplify the process. These included changes to specification of variables, component linkages, and emphases by trip purpose. The linkages in the work models were those least tampered with, whereas those of the nonwork models were substantially changed. The feedback loops from nonwork to work-trip models were removed, the structure of the nonwork distribution models was changed to the gravity form, and the only logit form used in nonwork models was for mode choice. The idea was to keep the main structure of work-trip models and to introduce warranted simplifications wherever possible.

The model development process covers two domains. The first includes the individual components and the second contains the model system. Six distinct steps in the development process span the two domains. Model specification, estimation, and disaggregate validation produce a candidate component model. Market segmentation, software preparation, and aggregate base-year validation are used to package the components into a forecasting model system.

Development of Component Models

Component models perform individual functions in the model system and were therefore uniquely treated in the update process. The six sequential steps of model development mentioned above may not apply to all components; in addition, there are varying degrees of partial or complete recycling throughout. The terminology and the process are geared more to disaggregate models than to aggregate components because the latter require fewer steps than the former.

Component model development is described in the following sections.

Model Specification

Model specification advances a hypothesis about the representation of the phenomena being modeled. It requires the identification of the component function and the dependent and explanatory variables and the selection of a mathematical form for the model. The function may pertain to such factors as automobile ownership prediction, trip attraction estimation, and mode choice simulation. Different combinations of socioeconomic variables, transportation level-of-service variables, or urban growth density variables have been used for different components. Four mathematical forms have been used. Linear regression is used for trip generation production models. Trip rates are used for some attraction models. Logit is used for automobile ownership, work trip distribution, and all mode choice models. Finally the gravity type is used for nonwork trip distribution models. Model specification applies equally to disaggregate and aggregate models.

Coefficient Estimation

Coefficient estimation is the process of applying the observed behavior reflected in the data base to the hypothesis advanced in the previous section. It uses statistical data-fitting techniques to quantify the relationship between the dependent and the independent variables. It produces the coefficients and constants of linear regression or logit utility functions. For aggregate models, *calibration* of gravity model friction factors is a more conventional, yet analogous, term.

Estimation is done by preparing special input files for use in special packages like SAS (multivariate analysis), LOGIT (maximum likelihood logit estimation), or AGM-UTPS (gravity calibration and application program). The resulting coefficients are reviewed for correct sign, reasonable size, and statistical significance. The results suggest either recycling through the previous step or acceptance of a candidate model for subsequent testing steps.

Disaggregate Validation

Disaggregate validation is unique to disaggregate models and involves applying the estimated coefficients to a sample of households from the 1981 survey to simulate their choice behavior. The predicted choices are compared with the reported choices to detect any biases by several socioeconomic stratifications. The results may suggest recycling back to specification, estimation, or acceptance of the component model for subsequent testing.

Development of the Regional Model System

Development of the regional model system was begun after selecting the candidate component models. At this stage, both the disaggregate and the aggregate components were in their semifinal versions. The three steps that compose the regional model system development are described briefly as a continuation of the discussion in the previous section.

Market Segmentation

Market segmentation involves adaptation of the disaggregate model coefficients for forecasting by market segment. In the conventional aggregate model systems, average zonal values are used in forecasting. In the MTC model system, the use of disaggregate models is accompanied by a number of stratifications in which group averages of household characteristics are used instead of zonal averages. The process involves analyzing the variables used in each component model to ascertain the need for revising the input zonal averages to reflect a market segment or to compute market-segment-specific coefficients based on the regional or county variations of household characteristics by market segment. The segmentation varies by component or group of component models. In total, the following segmentations are used: households with workers versus all households; primary workers versus secondary workers; three-income groups; and three-automobile ownership levels.

Software Preparation

Software preparation consists of revising, rewriting, or inserting a special code in existing programs to implement each component model equation on a particular computer. Each component model is implemented in one or more data processing "steps" by one or more data processing programs written in FORTRAN and compatible with the UTPS software. In the MTC model system update, most of the computer programs were rewritten to accommodate the new 1980–1981 models. Although the same framework, style, and file-naming conventions were used, consolidation of a number of steps and programs was undertaken to improve efficiency.

Aggregate Base-Year Validation

Aggregate base-year validation involves simulation of the 1980 base-year travel through the model system, comparison of the simulated choices with independent observed estimates, and calibration-adjustment of model constants to reasonably match observed choices or travel patterns. After market segmentation and model implementation in the software, the model system package was run on the 1980 data base to produce a simulation by each component model. The results of each model prediction were analyzed and compared with the most reliable and available 1980 observed data. The analyses led to either a recycling back to the specification-estimation steps or acceptance of the model with or without constant adjustments. Changes to these alternative-specific constants reduce prediction errors in the forecasting process. The errors can be attributed to a number of factors, including weakness in the underlying theory of the model structure, absence of important but unavailable or nonforecastable variables from model equations, biases in survey data, misrepresentation of time and cost level-of-service data, error in the base-year employment data, misrepresentation of captivity to alternative choices or modes, the regional averaging effect in model estimation, and deviation of actual human behavior from rational choices presumed by the models.

To validate the model system in one continuous cycle and at the same time eliminate compounding of errors from one component model to another, a separate analysis was done at the end of each step to validate each model before proceeding to the next.

For work-trip and mobility models, the MTC travel model update effort included several cycles through the three-component model development steps and two full cycles through model system validation. Disaggregate nonwork model components were developed in the same manner as the work-trip models. Aggregate nonwork models were developed in the traditional manner of gravity model calibration. One cycle of base-year aggregate simulation was undertaken. The models were aggregately validated to the 1980–1981 survey trip tables by mode. Although the survey had a small sample resulting in sparse and lumpy trip table entries, it was the only aggregate data base available to which to validate. It certainly was not as reliable as the census journey-to-work tables but appeared to adequately represent aggregate county modal shares.

MODEL CHARACTERISTICS

The 1980–1981 travel model update resulted in a bank of component models and networks to draw on for planning studies and special applications. In particular, the regional MTCFCAST-80/81 is packaged to represent state-of-the-art systems for demand forecasting and, together with its UTPS network package, represents a sophisticated and practical system. Travel demand model components and component linkages are shown in Figure 1. The 24 component models, their acronyms, and mathematical forms are shown in Table 1. The bank of alternative models provides for a number of conventional models (FRATAR, gravity, etc.), which are used side by side with MTCFCAST-80/81 for generating alternative forecasts to assess reasonableness, establish ranges, and bring about acceptability of such forecasts. The objective of this section is to report the highlights of unique characteristics of component demand models, regional model systems, and some recent network representation improvements.

Characteristics of Component Models

For convenience, the demand models are grouped into four functional areas, and their special characteristics are summarized accordingly in the following sections.

Detailed specification and model estimation results are summarized in three MTC technical summaries or working papers (*13–15*). Home-based work trip and mobility models are fully described elsewhere (*13*). Nonwork trip generation

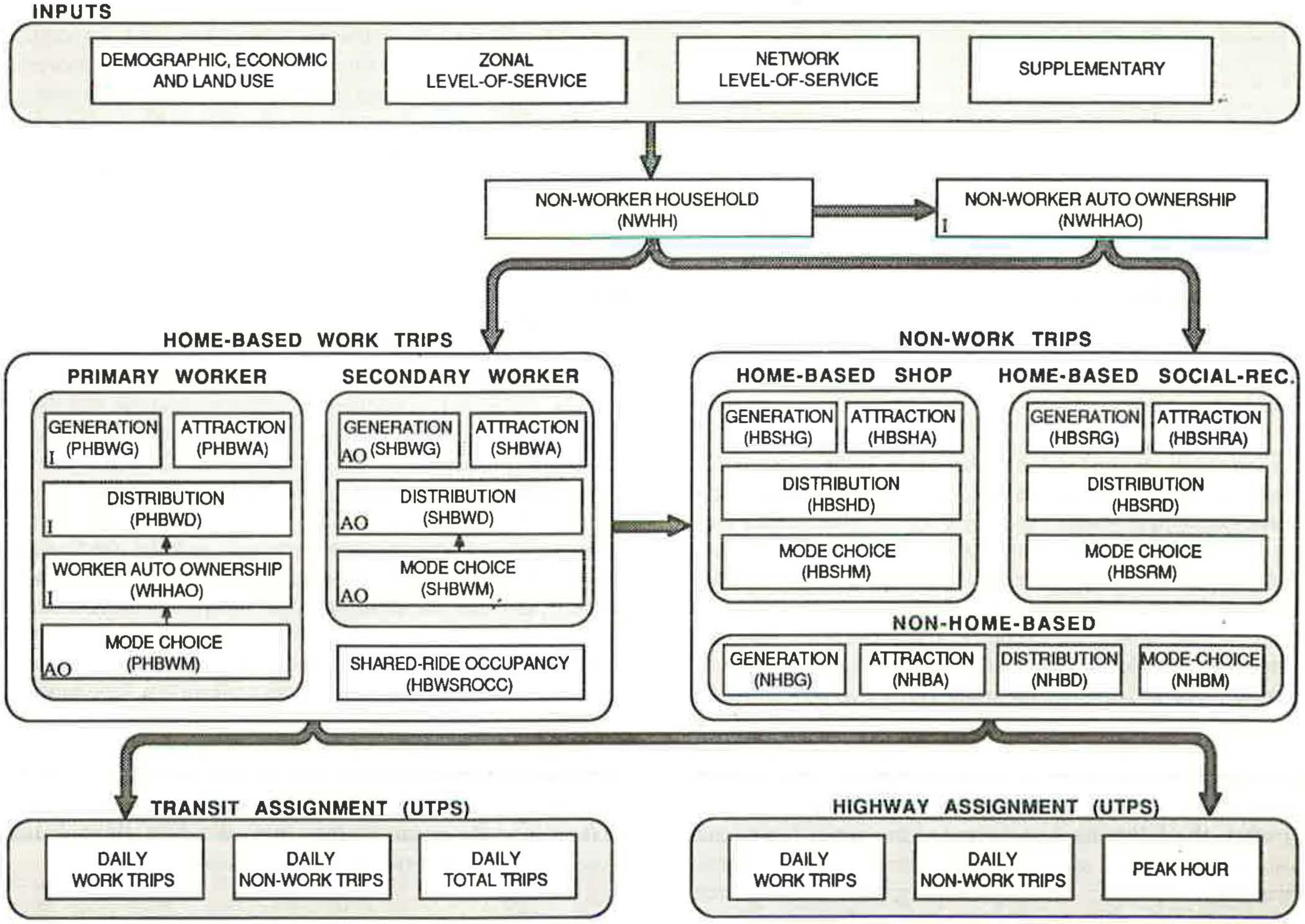

FIGURE 1 Regional travel forecasting model system (MTCFCAST-80/81).

and trip distribution models have been analyzed previously (*14*), as have final specifications for nonwork mode choice models (*15*). Given that the scope of this paper concerns travel model systems rather than detailed model components, estimation results are not presented here.

Mobility Block Models

The mobility block of models consists of the worker-nonworker household, nonworker household automobile ownership, and worker household automobile ownership models. These models use the most predictable socioeconomic, housing type, and density variables available from ABAG, which show a logical relationship to the independent variables they forecast and statistical significance of the estimated coefficients.

Trip Generation-Attraction Models

The trip generation-attraction models for work trips use the most basic units of observation—the worker at place of residence and job at place of work. In addition, the trip generation production models add socioeconomic and density variables to reflect other zonal characteristics.

For home-based nonwork trip generation production models, the most relevant socioeconomic variables (income, household size, and automobile ownership) are used to predict trips per household. For trip attractions, the most relevant sector employment and other variables are used to predict trip attractions. Similarly, a variety of employment sector variables to predict non-home-based trips are used.

Trip Distribution Models

Trip distribution models for work trips are of the logit form and are probabilistic in their destination choice. They incorporate the traditional attraction balancing and trip length matching to observed behavior as well as *K*-factors. Most notably they include composite accessibility variables by mode and automobile ownership. These accessibility variables are derived from lower-level models in the forecasting process and are fed to upper-level models as generalized variables that enable travel time and travel cost by mode to influence trip distribution. They are derived from a rigorous theory consistent with modal and automobile ownership probability.

Mode Choice Models

The mode choice model for work trips is a disaggregate logit model that predicts drive alone, shared ride with two occu-

TABLE 1 DEMAND MODEL SYSTEM COMPONENTS (MTCFCAST-80/81)

Model	Form	Description
1. NWHH	Logit	Worker/Non-Worker Household
2. NWHHAO	Logit	Non-worker Household Auto Ownership
3. WHHAO	Logit	Worker Household Auto Ownership
4. PHBWG	Linear	Primary Worker Home-Based Work Trip Generation
5. PHBWA*	Rate	Primary Worker Home-Based Work Trip Attraction
6. PHBWD	Logit	Primary Worker Home-Based Work Trip Distribution
7. PHBWM	Logit	Primary Worker Home-Based Work Mode Choice
8. SHBWG	Linear	Secondary Worker Home-Based Work Trip Generation
9. SHBWA*	Rate	Secondary Worker Home-Based Work Trip Attraction
10. SHBWD	Logit	Secondary Worker Home-Based Work Trip Distribution
11. SHBWM	Logit	Secondary Worker Home-Based Work Mode Choice
12. HBWSROCC	Linear	Home-Based Work Shared Ride Occupancy
13. HBSHG	Linear	Home-Based Shopping (Other) Trip Generation
14. HBSHA*	Linear	Home-Based Shopping (Other) Trip Attraction
15. HBSHD	Gravity	Home-Based Shopping (Other) Trip Distribution
16. HBSHM	Logit	Home-Based Shopping (Other) Mode Choice
17. HBSRG	Linear	Home-Based Social-Recreation Trip Generation
18. HBSRA*	Linear	Home-Based Social-Recreation Trip Attraction
19. HBSRD	Gravity	Home-Based Social-Recreation Trip Distribution
20. HBSRM	Logit	Home-Based Social-Recreation Mode Choice
21. NHBG*	Linear	Non-Home-Based Trip Generation
22. NHBA*	Linear	Non-Home-Based Trip Attraction
23. NHBD	Gravity	Non-Home-Based Trip Distribution
24. NHBM	Logit	Non-Home-Based Mode Choice

* Aggregate Model

pants, shared ride with three or more occupants, and transit passenger modes. Great care was taken in the development of this model because of its importance for transit planning. The model was estimated from survey samples and aggregately validated to replicate 1980 census journey-to-work modal shares through adjustment of modal constants. The main variables and features of the mode choice model are

1. Socioeconomic variables that are forecast by ABAG (income, household size, workers per household).
2. Automobiles per household forecast internally by the model system.
3. Mode-specific dummy variables.
4. Natural logarithm of total employment density as a continuous variable to reflect CBD characteristics in preference to judgmental CBD geographical definition.
5. Time and cost peak level-of-service matrices segmented to in-vehicle and out-of-vehicle travel time.
6. Two variables to reflect mode of access to transit stations. The first is the automobile access dummy with a 0/1 value to reflect the negative consequence of the automobile access requirement in a transit journey. The second is the household automobile ownership for trips requiring automobile access to transit. The latter has a positive consequence that mediates the negative ones as automobile ownership rises.
7. Stratification by primary and secondary worker and application to segmented person trip tables by three-automobile ownership groups. The mode choice application uses transit level-of-service matrices derived from the walk-only mode of access to transit for households owning no automobiles.

The nonwork mode choice models are simpler logit models yet include socioeconomic variables, total travel times, travel costs, and various employment and residential density variables.

Model System Characteristics

The regional travel model system MTCFCAST-80/81 is a packaged set of component models that convert logit models, developed from sample data, to an aggregate forecasting system (*16*). These new models are integrated with the conventional MPO-type models in a sophisticated process that produces what appears to be a product in a conventional format. The resulting system has the following unique characteristics:

1. The updated models better represent travel behavior through the rich 1980–1981 data base. Coefficients for the entire model system are estimated with large samples and extensive specification and statistical testing. Furthermore, the entire system is validated to the 1980–1981 observed travel behavior from the mobility block through trip assignment.
2. The model system relies on individual representation of modal level-of-service time and cost matrices for a practical representation of modes as well as the upward probabilistic representation of feedback between mode choice, automobile ownership, and work-trip distribution models.

Through the use of logsum variables, the joint decision travel behavior process is correctly represented by the apparent individual decision step of the conventional process.

3. The model system avoids the use of imaginary average traveler behavior through the use of market segmentation. It uses three income groups for the first part of the work-trip model sequence and three automobile ownership groups for the balance. By doing so, the travel decisions of these regions' residents are better represented than with average zonal characteristics.

Network and Trip Assignment Modeling

The MTC model system relies on UTPS for its network representation and trip assignment process. The characteristics of this system are well documented in UMTA manuals on this package. One unique improvement in the UROAD traffic assignment program is the representation of high-occupancy-vehicle (HOV) lanes and subsequent separate trip assignment to mixed flow and HOV facilities. The coding of HOV facilities using a "parallel" approach has been fully tested and implemented by MTC staff using Bay Area networks and trip tables. The results have been encouraging and useful in evaluating the impact of HOV improvement proposals. The parallel coding approach uses separate links for HOV lanes parallel to the mixed-flow adjacent facilities. This allows for coding separate speeds for the two types of facilities. After recycling through mode choice, new speeds are estimated for these facilities using capacity restraint results. Both speed estimation and volume assignment are reported separately to allow for realistic representation of the actual operation of these facilities. The improved coding procedures allow for different definitions of HOV operations in the region. They can be represented as allowing two-or-more occupants or three-or-more occupants in the vehicle.

Transit assignment incorporates two improvements. First is the use of a walk-only transit path in the process. This is done to allow market segmentation in the transit assignment where transit trip tables (out of the mode choice model) for the zero-automobile-ownership group can be assigned only to a path that uses walk-only centroid connectors to transit stations or bus stops.

The second improvement in transit assignment is the prevention of long automobile connectors to a transit station followed by a short hop on a line-haul system to the desired destination. This improvement is done through a series of logical checks to trip tables and network paths to identify unreasonable transit trips, divert them from the automobile-access transit path, and add them to the walk-only path.

CONCLUSIONS

Travel demand forecasting at MTC combines practical needs to provide long-range travel forecasts with the theoretical research and development work associated with disaggregate model estimation. Balancing the practical forecasting aspects with model development research provides Bay Area researchers and planners ample opportunity to test alternative model structures as well as to update the models as needs arise.

The MTCFCAST-80/81 travel forecasting system represents a major effort to build on past model structures with updated data bases. Simplifications to the MTCFCAST system were introduced as warranted except in the case of the work trip and mobility model sequences that had the fewest modifications. New demands on the model system to distinguish between two-occupant and three-plus-occupant carpools led to the estimation of a four-mode home-based work-mode choice model. Previous Bay Area models considered only three modes: drive alone, transit, and shared-ride two-plus occupant.

On the negative side, the sparseness of the 1981 travel survey data base proved to be a challenge in the estimation of disaggregate choice models, especially nonwork mode choice models. Given the overwhelming automobile choice predominance for nonwork trip purposes and the small sample size, the resulting nonwork mode choice models were simple in their final specifications. For example, in-vehicle and out-of-vehicle travel times were aggregated into a generic total time variable given unsatisfactory estimation results when travel times were disaggregated.

Next steps at MTC will include a new household travel survey to coincide with the 1990 census. The sample size of the 1990 survey will be determined in terms of balancing fiscal constraints with the demand for quality data necessary for estimating robust travel demand models. Lessons from developing travel demand models using the 1980 data base will be passed on to the 1990s. Continuity is the greatest challenge and benefit for regional transportation planning agencies charged with the responsibilities of providing skills and tools for travel demand forecasting.

ACKNOWLEDGMENTS

The work reported in this paper was financed in part through grants from UMTA, U.S. Department of Transportation, under the Urban Mass Transportation Act of 1964, as amended; and through grants from the Federal Highway Administration and the California Department of Transportation. The authors are indebted to Patrick Hackett, Senior Analyst of MTC, for providing much-needed computer programming support. Graphics were prepared by Michael Fram. Brenda Germany typed the manuscript.

REFERENCES

1. H. P. H. Kollo and E. C. Sullivan. *Trip Generation Model Development.* BATSC Technical Report 229. Bay Area Transportation Study Commission, Berkeley, Calif., Nov. 1969.
2. H. P. H. Kollo. *Modal Split Model Development.* BATSC Technical Report 227. Bay Area Transportation Study Commission, Berkeley, Calif., Nov. 1969.
3. Cambridge Systematics, Incorporated. *Travel Model Development Project: Phase 2 Final Report,* Vol. 1–3. Metropolitan Transportation Commission, Berkeley, Calif., 1980.
4. E. R. Ruiter and M. E. Ben-Akiva. Disaggregate Travel Demand Models for the San Francisco Bay Area: System Structure, Component Models, and Application Procedures. In *Transportation Research Record 677,* TRB, National Research Council, Washington, D.C., 1978, pp. 121–128.
5. M. E. Ben-Akiva, L. Sherman, and B. Kullman. Non-Home-Based Models. In *Transportation Research Record 677,* TRB, National Research Council, Washington, D.C., 1978, pp. 128–133.
6. M. L. Manheim. *Fundamentals of Transportation Systems Analysis. Vol. 1: Basic Concepts.* MIT Press, Cambridge, Mass., 1979.

7. M. D. Meyer and E. J. Miller. *Urban Transportation Planning: A Decision-Oriented Approach*. McGraw-Hill, New York, 1984.
8. M. E. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, Mass., 1985.
9. M. M. Reynolds, S. M. Flynn, and D. B. Reinke. 1981 San Francisco Bay Area Travel Survey. In *Transportation Research Record 877*, TRB, National Research Council, Washington, D.C., 1982, pp. 51–58.
10. *Sample Weighting and Trip Expansion: 1981 MTC Travel Survey*. Working Paper 4A. Metropolitan Transportation Commission, Berkeley, Calif., 1983.
11. *1980 Regional Travel Characteristics: 1981 MTC Travel Survey*. Working Paper 8. Metropolitan Transportation Commission, Berkeley, Calif., 1983.
12. H. P. H. Kollo and C. L. Purvis. Changes in Regional Travel Characteristics in the San Francisco Bay Area: 1960–1981. In *Transportation Research Record 987*, TRB, National Research Council, Washington, D.C., 1984, pp. 57–66.
13. *Home-Based Work Trip Models—Final Disaggregate Version: Travel Model Development with 1980/81 Data Base*. Working Paper 2. Metropolitan Transportation Commission, Oakland, Calif., revised Dec. 1987.
14. H. P. H. Kollo and C. L. Purvis. *Non-Work Trip Models: Final Version, Technical Summary, Travel Model Update with 1980/81 Data Base*. Metropolitan Transportation Commission, Oakland, Calif., Aug. 1987.
15. *Non-Work Trip Mode Choice Models—Final Disaggregate Version: Travel Model Development with 1980/81 Data Base*. Technical Summary. Metropolitan Transportation Commission, Oakland, Calif., revised Feb. 1988.
16. *Regional Travel Forecasting Model System: MTCFCAST-80/81. Travel Model Development with 1980/81 Data Base, Technical Summary*. Metropolitan Transportation Commission, Oakland, Calif., March 1988.