TRANSPORTATION RESEARCH RECORD

No. 1493

Planning and Administration

Travel Demand Forecasting, Travel Behavior Analysis, Time-Sensitive Transportation, and Traffic Assignment Methods

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1995

Transportation Research Record 1493
ISSN 0361-1981
ISBN 0-309-06170-9
Price: $45.00

Subscriber Category
IA planning and administration

Printed in the United States of America

Sponsorship of Transportation Research Record 1493

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

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

Committee on Passenger Travel Demand Forecasting
Chairman: Eric Ivan Pas, Duke University
Kay W. Axhausen, Moshe E. Ben-Akiva, Chandra R. Bhat, Daniel Brand, William A. Davidson, Christopher R. Fleet, David A. Hensher, Alan Joel Horowitz, Ron Jensen-Fisher, Peter M. Jones, Frank S. Koppelman, David L. Kurth, T. Keith Lawton, Fred L. Mannering, Eric J. Miller, Maren L. Outwater, Ram M. Pendyala, Joseph N. Prashker, Charles L. Purvis, Martin G. Richards, Thomas F. Rossi, Gordon W. Schultz, Peter R. Stopher, Jan van der Waard

Committee on Traveler Behavior and Values
Chairman: Ryuichi Kitamura, University of California
Julian M. Benjamin, Chandra R. Bhat, Elizabeth Deakin, Thomas F. Golob, Konstadinos G. Goulias, David T. Hartgen, Greig W. Harvey, Joel L. Horowitz, Dane Ismart, Peter M. Jones, Lidia P. Kostyniuk, Clarisse V. Lula, Hani S. Mahmassani, Patricia L. Mokhtarian, Alan E. Pisarski, Panos D. Prevedouros, William L. Schroeer, Frank Southworth, Peter R. Stopher, Mary Lynn Tischer, Cy Ulberg, Martin Wachs, Edward Weiner, Jeffrey M. Zupan

Committee on Transportation Supply Analysis
Chairman: Randolph W. Hall, University of Southern California
M. Hadi Baaj, Michael G. H. Bell, David Bernstein, David E. Boyce, Ennio Cascetta, Michael J. Cassidy, Yupo Chan, Teodor Gabriel Crainic, Carlos F. Daganzo, Robert B. Dial, Terry L. Friesz, Michel Gendreau, Ali E. Haghani, Ronald Hagquist, Rudi Hamerslag, Yasunori Iida, Bruce N. Janson, R. Jayakrishnan, Haris N. Koutsopoulos, Hani S. Mahmassani, Chryssi Malandraki, Fred L. Mannering, Mark R. McCord, Eric J. Miller, Anna Nagurney, Srinivas Peeta, Earl R. Ruiter, Yorgos J. Stephanedes, Konstantinos Zograsos

Transportation Research Board Staff
Robert E. Spicher, Director, Technical Activities
James A. Scott, Transportation Planner
Nancy A. Ackerman, Director, Reports and Editorial Services

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1994.

Transportation Research Record 1493

Contents

Foreword

The papers in this volume focus on mode choice models, traffic assignment, hazard models, and various modeling systems to improve urban transportation planning, a model of origin-destination route choice, and trip generation for shopping travel.

A series of papers focuses on residential relocation, impact of mobility policies, household travel behavior, driver's route choice, and multipath transit assignment.

Another set of papers addresses travel time, including commercial vehicle operations, route choice behavior, day-to-day dynamic simulation, data collection procedures, in-time use research, temporal variations on the allocation of time, and travel time uncertainty.

Analysis of the Temporal Transferability of Disaggregate Work Trip Mode Choice Models

DANIEL A. BADOE AND ERIC J. MILLER

An empirical study is presented of the long-range temporal transferability properties within a fixed geographic area of disaggregate logit models of work trip mode choice. The study area is the greater Toronto area, Ontario, Canada. The two temporal contexts are 1964 and 1986, with models estimated from 1964 data being used to predict 1986 travel choices. In addition to the very long transfer period (which does not appear to have been previously examined), a major feature of this study is that a wide variety of model specifications, ranging from the simplest possible market share model to a complex market segmentation model, are tested to investigate the relationship between model specification and transferability. Major findings of the study include (*a*) as in most transferability studies, model parameters are not temporally stable; (*b*) pragmatically the transferred models provide considerable useful information about application context travel behavior; (*c*) in general, improved model specification improves the extent of the model's transferability; (*d*) an important exception to Point c is the complex market segment model, which appears to be "overspecified" and, in the face of changing contextual factors during the 22-year period predicts 1986 conditions quite poorly; (*e*) Point c notwithstanding, simple level-of-service models perform very well in terms of their spatially aggregate predictions (which are often of primary practical importance to planners); (*f*) the models that best fit the estimation context (1964) data do not always transfer the best to 1986 conditions; and (*g*) "transfer scaling," in which modal utility constants and scales are updated, can significantly improve model transferability.

An important expected benefit from use of random utility models in transport modeling is transferability, that is, application of a model to a context different from which it was estimated. This expectation is based on the belief, first, that these models better represent the travel decision-making process and, second, that in the estimated model parameters the values associated with the different socioeconomic classes are built in. Hence, once a model is well specified to capture the decision process in one context, it should be applicable in other contexts so long as the basic nature of the decision-making process remains the same.

Consequently, transferability has been a subject of research interest for the following reasons: first, if it is feasible, the costs and time associated with transport decision-making, in a number of instances, can be reduced significantly; and second, it provides direct evidence of how well models that were estimated in one context perform in forecasting free of errors that would arise from having to forecast explanatory variables, thus making a statement about the range of validity of these models. Several empirical studies have been conducted to assess the effectiveness of model transfer from one context to another (*1–7*). Some of these studies have examined model transfer from one spatial context to another (*1,3,5,8,9*), whereas others have examined the temporal transfer of these models (*2,4,10–12*). The temporal dimension of transferability is the focus of this paper.

The assessment of transferability in the temporal domain has been mixed. The studies reported elsewhere (*2,4,10,12*), even though in some cases they involved simple specifications [e.g., Hensher and Johnson (*2*) used only level-of-service variables in their models], found disaggregate demand model explanatory variable coefficient estimates to show stability and provide a great degree of useful information in the transfer context and concluded that the developed models were temporally transferable. On the other hand, the transferability studies of Talvitie and Kirshner (*3*) and Train (*11*) reject temporal transferability. Train found the forecast errors from transferring estimated models on a pre-BART (Bay Area Rapid Transit) context to a post-BART context to be large and therefore rejected temporal transferability. Train's study, however, was clouded by problems of introduced new modes; therefore his findings are not entirely surprising. Talvitie and Kirshner assessed transferability on the basis of a statistical test of the set of model parameters from the pre-BART context being equal to the set of post-BART model parameters (this included the modal constant terms, which are context specific). As argued by Ben-Akiva (*13*) and Koppelman and Wilmot (*5*), assessing model transferability only on the basis of the set of model parameters being equal in the two contexts is stringent and unlikely to be met because no model is perfectly specified; as a result, all models are in principle context dependent (*6*). A more pragmatic evaluation of transferability is achieved by assessing the extent of useful information provided in an application context by transferred models (*6,13*). This viewpoint for assessing model transferability is also adopted in this paper.

Basically all the temporal transferability studies to date have been limited to short intervening times between estimation and application contexts, where differences in urban conditions between the two contexts are unlikely to be large. Nevertheless, these models are also applied in long-range forecasting in which significant changes in urban conditions occur. Horowitz (*14*, p.145) writes:

> An issue in temporal transferability that has arisen relatively recently concerns whether random utility travel demand models are likely to be transferable over time in periods of significant macroeconomic structural change, such as appear to be occurring now in some Western countries. There is, at present, no empirical evidence on this issue.

This paper examines the temporal transferability of morning peak period work-trip disaggregate multinomial logit mode choice models in the greater Toronto area (GTA), Ontario, Canada. Three travel

D. A. Badoe, Joint Program in Transportation, University of Toronto, 42 St. George Street, Toronto, Ontario, M5S 2E4 Canada. E. J. Miller, Department of Civil Engineering, University of Toronto, 35 St. George Street, Toronto, Ontario, M5S 1A4 Canada.

modes are considered: automobile drive, public transit, and walk modes. The two urban contexts used in the study have an intervening period of 22 years between them, during which significant urban changes occurred. The validity of the assumptions inherent in the use of cross-sectional random utility models in such long-term temporal transfer has not yet been rigorously tested. From a theoretical perspective, Horowitz (*14*) points out that the ability of random utility models to transfer successfully over time during periods of structural change depends on whether such change entails substantial alterations of people's tastes or whether it consists mainly of changes in the attributes of the alternatives people face. In the former case, Horowitz states that it is unlikely that models can be transferred, whereas there is reason for cautious optimism in the latter case. He goes on to state (*14,* p.145):

> . . . if structural change influences mainly the attributes of available travel alternatives, then disaggregate random utility models can be expected to be transferable if they are free of serious specification error and if their explanatory variables encompass all attributes relevant to the choices of interest whose levels change significantly.

The other area in which this paper differs from existing empirical temporal transferability studies is in model specification; a single areawide model specification is not assumed, a priori, to be the most appropriate to capture traveler mode choice behavior. Instead, alternate specifications are explored. [It is noted, for example, that Train (*11*) and Koppelman and Wilmot (*9*) tested alternate specifications; however, the tested specifications had the underlying assumption that all travelers placed the same weight on transport system attributes.] Some of these specifications allow for taste differences among defined subgroups in the travel market. This testing of alternate specifications permits an assessment of the relationship between long-term transfer effectiveness and model specification.

The impact reestimation of modal constants and utility scale parameter has on transfer effectiveness is also discussed, thus allowing comments on whether tastes changed over time in response to significant macroeconomic changes.

The next section of this paper describes the two data sets used for the analysis. The section on comparison of urban structure attributes discusses briefly the differences in urban conditions between 1964 and 1986. The section on model specification presents the alternate model specifications investigated. The section on model estimation results presents the statistical estimation results of the estimation context models. The section on evaluation of transferability presents the results and discussion of the various transferability tests conducted. The impact reestimating modal constants or utility scale parameters, or both, has on transfer effectiveness is discussed in the section on updating constants or utility scale parameter. Finally, the conclusions and findings drawn from this study are outlined.

DATA

The two sources of data for this study are the 1964 Metropolitan Toronto and Regions Transport Study (MTARTS) data base and 1986 Transportation Tomorrow Survey (TTS) data base. The 1964 data were collected in a home interview survey conducted in metropolitan Toronto and its neighboring regions. The total usable questionnaires from this survey totalled 24,000, representing 3.3 percent of all households in the survey area. It provides detailed information on trips and personal characteristics of all household members in the sample. The 1986 TTS data, collected in a telephone interview survey conducted throughout the entire GTA, also provide detailed information on trips and personal characteristics of each household member in the sample. The number of usable household questionnaires totalled 67,000, representing 4 percent of all households in the sampling frame. The data sets do not contain identical information. For example, the 1964 survey collected information on occupation of household members and household income, whereas the 1986 survey did not. These data inconsistencies are considered in model specification. Census data obtained from Statistics Canada are used to augment the travel survey data in the brief descriptive comparison of urban conditions.

All level-of-service data required for model development, with the exception of parking costs and transit fares, were generated using computerized representations of the GTA automobile and transit networks maintained within the EMME/2 modeling system.

The 1964 travel data base is used for estimation of models that are to be transferred. The 1986 data base represents the travel context to which the estimated 1964 models are transferred for evaluation of transferability.

Although automobile passenger, automobile access to transit (park and ride or kiss and ride), and (in 1986) commuter rail modes were also observed to be used by workers in the data bases, these modes were excluded from this analysis to reduce modeling complexity with respect to specification, decision structure (e.g., avoidance of nested decision structures associated with access mode choice), and introduction of new modes (the commuter rail service did not exist in 1964).

COMPARISON OF 1964 AND 1986 URBAN CONDITIONS

Table 1 presents figures on the various characteristics of urban structure in the GTA for 1964 and 1986. The population of the GTA grew from about 2.7 million in 1964 to 4.1 million in 1986, representing a 53 percent increase in the 22-year period, whereas the number of households grew from 0.71 million to 1.47 million, a 106 percent increase. Average household size thus declined from 3.7 persons per household to 2.8 persons per household. A predictable outcome was the increase in percentage of single- and two-person households.

The labor force participation rate for females rose from 45 percent in 1971 to 66 percent in 1986, with the corresponding figures for males being 78 and 81 percent, respectively. This contributed to an increase in the proportion of multiple-worker households. The rate of driver license ownership among female workers also rose from 43.4 percent in 1964 to 77.8 percent in 1986. The corresponding figures for males were 88.6 percent and 93.8 percent, respectively. Private car registration in the GTA rose dramatically from 0.54 million in 1964 to 2 million in 1986, representing close to a 300 percent increase. Household car ownership consequently rose from 0.80 cars per household to 1.4 cars per household in the respective years.

The economic base of the GTA also changed, with the service industry superseding the manufacturing sector as the major employment source for GTA residents. Location patterns for these two industry types are different. The service industry is oriented more to the central business district (CBD), whereas the manufacturing industry, which in 1964 was largely located within the bounds of metropolitan Toronto, is primarily located in the subur-

TABLE 1 Comparison of Urban Attributes of GTA in 1964 and 1986

Attribute	Year	
	1964	**1986**
Population (thousands)	2,657	4,063
Average Weekday Travel (thousands)	3,800	8,800
Average Household Size	3.7	2.8
Private Auto Registration (thousands)	542	1,996
Transit Route Kilometres	953	1345
Transit Vehicle Kilometres (millions)	88	189

Source: TTC Annual Report, 1964 and 1986
Canada Statistics, Road Motor Vehicle Registration
1964 MTARTS and 1986 TTS Travel Survey Data

ban areas of neighboring regions to metropolitan Toronto, where space is available and cheap. Decentralization resulted in the percentage of total population residents in the suburban regional municipalities rising from 34 to 47 percent. These spatial trends in employment and residential locations in turn altered trip distribution patterns within the GTA.

The transport system also experienced expansion. However, the balance of investment was in favor of public transport, which increased its output, measured in transit vehicle kilometers, from 88 million in 1964 to 189 million in 1986.

Average weekday travel in 1964 was about 3.8 million trips, whereas in 1986 this was about 8.8 million trips, representing a 158 percent growth in travel. Notwithstanding the decline in average household size, the number of trips made per household rose from 5.50 to 5.85, and the number of trips per person rose from 1.4 to 2.1. Car use increased by 120 percent from 2.2 million trips in 1964 to 4.8 million trips in 1986. However, the passengers carried in these cars increased less than 60 percent from 0.8 million trips to 1.3 million trips in the respective years, resulting in a decline in the car occupancy rate. Use of public transport increased over 90 percent, from 0.7 million daily trips to 1.35 million daily trips. The average work trip length (Euclidean distance) increased from 7.9 km in 1964 to 11.5 km in 1986, the increase being particularly pronounced for trips by car and transit.

MODEL SPECIFICATIONS

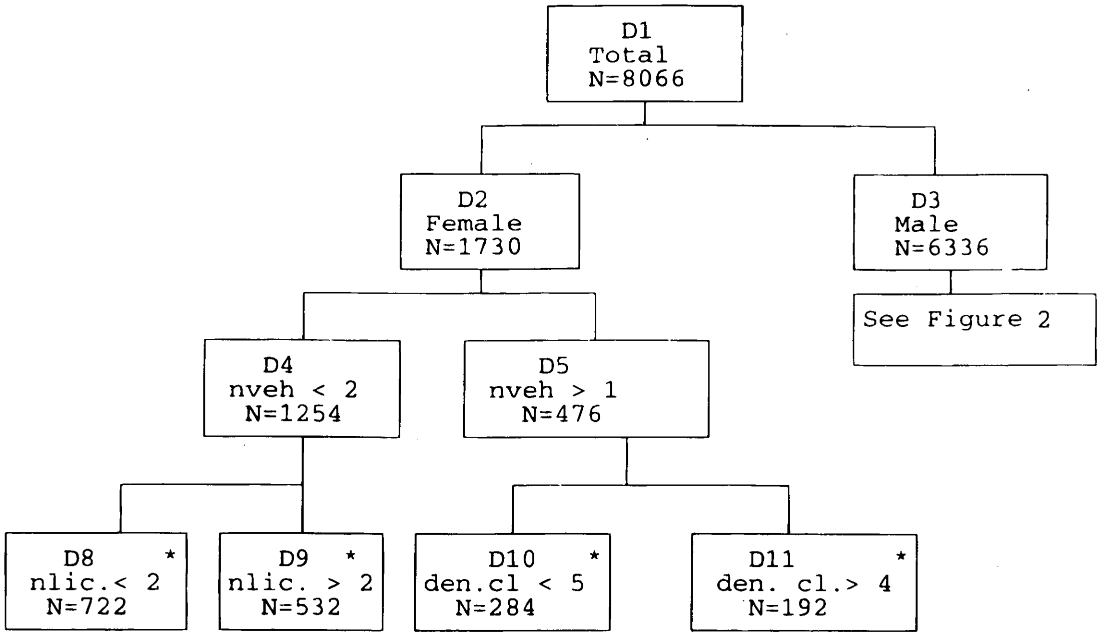
Seven different model specifications are explored. The first is a simple market share model, with the interpretation that each of the considered modes retains its relative share in the forecast context or that no explanatory variables are necessary to explain choice variations in the forecast context. It gives a lower bound on model transfer performance. The second and third models are simple level-of-service models. The first of them treats all the variables, with the exception of in-vehicle costs, as mode specific. The second level-of-service model treats the in-vehicle cost and in-vehicle time of the variables as generic attributes in the automobile drive and transit utilities. Further, it assigns the same importance weight to transit wait time and transit access and egress times. The fourth is termed a fully specified model. In addition to level-of-service attributes, it includes spatial, personal, and household characteristics of the tripmaker.

These four models assume the same coefficient estimates for all travelers in the GTA. The next three model specifications are defined for subgroups of workers that are determined to be relatively internally taste homogeneous. In line with this, the fifth model uses a heuristic segmentation procedure, which essentially consists of applying the automatic interaction detector (*15*) with multinomial logit models to identify 10 multivariately defined market segments with relatively homogeneous tastes. These mutually exclusive segments, which are defined by socioeconomic and spatial variables, are shown in Figures 1 and 2. Simple level-of-service models, similar in specification to the first level-of-service model mentioned, are estimated using data from each subgroup. For a complete description of the segmentation procedure used, see work by Badoe (*16*). Although such an extensive segmentation scheme would not generally be practical in most forecasting applications, it was supportable in this study given the large data sets available. Given this, it was felt that as a research exercise it was worthwhile to explore the impact that multivariate segmentation would have on model performance relative to more conventional nonsegmented or univariate segmentation schemes.

The sixth model takes the first pair of subgroups to emerge from application of the segmentation procedure mentioned earlier to the 1964 data (here the entire sample is stratified by gender to yield subgroups of male and female workers) and estimates models similar in specification to Model 4 mentioned earlier but excluding the gender alternative–specific socioeconomic variables on the two gender worker groups. The seventh model takes the first set of homogeneous subgroups to emerge from application of the segmentation procedure to the 1986 data set. In this case, the worker subgroups were defined according to household automobile ownership level. Models similar in specifications to Model 4 were estimated on the obtained subgroups. The structure of all the models mentioned earlier is the multinomial logit model with three modes: automobile drive, transit, and walk.

MODEL ESTIMATION RESULTS

Table 2 defines the variables included in the models considered. Estimation results for these models are presented in Tables 3 and 4. Most of the models' estimated parameters are statistically well determined and have signs consistent with a priori expectations.



Abbreviations

nveh.	number of vehicles available to household.
den.cl.	trip-end density class (ranges from 1 to 6, higher class numbers mean greater trip-end density).
nlic.	number of persons in household with driver's licence.
orig.	work-trip origin.
dest.	work-trip destination.

FIGURE 1 Multivariately defined market segments, Part 1; selected samples indicated by asterisk.

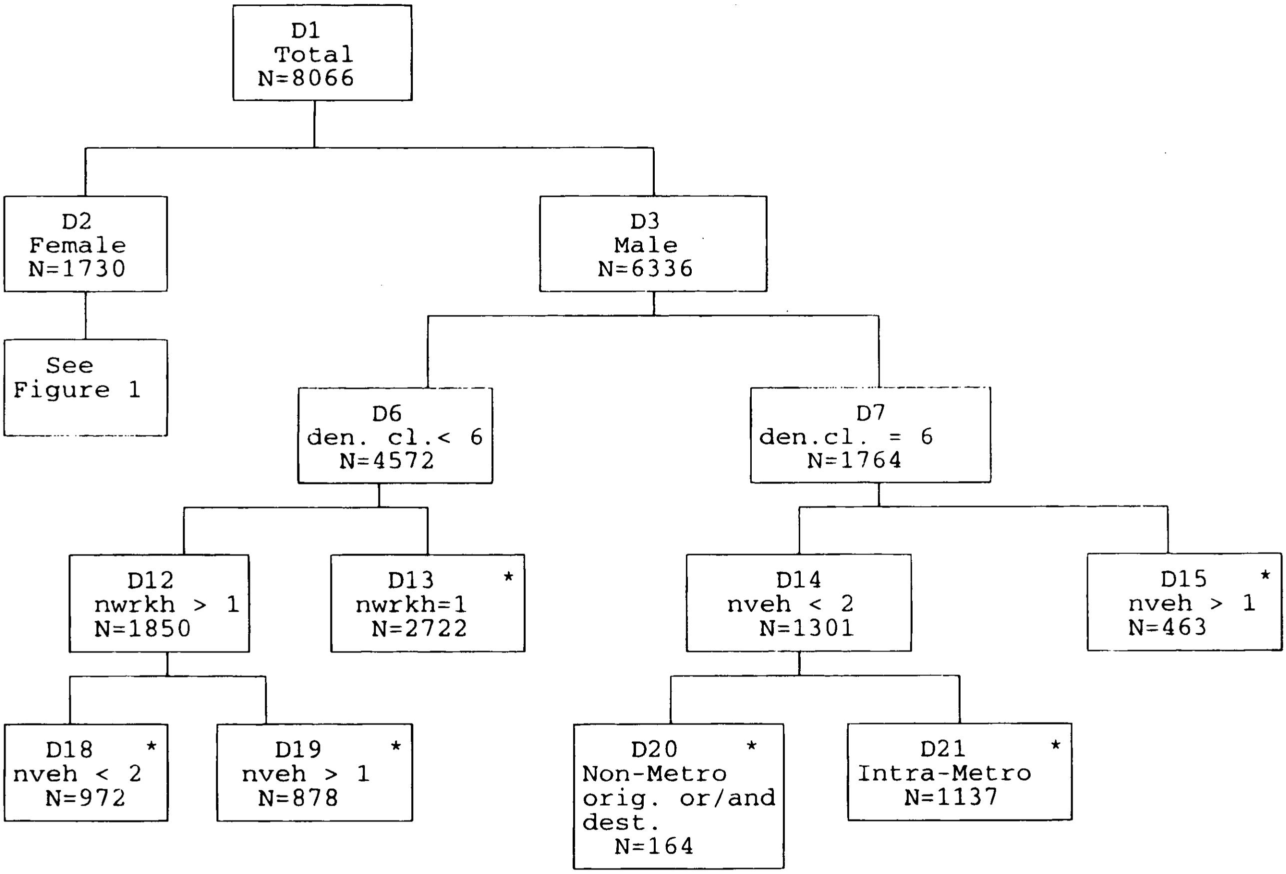
However, the multivariately defined market segment models have some parameters of counterintuitive sign. For forecasting purposes, the affected models are reestimated constraining parameters of counterintuitive sign to 0 values.

Log-likelihood values for these models indicate the multivariately defined segment models to give the best fit to the 1964 data. This is followed by the gender-based models and then the conditional household automobile ownership models, and so forth, ranked according to log-likelihood value. However, when penalty is applied for the number of estimated parameters, as given by the adjusted likelihood ratio index, the gender-based models and the multivariately defined segment models have similar goodness-of-fit.

MEASURES OF TRANSFERABILITY

The following are criteria for judging model transferability:

1. Statistical similarity of estimation and application model coefficients: A nested likelihood ratio test is conducted for this purpose.
2. Ability of the transferred model to replicate individual choice in the application context: In absolute terms, performance here is assessed by the transfer log-likelihood value, which indicates relative disaggregate prediction performance of alternate model specifications. Relative performance measures that indicate how well the 1964 models perform in disaggregate prediction relative to locally estimated similarly specified models on the 1986 data set are provided by the transfer index (TI) and transfer goodness-of-fit measures (*5*). The transfer index has a maximum value of 1.0.
3. Ability of the transferred model to replicate observed aggregate shares: Aggregate predictions of mode use are obtained for seven destination regions of the GTA, which comprise the six constituent municipalities of the GTA, with metropolitan Toronto (by far the largest of the six regions) being split in two, for example, the CBD (Planning District 1) and the remaining districts of metropolitan Toronto. Root-mean-square error (RMSE) and mean absolute error (MAE) values, which are absolute measures of predictive accuracy, are computed using the aggregate predictions to assess forecast accuracy. The relative aggregate transfer error (RATE), which is the ratio of the RMSE value from application of the transferred model in the application context to the RMSE value from a locally estimated model on the application context is also computed. Expressions for these error measures can be found elsewhere (*5*). In addition, 95 percent prediction intervals are constructed (*17*) to determine whether the intervals given by each of these models include the observed mode use for each destination region. The rationale for obtaining confidence intervals is that the 1986 forecasts are based on relationships between dependent and independent variables that are not precise but subject to random errors. A point estimate alone would therefore be suggestive of a precise relationship not subject to random errors. Thus, the observed mode use



Abbreviations

nveh.	number of vehicles available to household.
den.cl.	trip-end density class (ranges from 1 to 6, higher class numbers mean greater trip-end density).
nwrkh.	number of workers in household.
orig.	work-trip origin.
dest.	work-trip destination.

FIGURE 2 Multivariately defined market segments, Part 2; selected samples indicated by asterisk.

value, if the confidence interval bounds it, confirms the appropriateness of the model specification.

RESULTS

Test of Parameter Equality

As in other transferability studies, results of the nested likelihood ratio test of 1964 and 1986 model parameters being statistically equal (Table 5, Column 2) reject the null hypothesis for all model specifications. As discussed earlier, this is not surprising given the errors in the modeling procedure. The emphasis is thus on the more pragmatic measures of transferability assessment reported below.

Disaggregate Measures of Transferability

The transfer log-likelihood values for each model are indicated in the Column 3 of Table 5. The worker mode choice model conditional on household automobile ownership level, with a log-likelihood value of −10,304, is found to give the best disaggregate predictions on the observed 1986 data. This is followed by the gender segment models and then the single areawide fully specified model. The multivariately defined segment models, which gave the best data fit in the estimation context, performed quite poorly yielding a log-likelihood value of −11,366. With this exception, however, improved model specification in general translates into improved disaggregate predictive performance in the application context.

TI values range from 0.132 for the multivariately defined segment models to 0.894 for Level-of-Service Model I, indicating that some of the 1964 models provide a significant component of information obtained from local 1986 models. The less well specified models have higher TI values than the better-specified models. Computed transfer goodness-of-fit measures (Table 5, Column 5), in general, compare favorably with the goodness-of-fit index values given by the locally estimated models on the 1986 data set (Column 6). The negative goodness-of-fit value for the transferred 1964 market share model means its log-likelihood is lower than the log-likelihood given by the 1986 market share model.

TABLE 2 Definition of Variables Specified in Mode Choice Models in Table 3

dauto	= 1 in auto-drive mode utility (modal constant); = 0 otherwise
dwalk	= 1 in walk mode utility (modal constant); = 0 otherwise
aivtt	= auto in-vehicle travel time (min.), enters into auto-drive mode utility; = 0 otherwise
ivtt	= aivtt (auto in-vehicle travel time) in auto-drive mode utility; = tivtt (transit in-vehicle travel time in transit mode utility; = 0 in walk mode utility.
tivtt	= transit in-vehicle travel time (min.) in transit mode utility; = 0 otherwise
twait	= transit wait time (min.) in transit mode utility; = 0 otherwise
twalk	= transit access + egress + transfer time (min.) in transit mode utility; = 0 otherwise
tovtt	= transit out-of-vehicle time (min.) (access+egress+transfer+wait times) in transit mode utility; = 0 otherwise
ivtc	= aivtc (the auto in-vehicle travel costs ($) for auto-drive mode); = 0 in walk mode utility; = tfare (the transit fare) in transit mode utility.
apkcst	= auto daily parking cost ($) in auto-drive mode utility; = 0 otherwise
wdist	= walk distance (km.) in walk mode utility; = 0 otherwise
avplic	= number of vehicles per licensed person in household. Enters into auto-drive mode utility; = 0 otherwise
wcbd	= 1 in walk utility if worker's employment location is in Central Business District; = 0 otherwise
amal	= 1 in auto-drive utility, if worker is male; = 0 otherwise
tcbd	= 1 in transit utility if worker's employment location is in Central Business District; = 0 otherwise
tgend	= 1 in transit mode utility, if worker is female; = 0 otherwise

Aggregate Measures of Transferability

Aggregate transfer measure values based on aggregate predictions given by naively transferred models from the 1964 context are presented in Table 5. Even though RMSE and MAE penalize the prediction error differently, they both indicate that the Level-of-Service Model I, which treats nearly all the system attributes as alternative specific, notwithstanding its simplicity in specification, to yield the best spatial predictions of mode use. Level-of-Service Model II and the worker choice model conditional on automobile ownership level have comparable aggregate forecast performance. The forecast performance of the single areawide fully specified model and the multivariately defined segment models are disappointing given their superior specification to the simple level-of-service models.

RATE values range from 1.0 for the market share model to 5.6 for the market segment models. In general, the better the model specification the higher the RATE value. This is understandable, given the fact that a well-specified model estimated on the 1986 context yields far superior aggregate predictions compared with a poorly specified local 1986 model.

Table 6 is a summary table that shows whether the confidence intervals for predicted mode use by destination region, given by each of the 1964 models, does include the observed mode use. Where the confidence interval given by a particular model type bounds the observed mode use for a destination region, the abbreviation for that model is recorded in the cell described by that mode and destination region. The results indicate that with the exception of the market share model, most models do not yield confidence intervals that include the observed walk mode use for most of the destination regions. This, however, improves for the transit and automobile drive modes. Hamilton is the only region for which the confidence interval given by most of the models bounds the observed mode use for all the three modes. Interestingly this destination region, which is self-contained in terms of trip distribution, underwent comparatively minor change in urban conditions in the 22-year period (*16*). None of the models yields modal confidence intervals that bound the observed for work trips destined to Planning District 1, the downtown district of Toronto. This is disappointing because in any long-range planning mode split of trips to this district would be of considerable interest. The three models here, which are of superior performance compared with the others, are the Level-of-Service Model I, the gender segment models, and the automobile ownership models.

Reestimation of Modal Constants and Utility Scale

Modal constants are in principle context specific because they capture those aspects of the choice process for which the included model explanatory attributes do not account. Hence, their transferability from one context to another is expected to be weak. Thus, these con-

TABLE 3 1964 Model Parameter Estimates

Variables	Market Share	Level of Service (I)	Level of Service (II)	Fully Specified	Male Model	Female Model	0 Veh. Model	1 Veh. Model	2+ Veh. Model
dauto	1.348	0.090 *	-0.133 *	-1.266	-0.583	-2.298		-0.978	-1.636
dwalk	-1.139	0.924	0.626	1.592	1.050	1.927	0.661 *	1.423	1.597
aivtt		-0.031		-0.009	-0.009	-0.014		-0.015	
ivtt			-0.037						
tivtt		-0.043		-0.029	-0.036	-0.011 *		-0.033	-0.030
twait		-0.205		-0.202	-0.209	-0.202	-0.284	-0.182	-0.216
twalk		-0.046		-0.026	-0.037		-0.051 *	-0.032	-0.003 *
tovtt			-0.123						
wdist		-1.961	-1.918	-1.884	-1.675	-2.347	-2.253	-1.758	-1.810
ivtc		-0.389	-0.040 *	-0.388	-0.386	-0.468	-2.190	-0.278	-0.695
pkcst		-0.333	-0.314	-0.282	-0.273	-0.332		-0.317	-0.199
avplic					1.540	3.487		1.523	2.739
amal								0.672	0.649
tcbd					1.252	1.014	0.293 *	1.061	1.559
tgend							0.546	0.973	0.734
wcbd					0.773	0.984		0.773	1.304
No. of Obs.	8066	8066	8066	8066	6336	1730	640	5150	2276
Log-Likelihood at Zero	-5929.6	-5929.6	-5929.6	-5929.6	-4677.6	-1251.9	-434.7	-3836.7	-1658.2
Log-Likelihood at Conv.	-3847.3	-2839.4	-2883.4	-2590.5	-1990.7	-566.1	-165.1	-1785.3	-614.8
Adjusted Likelihood Ratio Index	0.3511	0.5204	0.5134	0.5625	0.5737	0.5452	0.6159	0.5336	0.6274

veh. - number of vehicles available to household.

Note: Parameter estimates with asterisk (*) sign are insignificant at the 5% level.

TABLE 4 Parameter Estimates of Multivariately Defined Market Segment Models

Sample	Parameters								
	dauto	dwalk	aivtt	tivtt	twait	twalk	wdist	ivtc	apkcst
D8	1.280	2.508	-0.051	-0.017*	-0.244	0.012*	-2.666	-0.109*	-0.308
D9	-0.019*	2.563	-0.012*	0.003*	-0.199	0.036*	-2.297	-0.159*	-0.951
D10	-0.951*	0.201*	-0.108	-0.098	-0.175	-0.115*	-2.675	-1.387	-0.282*
D11	-0.233*	0.415*	0.002*	-0.030*	-0.189	0.082*	-1.527	-1.878	-0.281
D13	1.208	1.254	-0.040	-0.038	-0.168	-0.066*	-1.518	0.039*	-0.073*
D15	-1.978	-1.881*	0.006*	-0.039	-0.498	-0.065*	-1.587	-0.916	-0.326
D18	-1.092	-0.463*	-0.023	-0.039	-0.235	-0.121	-1.739	-0.926	-0.089*
D19	2.884	3.862	-0.048	-0.025*	-0.163	0.049*	-2.139	-0.363*	-0.172*
D20	6.333*	5.952*	-0.097*	-0.000*	0.051*	0.038*	-1.171	1.804*	-0.216*
D21	-1.069	1.740	-0.005*	-0.057	-0.093	-0.017*	-2.376	-0.429	-0.337

Note: Parameter estimates with asterisk sign are insignificant at the 5% level.

stant terms are reestimated while the remaining utility function parameters are transferred to explore how well these models would have performed free of these purely contextual parameters. In another scenario, the remaining utility function parameters are rescaled. Rescaling is equivalent to reestimating the variances of the distributions of the random utility components. The necessary mathematics for this can be found elsewhere (*7*). The intent of the analysis here is to investigate whether shifts in constants or scale, or both, are responsible for the models not yielding better transfer performance.

Evaluating transfer assessment measures reported in Table 7 after reestimating the modal constants using information from the application context indicates that this results in considerable improvement in model predictive performance. Across models, the log-likelihood values increase significantly compared with their naive

TABLE 5 Results from Transferring 1964 Models to 1986 Application Context

1964 Model Type	Nested Likelihood Ratio	Absolute and Relative Disaggregate Transfer Measures				Aggregate Transfer Measures		
		$(LL_{86}(\theta_{64})$	TI	$Rho(\theta_{64})$	$Rho(\theta_{86})$	$RMSE_{64}$	MAE	RATE
Market Share	45	-15115		-0.008	0.000	0.35	0.28	1.01
Level of Service (I)	158	-11352	0.894	0.243	0.272	0.09	0.06	1.82
Level of Service (II)	107	-11378	0.891	0.241	0.271	0.12	0.08	2.21
Fully Specified	371	-10787	0.759	0.281	0.370	0.18	0.10	5.21
Auto Ownership	347	-10304	0.787	0.259	0.329	0.13	0.07	4.09
Gender Segments	306	-10494	0.789	0.280	0.355	0.17	0.09	4.92
Market Segments	367	-11366	0.132	0.024	0.179	0.16	0.10	5.57

$$TI = \frac{LL_{86}(\theta_{64}) - LL_{86}(c_{86})}{LL_{86}(\theta_{86}) - LL_{86}(c_{86})}$$

where TI is the Transfer Index Computed with 1986 context Log-Likelihood given by 1986 Market Share model (c_{86}) as Base; and $LL_{86}(\theta_i)$ is the 1986 context log-likelihood using parameters θ_i from year i ($i \in \{64,86\}$)

$Rho(\theta_i)$ is the Likelihood Ratio Index computed with $LL(c_{86})$ using model parameters θ_i from Year i ($i \in \{64,86\}$).

$$RMSE = \left[\frac{\sum_{mg} \frac{(\hat{N}_{mg} - N_{mg})^2}{\hat{N}_{mg}}}{\sum_{mg} \hat{N}_{mg}} \right]^{\frac{1}{2}}$$

where N_{mg} and N_{mg} are the number of persons observed and predicted to choose alternative m from group g respectively.

$$RATE = \frac{RMSE_{64}}{RMSE_{86}}$$

where $RMSE_i$ represents the 1986 context root-mean-square error computed with estimated model parameters from year i (i could be 64 or 86).

TABLE 6 Models Yielding Confidence Intervals That Include Observed Mode Use

Destination Region	Mode		
	Auto Drive	Transit	Walk
PD1			MC
R.O.M	AO, GEN	LOS I,FS, AO, GEN	
Durham		GEN, MS	MC
York	LOS I, LOS II, AO, GEN, MS	LOS I, MS, LOS II, FS, AO, GEN	
Peel		LOS I, LOS II, AO, GEN, MS	
Halton			MC
Hamilton	LOS I, FS, AO, LOS II, GEN, MS	LOS I, LOS II, FS, AO, GEN, MS	MC, AO, FS, LOS I, GEN, MS

Model Definition

MC - Market Share Model
LOS I - Level of Service Model (I)
LOS II - Level of Service Model (II)
FS - Fully Specified Model
AO - Choice Model Conditional on Auto-Ownership
GEN - Choice Model Conditional on Worker Gender
MS - Multivariately Defined Segment Models

Destination Region Definition

PD1 - Planning District 1
R.O.M - Remaining Planning Districts of Metro Toronto Region

transfer performance, particularly so, for the multivariately defined segment models. TI values correspondingly show an increase across specification.

Aggregate error measure values decline for all models and, consequently, the RATE values also decline very significantly. As in an earlier case, the improvement in aggregate performance is particularly pronounced for the multivariately defined segment models. This is because far more model parameters are reestimated for this model compared with the remaining models. Notwithstanding its big improvement, the multivariately defined segments' models yield a lower transfer log-likelihood than the gender or conditional automobile ownership models. Aggregate error measures are smaller though, but this is largely due to the update of several modal constant terms. From a practical viewpoint, updating such a model compared with the others would require substantially far more data and therefore would be unattractive.

Rescaling the model parameters yields additional significant improvement in transfer log-likelihood values, and hence TI, with the TI values for the simple level-of-service models attaining value close to I (Table 7). This would suggest that the scale parameter changed between the two temporal contexts presumably in response to the significant changes in urban character. The implication of these results is that if the constants or scales, or both, can be updated, then existing models estimated on richer data sets, collected at periods when more resources were available, can be employed in forecasting in present-day contexts. This issue is addressed in more detail in a paper by Badoe and Miller in this Record. TI values after reestimation of the modal constants and utility scale parameter do not attain a value of 1 for the better-specified models, which suggest that in addition to scale and constants changing from one context to the other, the underlying utility function parameters also may have changed.

Overall, both disaggregate and aggregate transferability measures show fairly similar trends in results. That is, in general as specification is improved the absolute and relative disaggregate transfer measures show increase, whereas the aggregate error measures show a decline in magnitude.

SUMMARY AND CONCLUSIONS

This paper examines the transferability of disaggregate demand models for a fixed urban area at two points in time—1964 and 1986—with major differences in urban conditions. It also examines the issue of model specification and transfer effectiveness. Alternate model specifications ranging from simple to complex were estimated on the 1964 work trip data, which represented the estimation context. These models were then naively transferred to the 1986 application context for forecast purposes.

Pure statistical tests of model parameters from the two urban contexts being equal reject the null hypothesis of equality, indicating that model parameters have not remained stable over time. Thus, from a theoretical viewpoint, long-range transferability is rejected. However, from a pragmatic perspective, relative measures of trans-

TABLE 7 Measures of Transferability after Updating Estimation Context Models

1964 Model Type	Transferability Measures after Re-estimation of Constants					Transferability Measures after Re-estimation of Modal Constants and Scale Parameter	
	Transfer Log-Likelihood	Transfer Index	Root Mean Square Error	Mean Absolute Error	Relative Aggregate Transfer Error	Transfer Log-Likelihood	Transfer Index
	$(LL_{86}(\theta_{64})$	$TI(c_{86})$	RMSE	MAE	RATE	$(LL_{86}(\theta_{64})$	$TI(c_{86})$
Market Share	-14996		0.35	0.24	1.00	-14996	
Level of Service (I)	-11229	0.92	0.06	0.03	1.20	-11042	0.97
Level of Service (II)	-11143	0.95	0.07	0.04	1.32	-10992	0.99
Fully Specified	-10323	0.84	0.11	0.07	3.03	-10070	0.89
Auto Ownership	-10039	0.84	0.09	0.06	2.71	-9737	0.91
Gender	-10154	0.85	0.12	0.08	3.32	-9894	0.90
Market Segment	-10451	0.57	0.03	0.01	1.11	-9810	0.88

$TI(c_{86})$ - Transfer Index Computed with 1986 Log-Likelihood of Market Share model as Base.

$Rho(\theta_i)$ - Likelihood Ratio Index computed on 1986 Data Using Model Parameters of Year *i*, ($i \in \{64,86\}$).

ferability indicate that the transferred models yield useful information in the application context. TI values show that with the exception of the models of the multivariate segments, transferred models provide at least 76 percent of the log-likelihood provided by locally estimated 1986 models and, with updating, this percent figure rises to 84. RMSE values, which range from 0.09 to 0.18, are comparable to values encountered in the literature for short-range and interurban transferability. RATE values show that use of naively transferred models result in comparatively significant aggregate error, with this error increasing with improved specification. Updating the modal constant terms significantly reduces this aggregate error.

Consistent with the findings of other transferability studies, model transferability is found to improve with improved model specification. However, the best fitting model in the estimation context did not give the best predictive performance on transfer to the application context. The choice models conditional on household automobile ownership level generally give the best model transfer performance, especially when disaggregate measures are used to evaluate transfer performance. However, simple level-of-service specifications appear to be surprisingly robust and performed very well at aggregate levels of typical planning interest. Full market segmentation specifications in the face of changing contextual factors resulted in poor transferability performance. A possible reason for this might be that extensive segmentation resulted in the models being so "trained" to the urban conditions of the estimation contexts that under the major changes that occurred in urban character, the models lost severely in predictive power.

As indicated earlier, reestimation of the modal constants or utility scale translates to significant improvements in model predictive performance, suggesting that these parameters may not have remained stable over time. Where possible, the evidence presented in this work suggests that this subset of parameters (at a minimum the modal constants) be reestimated to enhance transferability.

In sum, the reported results and findings in this work are generally consistent with and support those reported in the literature on short-term temporal transferability analysis.

ACKNOWLEDGMENTS

The work reported in this paper was funded by an operating research grant from the Natural Science and Engineering Research Council of Canada. In addition, funding for a portion of this work was provided by the Ontario Joint Transportation Research Program of the Ministry of Transportation, Ontario, on behalf of the ministry's Travel Demand Research Office. Access to the travel behavior data bases and EMME/2 network modeling system was provided by the University of Toronto Joint Program in Transportation.

REFERENCES

1. Atherton, T. J., and M. E. Ben-Akiva. Transferability and Updating of Disaggregate Travel Demand Models. In *Transportation Research Record 610,* TRB, National Research Council, Washington, D.C., 1976, pp. 12–18.
2. Hensher, D. A., and L. W. Johnson. A Two-Period Analysis of Commuter Mode Choice: The Predictive Capability of Individual Choice Models. *The Logistic and Transportation Review,* Vol. 13, No. 4, 1977, pp. 361–375.
3. Talvitie, A., and D. Kirshner. Specification, Transferability and the Effect of Data Outliers in Modeling the Choice of Mode in Urban Travel. *Transportation,* Vol. 7, No. 3, 1978, pp. 311–332.
4. Silman, L. A. The Time Stability of a Model-Split Model for Tel-Aviv. *Environment and Planning A,* Vol. 13, 1981, pp. 751–762.
5. Koppelman, F. S., and C. G. Wilmot. Transferability Analysis of Disaggregate Choice Models. In *Transportation Research Record 895,* TRB, National Research Council, Washington, D.C., 1982, pp. 18–24.
6. Gunn, H. F., M. E. Ben-Akiva, and M. A. Bradley. Tests of the Scaling Approach to Transferring Disaggregate Travel Demand Models. In *Transportation Research Record 1037,* TRB, National Research Council, Washington, D.C., 1985, pp. 21–30.
7. Koppelman, F. S., G. Kuah, and C. G. Wilmot. Transfer Model Updating with Disaggregate Data. In *Transportation Research Record 1037,* TRB, National Research Council, Washington, D.C., 1985, pp. 102–107.
8. McCoomb, L. A. Analysis of the Transferability of Disaggregate Demand Models Among Ten Canadian Cities. *Transportation Forum,* Vol. 3-1, 1986, pp. 102–107.

9. Koppelman, F. S., and C. G. Wilmot. The Effect of Omission of Variables on Choice Model Transferability. *Transportation Research,* Vol. 20B, No. 3, 1986, pp. 205–213.
10. Parody, T. E. An Analysis of Disaggregate Mode Choice Models in Prediction. In *Transportation Research Record 637,* TRB, National Research Council, Washington, D.C., 1977, pp. 51–57.
11. Train, K. E. A Comparison of the Predictive Ability of Mode Choice Models with Various Levels of Complexity. *Transportation Research,* Vol. 13A, No. 1, 1979, pp. 11–16.
12. McCarthy, P. S. Further Evidence on the Temporal Stability of Disaggregate Travel Demand Models. *Transportation Research,* Vol. 16B, No. 4, 1982, pp. 263–278.
13. Ben-Akiva, M. E. Issues in Transferring and Updating Travel and Updating Travel-Behavior Models. In *New Horizons in Travel-Behaviour Research* (P. R. Stopher, A. H. Meyburg, and W. Brog, eds.), D.C. Heath and Co., Lexington, Mass., 1981.
14. Horowitz, J. L. Random Utility Travel Demand Models. In *Transportation and Mobility in an Era of Transition* (G. R. Jansen, P. Nijkamp, and C. J. Ruijgrok, eds.), Elsevier Science Publishers B.V., Amsterdam, 1985, pp. 141–156.
15. Sonquist, J. A., E. L. Baker, and J. H. Morgan. *Searching for Structure.* Institute for Social Research, University of Michigan, Ann Arbor, 1971.
16. Badoe, D. A. *An Investigation into the Long Range Transferability of Work-Trip Discrete Choice Models,* Ph.D. dissertation. Department of Civil Engineering, University of Toronto, Ontario, Canada, 1994.
17. Daganzo, C. F. The Statistical Interpretation of Predictions with Disaggregate Demand Models. *Transportation Science,* Vol. 13, No. 1, 1979, pp. 1–12.

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

Trip Generation for Shopping Travel

KWAKU AGYEMANG-DUAH, WILLIAM P. ANDERSON, AND FRED L. HALL

The effect of the geographic location of households on weekday, home-based shopping trips in the greater Toronto area (GTA) is reported. Five zones within the GTA were chosen to reflect different types of location and accessibility. An ordered response model, which maintains the ordinal nature in trip-making decisions, was used in the analysis. The statistical results show that, after controlling for a household's socio-demographic characteristics, a household's location within the metropolitan area has some effect on its weekday, home-based shopping trip generation. In particular, households located in the older urban area are likely to make fewer trips than those living in the suburbs.

The relative importance of discretionary travel (defined as all non-work travel for shopping or social or recreational purposes) has grown over the years and has also captured the attention of both policy makers and transportation demand modelers. In large metropolitan areas, the ratio of discretionary trips to mandatory trips (work and school) is often greater than 1 (*1*). In the Transportation Tomorrow Survey (TTS) in the greater Toronto area (GTA) in 1986, 68 percent of all household trips were for discretionary purposes. The National Personal Transportation Survey in the United States indicated that the number of discretionary trips grew faster than the number of work trips between 1977 and 1988, with discretionary trips making up three-fourths of all household trips in 1988 (*2*). A recent study in the regional municipality of Ottawa-Carleton, Canada, showed that shopping, leisure, and social trips accounted for more than 52 percent of total trips (*3*).

Despite their sheer volume, discretionary trips have been treated crudely in most operational models. For instance, one way to estimate the number of discretionary trips is by applying a constant factor to the number of work and school (mandatory) trips. Discretionary travel, however, may have different temporal and spatial patterns than mandatory travel. Studies on work and school trips focus on maximum peak periods because their purpose is primarily to aid in facility design. The bulk of discretionary trips, however, take place after the morning and evening rush periods when most work trips are over (*4*). Compared with work and school trips, the number of discretionary trips may be more sensitive to such factors as the cost of travel, accessibility, or the land use pattern, all of which tend to vary spatially within a metropolitan area.

In light of the ongoing shift in the focus of transportation planning from plans to build more infrastructure to plans aimed at modifying travel behavior, the development of better models of discretionary travel should be high on the transportation research agenda. The purpose of this paper is to start moving toward improved trip generation models for discretionary travel that are more responsive to locational factors.

The rest of the paper is organized as follows. The next section provides the background for the study. It contains a review of the Urban Transportation Modeling System (UTMS) methods for trip generation analysis and of some past studies of the relationship between trip frequency and the location of trip makers. The next sections discuss the following: (*a*) the data and the rationale for selecting the locations used; (*b*) a brief description of the analytical method used and how it addresses the weaknesses identified in the UTMS approaches; (*c*) a discussion of the statistical results; and (*d*) conclusions.

Department of Geography, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada.

STUDY BACKGROUND

Three things are discussed in this background section: current modeling approaches; the nature of explanatory variables currently used for discretionary trip generation; and recent studies that directly address the relationship between trip generation and location.

Modeling Approaches

Regression models and category analysis are the two main methods used for trip generation in the UTMS. Regression models treat the number of trips generated per household (or individual) as a linear function of a set of explanatory variables. Category analysis divides households into categories on the basis of a cross classification of their characteristics and applies a constant trip generation rate for each category. Both methods have a number of shortcomings.

One problem with the standard regression model is the lack of any built-in upper limit to household trips as the values of explanatory values, such as household size and vehicle ownership, increase. There is also the possibility of the regression models predicting negative trips. In an attempt to deal with these problems, the regression model is sometimes given a probabilistic interpretation. Greene (*5*) has noted, however, that such a model can predict probabilities greater than 1 or less than 0.

The difficulty with category analysis is the lack of any effective way to choose the best groupings of household characteristics and hence the best categories. One way is to minimize the standard deviations among the categories. In situations in which there are many variables and hence many categories, this involves extensive trial and error. Hutchinson (*1*) describes a study by Vandertol using trip data from Hamilton, Ontario, Canada, that produced wide margins of error for households within various categories. The error margins range from 10 percent of the average trip rate for one-worker households to 37 percent for four-worker households. (Although the analysis was based on work trip data, it illustrates the problem of defining the best categories.) Another drawback of category analysis is the lack of inferential statistics. In the absence of such measures, there is no way to assess the statistical significance of the explanatory variables in trip generation.

A problem with both models is that they treat the number of trips per household as a continuous dependent variable. One can of course make a statistical defense of this, but to develop a behavioral basis for trip generation, the dependent variable must be discrete rather than continuous. One possible solution to this problem is to use the poisson regression model in place of the linear regression model. The poisson regression model has been shown to be appropriate in applications to count data, especially when the count for some observations is small or 0 (*6*). An alternative solution is to use one of the family of discrete choice models, which are based on a probabilistic theory of choice among a finite set of options.

Additionally, there is a definite order to the trip-making decision. If a person makes two trips, that person also necessarily makes one trip. The ordinal nature of the trip-making decision is not, and cannot be, captured by either of the regression or category approaches or by the Poisson regression model. The ordered categorical property of the outcomes of the trip-making decision makes it imperative to look for an alternative approach that can exploit the ordering of the information. The ordered response model, a type of discrete choice model that maintains the ordinal nature in the dependent variable in situations in which there are more than two responses, is therefore the best candidate for trip generation analysis. This approach is adopted in this study.

Nature of Explanatory Variables

The types of explanatory variables that are usually used in regression models and category analysis are either the socioeconomic characteristics of households within a zone (for example, income, car ownership, family size), or if these are not available, the characteristics of the zone itself (for instance, population and employment densities). Although cost of travel, accessibility or locational factors have been identified as influencing travel decisions, they are generally excluded from operational models. Ortuzar and Willumsen (*7*) report that attempts to incorporate accessibility measures into UTMS trip generation models have been unsuccessful, noting that the accessibility index is either nonsignificant or has the wrong sign in regression models.

A good indication of the range of explanatory variables currently in use is found in the extensive compilation of trip generation rates by ITE in 1987 (*8*). For generation of shopping trips from residential neighborhoods, for instance, regression models included in the ITE report use household size, the number of vehicles, and the number of dwelling units as explanatory variables. (The report does not distinguish between mandatory and discretionary trips so it is assumed that the model is applicable to all types of trip.) There is a suggestion in the ITE report that location might affect trip generation, but that is not explicitly followed up in the regression result tables. ITE noted that

> dwelling units that were larger in size, more expensive, or farther away from the central business district had a higher trip generation rate per unit than those smaller in size, less expensive, or closer to the CBD. However, other factors, such as geographic location and type of adjacent and nearby development, also had an effect on the trip generation rates. (*8*, p. 256)

The ITE trip generation rates employ adjustment factors for household size, vehicles owned, and density (dwelling units per acre). Although density might be correlated with distance from the central business district, the regression models used to produce the ITE trip generation rates do not take explicit account of location within the city.

Past Studies

Very few studies investigate the relationship between observed trip frequency and location within the city. Two studies that do are reviewed here. The first one is a study carried out in the Canadian regional municipality of Ottawa-Carleton (*3*). The objective of the study was to explore the observed relationship between transportation, land use, and the environment. (The review of the IBI study in this paper concentrates only on the relationship between trip rates and location within the study area.) The study region was divided into nine distinct areas according to similarity in land use mix and density patterns. The major conclusion is that there are no significant differences in the total trip (both work and nonwork) generation rate among the nine areas. The mean daily trips per person range from 2.57 to 3.11 in the nine areas.

This conclusion may be questioned on several grounds. For example, the dispersion of trip rates within areas may vary more than the mean number of trips. Additionally, a different conclusion might have been reached if separate analyses were done for mandatory and discretionary trips because the former is fairly inelastic to locational factors whereas the latter may not be. Thus the results do not really exclude the possibility of some variation in trip-making behavior over space—especially for discretionary trips.

Friedman et al. (*9*) examined trip frequency in older neighborhoods and the newer suburbs in the San Francisco Bay Area. Using 1980 travel data, the study revealed that the number of total trips per household in the two areas differs significantly: 9 and 11 trips for the older neighborhoods and the suburbs, respectively. The study failed to address the following two questions: To what extent does household size correlate with suburban living? Is the difference in trip frequency associated with differences in car ownership in the two areas? (This last question is important because the researchers reported marked differences in mode split for the two areas: 86 percent of trips were by automobile in the suburbs versus 64 percent in the older neighborhoods.) For this reason it is impossible to determine whether the results indicate a "pure" locational effect on trip generation or simply reflect differences in household characteristics over space.

This review has shown that there are problems with the existing approaches to modeling trip generation and that the results of studies on the trip generation-location relationship are inconclusive. The analysis that follows constitutes an attempt to address some of the methodological problems mentioned above and to provide new empirical evidence on the effect of locational factors on discretionary trip making.

DATA

The data for the analysis were obtained from the Transportation Tomorrow Survey conducted between September and December 1986 by the Joint Program in Transportation Studies, University of Toronto, and supported by the Ontario Ministry of Transportation. The TTS was a telephone interview survey of a random sample of 1.5 million households in the GTA, Canada. Completed, usable surveys were obtained for 61,453 households. The GTA is an expanded metropolitan definition that contains 3 of the 25 census metropolitan areas in Canada: Toronto census metropolitan area (CMA) and two contiguous CMAs: Oshawa and Hamilton. For the purpose of the TTS, the GTA was divided into 46 macrozones.

The survey collected data on the sociodemographic characteristics and weekday travel patterns of households. Household characteristics of interest for this analysis are household size, which is the number of persons in the household; the number of household members who are fully employed outside the home, who are employed part time outside the home, who are working at home, and who are unemployed; the number of children (under 16 years); the number of vehicles and the zone of residence. Unfortunately, the survey does not include information on household income or occupation. Census data on income for 1986 for each zone are available in 10 categories. These data cannot be used for detailed analysis, however, because an income level for each household is needed. Data on average zonal income, which are also available, were considered too gross and therefore unsuitable to use in the analysis. The total number of weekday, home-based shopping trips made by automobile and transit is used to calculate household trip generation rates. There were no walk (shopping) trips in the data set for the five zones studied. It was decided to include only home-based shopping trips in the analysis to allow a more direct behavioral interpretation of the results than would be possible with a broader definition of discretionary trips.

The use of observed trip rates raises the question of latent shopping travel demand. This is particularly so when the data used in the analysis were collected for one weekday, despite the fact that many shopping trips take place at the weekend. The unavailability of weekend shopping trip data, however, is less of a problem given that the goal of this study is to search for improved trip generation models rather than to predict the total number of shopping trips.

The use of weekday, home-based shopping trips raises the question of whether these trips constitute a major proportion of total shopping trips. TTS Report 5 provides a table of total (weekdays) shopping trips from each zone. As indicated in Table 1, home-based shopping trips are a relatively constant fraction of the total number of weekday shopping trips. Home-based shopping trips as a percentage of total shopping trips vary from 59 to 66 percent in the five zones, confirming the importance of home-based trips and the need to study them. One should note, however, that shopping trips are defined in the TTS report as all trips that have their destination purpose as shop. It is not clear whether this definition includes trip chains.

The principal hypothesis of this study is that geographic location is an important factor in determining trip generation rates. One simple reason is that location affects people's accessibility, defined as the ease of travel between one point and a set of other points. Ideally, household accessibility measures would have been included in the analysis. However, the data for calculating the accessibility indexes were not readily available. The location of each household in one of five zones within the urban area is therefore used as a proxy for accessibility—although it may also reflect other spatially variant factors such as "lifestyle" differences. Five zones were chosen to reflect different types of location and accessibility (Figure 1).

TABLE 1 Comparison of Weekday Home-Based and Total Shopping Trips per Household

Zone	Home-based	Total	%*
1	0.18	0.30	60
2	0.16	0.27	59
3	0.31	0.47	66
4	0.31	0.49	63
5	0.35	0.57	61

*percentage of weekday home-based shopping trips to total number of shopping trips reported.

Two zones (1 and 2) are within the older urban area and are well served with public transportation, including buses, trolleys, and subways. A third zone is in the inner postwar suburbs, and it is also well served by the transit system. Zones 4 and 5 represent locations that are recently developed suburbs superimposed on older towns. Each of the last three zones has good expressway access. Zones 4 and 5 have, in addition, a network of rural roads but relatively poor public transportation service. The total number of households interviewed in the TTS in the five planning zones were 10,867. Table 2 gives a profile of the five zones.

ORDERED RESPONSE MODEL

The model presented in this section is similar in structure to the probit model developed by McKelvey and Zavoina (*10*) for the analysis of Congressional voting on the 1965 Medicare Bill and by Bhat and Koppelman (*11*) for modeling household income and employment, but with a different set of assumptions. The ordered response model is an extension of the better-known binomial and multinomial logit models. The binomial logit model is used to predict the probability that a categorical variable will take on one of two possible values. In this case it does not matter whether the variable is measured on an ordinal or a nominal scale. The multinomial logit model predicts probabilities for three or more values that a categorical variable can take on. In this case, it is assumed that the variable is measured on a nominal scale. (A common application is the choice among three or more travel modes.) The ordered response model is appropriate when the categorical variable takes on three or more possible values that are subject to some logical ordering. For example, the categorical variable may be successive levels of educational attainment, ratings from an opinion survey, or employment status (unemployed, part-time employed, and full-time employed.) The number of trips generated from a household is clearly such an ordinally scaled categorical variable.

The ordered response model is based on the definition of an abstract score for each household, which can be interpreted in this application as the utility derived by a household from making shopping trips.

$$U_n = V_n + \epsilon_n \qquad (1)$$

where

U_n = "total" utility that household n derives from making trips,
V_n = systematic or "observed" utility, and
ϵ_n = random component.

The V_n is defined as a linear function of attributes of the household:

$$V_n = \beta X_n \qquad (2)$$

where β and X_n are, respectively, a vector of parameters and a vector of household attributes used as independent variables. (A more general specification would include attributes of the choice alternatives in X; however, no such attributes were employed in this analysis.) The random component is the part of the utility that is unknown to the researcher. It reflects the idiosyncrasies and tastes that vary randomly for each household together with the effect of omitted variables or measurement errors (*12*). The ordered response model assumes "local" instead of "global" utility maximization. Local utility maximization implies a choice situation in which each binary

FIGURE 1 Five macrozones in GTA.

decision consists of whether to accept the current value or "take one more" (*13*). The decision maker stops when the first local optimum is reached. Global utility maximization occurs when all alternatives in the choice set are simultaneously considered. The ordered response model was chosen over the ordered generalized extreme value model of Small (*14*), which maximizes global utility because of its simple mathematical structure, which makes it more convenient for applied analysis.

The model also defines a set of "cut points" associated with each of the possible outcomes. For example, suppose a household can make 0, 1, 2, . . . , J trips, where J is a maximum defined through inspection of the data. Define a cut point λ_1 such that household n will make zero trips if U_n is less than λ_1, or in probabilistic terms

$$P_{n0} = Pr(\beta X_n + \epsilon_n \leq \lambda_1) \tag{3a}$$

where P_{n0} is the probability that household n makes zero trips. The probability that the household makes one trip is now defined as the probability that U_n is greater than λ_1 but less than a second cut point λ_2:

TABLE 2 Profile of Five Zones

Indicators	Zone 1	Zone 2	Zone 3	Zone 4	Zone 5
No. of hld*	3210	3723	930	2410	594
Children**	0.46	0.30	0.66	0.79	0.74
Avg. hld size	2.70	2.13	3.22	3.17	3.06
Vehicle/hld	1.11	1.05	1.60	1.76	1.86
Avg. trip***	0.18	0.16	0.31	0.31	0.35

* total number of households interviewed
** average number of household members who were under 16 years
*** average number of home-based shopping trips per weekday
hld = household

$$P_{n1} = Pr(\lambda_1 < \beta X_n + \epsilon_n \leq \lambda_2) \quad (3b)$$

or, more generally,

$$P_{nj} = Pr(\lambda_j < \beta X_n + \epsilon_n \leq \lambda_{j+1}) \quad \text{for } j = 1, \ldots, J-1 \quad (3c)$$

and

$$P_{nj} = 1 - Pr(\beta X_n + \epsilon_n \leq \lambda_J) \quad (3d)$$

Because it is not possible to observe the values of the random components ϵ_n, the empirical model is derived by making an assumption about their distribution. The random components are assumed logistically distributed:

$$F(\epsilon_n) = 1/[1 + \exp(-\mu\epsilon_n)] \quad (4)$$

where μ is a positive scale parameter that is unobservable; therefore it is assumed that $\mu = 1$. Given these assumptions, an explicit form for Equation 3a can be written:

$$P_{n0} = 1/[1 + \exp(\beta X_n - \lambda_1)] \quad (5a)$$

$$P_{n1} = 1/[1 + \exp(\beta X_n - \lambda_2)] - 1/[1 + \exp(\beta X_n - \lambda_1)] \quad (5b)$$

$$P_{nj} = 1/[1 + \exp(\beta X_n - \lambda_{j+1})] - 1/[1 + \exp(\beta X_n - \lambda_j)] \quad \text{for } j = 2, 3, \ldots, J-1 \quad (5c)$$

$$P_{nj} = 1 - 1/[1 + \exp(\beta X_n - \lambda_J)] \quad (5d)$$

Estimates of β and $\lambda_1 \ldots \lambda_J$ may be obtained using the maximum likelihood method based on a set of observations (households) making 0, 1, . . . , or J trips for which the attribute data in X_n are available. An application of the ordered response model in travel choice situation was the analysis of trip generation behavior of 774 elderly persons in the Washington, D.C., metropolitan area (*15*).

The ordered response model has the following advantages over the standard regression model of trip generation. First, the property that choice probabilities are necessarily between 0 and 1 means that in prediction mode, the model cannot forecast negative or infinite trips. The second advantage is that the model predicts the whole distribution of the response levels unlike the standard regression approach, which will at best predict the mean of the dependent variable. These advantages of the ordered response model are in addition to what was stated earlier: that the model offers a way to exploit the ordering of information.

STATISTICAL ANALYSIS

The discussion of statistical analysis covers three main areas. First, a brief discussion of the variables used in the estimation of the model is presented. This is followed by a comparative analysis of alternative utility specification functions. Finally, the estimated results are discussed including tests of the estimated parameters and a comparative analysis to assess the overall fit of the model and to demonstrate the extent of zonal variation in trip-making behavior indicated by the model.

Variables Used

The total number of home-based shopping trips over a 24-hr period made by all persons in the household is used in the definition of observed probabilities. ("Trip" as used in the paper is defined as a one-way movement between two places.) If a household is observed to make two trips, the observed probability of making two trips is defined as 1 and the probability of making any other number of trips is defined as 0.

The explanatory variables may be put into two groups: household characteristics and zonal dummy variables. The household characteristics include household size, number of vehicles owned by the household, number of children, and employment status of household members. The household size is expected to be positively correlated with the number of trips because it should influence the level of demand for goods or services, or both. The presence of children in the family may have a dual influence on travel. On the one hand, it may lead to some restrictions on the time available for shopping. Alternatively, it may be regarded as a scale factor leading to increased shopping trips. (The inclusion of household size controls for this scale effect to some extent so that one might expect the number of children to have a negative effect.) Vehicle ownership dramatically improves mobility; hence one might expect more trips in a household with more cars.

The four categories of employment status—full time, part time, working at home, and unemployed—may exert different time budget constraints on shopping trips. Full-time and, to some extent, part-time work is expected to have a negative impact on weekday home-based shopping trips. There is no expectation of the nature of effects of working at home on shopping. Two opposing effects of unemployment may be hypothesized. One effect is that the unemployed person has more time and therefore can make more shopping trips. The other hypothesis is that because a person is unemployed, he or she does not have enough money for shopping.

The four zonal dummy variables were introduced into the ordered response model in both additive and interactive manner. Implicit in the use of additive dummy variables is the assumption that zonal effects are independent of the effect of any household characteristic. It is possible that, for example, household size will have a different impact on trip generation in one zone as opposed to another. To test this hypothesis, the zonal dummies were interacted with household size in the model.

Specification and Comparison of Two Utility Functions

Two utility functions were specified, leading to two types of model. In Model 1, the effects of household size, number of children, and number of vehicles are specified as dummy variables. The utility function in Model 2 is a restricted form of Model 1 in which these same variables were entered in generic form. ("Generic form" means that the explanatory factors are treated as continuous variables. Because of the computational difficulties of including large numbers of dummy variables, the employment variables are entered in generic form for both models.) The two models were estimated in STATA Version 3.0, which uses a Newton-Raphson algorithm. There was some difficulty in estimating Model 1 because of the small number of observations for households of a size greater than six, with five or more children, with more than four vehicles or households making five or more trips. These households were dropped from the data set. The omitted observations constitute only 1.5 percent of the whole data set, leaving 10,701 observations for the estimation of the models.

Using a backward stepwise procedure, all the interactive terms were dropped from both models at a significance level of 0.15, which leads to the conclusion that the dummy variables for the zones have an additive, independent effect on trip generation. The variable working at home was also eliminated from the utility functions as a result of a problem of collinearity with full- and part-time employment. The remaining variables were used to estimate the two models for comparison.

A likelihood ratio test was performed to test the hypothesis that the two models are equal. The test statistic used is $-2(L_2 - L_1)$ which is distributed chi-square. L_1 and L_2 are, respectively, log likelihood values for Model Types 1 and 2. A chi-square value of 32.97 with 10 degrees of freedom was found, which is significant at 0.01, indicating that the two models are unequal. Models 1 and 2 have pseudo R^2 values of 0.0463 and 0.0435, accordingly. (Pseudo R^2 for each model is defined as $1 - L(\beta)/L(c)$, where $L(\beta)$ and $L(c)$ correspond to the log likelihood of a model with all parameters and with only constants, respectively). Model 1 was chosen for further analysis because it had a higher log likelihood value, as evidenced in both the pseudo R^2 and the likelihood ratio test.

Estimated Results

There were two runs of Model 1. The first run had all the household size, number of children, and vehicle dummy variables. (Household size variable has a minimum value of 1 and a maximum of 6 and the number of children and vehicles in each ranges from 0 to 4.) Pairwise significance tests were separately performed for the estimated coefficients of household size and number of children and vehicle dummies. The results showed that the coefficients of all the number of vehicle dummy variables are significantly different from each other. However, household-size dummy variables specific to 4 through 6 and the coefficients for children dummy variables specific to 2 through 4 are not significantly different. Dummy variables specific to household size 4 through 6 and number of children 2 through 4 were therefore constrained to be equal, and the model was run again.

The estimated parameters, together with their standard errors and z-values (used rather than t-values because of the large sample size) for the second run are presented in Table 3. The estimated model is highly significant: a likelihood ratio test of the model against the hypothesis that all the coefficients except the cut points are 0 gives a chi-square value of 556 with 16 degrees of freedom.

As one would expect, the dummy variables for household sizes and number of vehicles are significant. The magnitude of the coefficients of these dummies increases with increasing household size and number of vehicles but at a decreasing rate. The implication is that household sizes and number of vehicles have nonlinear effects on discretionary trip generation.

Two of the three categories of employment status are negatively weighted. Full- and part-time employment is significant, which may be symptomatic of time budget constraints on weekday, home-based shopping trips. The relatively high negative coefficient of full-time employment is indicative of the severe limitations that this variable has on home-based, weekday shopping trips. The effect of unemployment is not statistically significant at 0.1.

The estimated parameters for the two dummy variables for children are negative and are significant. In interpreting the negative coefficients for the children dummies, one should not lose sight of the fact that the data were collected on the weekdays between September and December when children of school age were at school. Child care responsibilities might have had some time budget effects on trip making.

Zonal dummies specific to Zones 3, 4, and 5 are positive and significant, implying that these locations have an effect on trip making relative to the Base Zone 1. The coefficient of the dummy variable for Zone 2 is negative and not statistically significant. (The corresponding value for Zone 1 is 0 by construction.) Pairwise significance tests based on a quadratic approximation to the likelihood function were performed to determine whether the coefficients of these zonal dummy variables are equal. The test results indicate that the differences between the dummy variables for zone pairs 3-4, 3-5, and 4-5 are not significantly different from 0. The test, however, rejects the equality constraint imposed on Zone Pairs 2-3, 2-4, and 2-5. The implication is that Zones 3 through 5 show trip-making propensities distinctly different from those of Zones 1 and 2. There is the possibility that the difference in shopping trip frequency among the zones may be partially because of unobserved income effects. The 1986 average household incomes for Zones 1 and 2 are, respectively, Canadian $32,000 and $39,000. On the other hand, each of Zones 3 through 5 has a comparatively higher average household income of approximately Canadian $45,000 (*16*). However, it is not possible to draw any conclusions about the effect of income on shopping trip frequency in the absence of adequate, reliable data.

Assessment of Prediction Ability

The following exercise is conducted to illustrate the ability of the model to predict aggregate trip-making propensities and also to illustrate the contribution of the zonal dummy variables to the pre-

TABLE 3 Ordered Response Model Estimates

Variable name	Coefficient	Standard error	z-values
Cut point specific to			
trips=1(λ_1)	2.429	0.119	20.412
trips=2(λ_2)	3.873	0.125	30.984
trips=3(λ_3)	5.690	0.160	35.563
trips=4(λ_4)	7.135	0.252	28.310
Household size (HHS) dummy variables specific to:			
HHS=2	0.578	0.108	5.351
HHS=3	0.921	0.163	5.661
HHS=4	1.174	0.236	4.975
Household members:			
fully employed	-0.567	0.700	-8.095
working part-time	-0.234	0.084	-2.796
unemployed	0.085	0.064	1.319*
Children (CHD) dummy variables specific to:			
CHD=1	-0.354	0.091	-3.879
CHD=2	-0.533	0.112	-4.749
Vehicles (VEH) dummy variables specific to:			
VEH=1	0.587	0.960	6.114
VEH=2	0.885	0.108	8.184
VEH=3	1.170	0.143	8.192
VEH=4	1.524	0.214	7.126
Zone (ZN) dummy variables specific to:			
ZN=2	-0.019	0.074	-0.259*
ZN=3	0.457	0.099	4.614
ZN=4	0.446	0.077	5.768
ZN=5	0.562	0.112	5.010

Summary statistics

Number of observations	10701
Chi-square	556.3
Degree of freedom	16
Prob > chi-square	0.0000
Log likelihood (c)	-6033.79
Log likelihood (β)	-5755.64
Pseudo R^2	0.0461

z-values = coefficient / standard error
All variables except those marked by asterisk (*) are significant at 0.01
Trips=0, HHS=1, CHD=0, VEH=0 and ZN=1 were normalised to zero

dictive ability of the model. Define A_{kj} as the aggregate probability that households in Zone k generate j trips, calculated as a relative frequency:

$$A_{kj} = \sum_{n \in Z_k} \frac{P_{nj}}{N_k}$$

where

P_{nj} = probability that household n makes j trips,
Z_k = set of all observations in Zone k, and
N_k = number of observations in Zone k.

A_{kj} is calculated for k = 1, 2, 3, 4, 5 and j = 0, 1, 2, 3, 4. This calculation is done first on the observed number of trips for households in the data and then on the fitted trip-making probabilities for the same households on the basis of the estimated model. For the purpose of comparison, these observed and predicted probabilities are presented in columns 2 and 3 of Table 4. The results suggest that the model should perform well for the purpose of estimating aggregate trip generation from zones.

To assess the contribution of the zonal dummy variables to the accuracy of prediction, the model was reestimated with the zonal dummies omitted from the specification. Aggregate probabilities calculated on the basis of this model are presented in the fourth column of Table 4. There is some zonal variation in these fitted probabilities, which occurs because of differences in household characteristics in various parts of the metropolitan area. However, these probabilities do not correspond to the observed probabilities nearly as well as those calculated from the original model. This indicates that, even after controlling for spatial variations in household characteristics, there are differences in trip-making behavior at

TABLE 4 Observed and Fitted Aggregate Probabilities

Trips	Observed	Model with zonal dummy variables	Model without zonal dummy variables
Zone 1			
0	0.8677	0.8672	0.8470
1	0.0938	0.0967	0.1100
2	0.0319	0.0300	0.0355
3	0.0054	0.0047	0.0056
4	0.0013	0.0014	0.0017
Zone 2			
0	0.8726	0.8733	0.8529
1	0.1001	0.0924	0.1060
2	0.0246	0.0285	0.0342
3	0.0022	0.0044	0.0054
4	0.0005	0.0014	0.0017
Zone 3			
0	0.7854	0.7838	0.8206
1	0.1427	0.1525	0.1281
2	0.0619	0.0526	0.0424
3	0.0077	0.0085	0.0068
4	0.0022	0.0026	0.0021
Zone 4			
0	0.7886	0.7866	0.8185
1	0.1441	0.1511	0.1297
2	0.0532	0.0515	0.0428
3	0.0106	0.0083	0.0068
4	0.0034	0.0026	0.0021
Zone 5			
0	0.7435	0.7452	0.8003
1	0.1842	0.1771	0.1418
2	0.0534	0.0639	0.0479
3	0.0120	0.0104	0.0077
4	0.0069	0.0033	0.0024

Columns may not add to one due to rounding error.

different locations. These differences may be because of differences in accessibility or other spatially variant factors.

CONCLUSIONS

The objective of this paper was to investigate the effects of location on discretionary household trip generation. Data on weekday, home-based shopping trips, socioeconomic characteristics, and location of households in five widely spaced zones in the GTA were obtained from the TTS. Weekday, home-based shopping trips constitute 59 percent or more of total shopping trips in each zone. An ordered response model was used to analyze the data. Household size, number of vehicles and children, employment status and location of households were included as explanatory variables in the analysis.

The results of the analysis, in terms of the likelihood ratio test of all the explanatory variables, suggest that the estimated model is significant. The z-scores indicate that full- and part-time employment and the dummy variables for household sizes, number of children, number of vehicles, and for Zones 3 through 5 produce significant effects on weekday, home-based shopping travel behavior. The significance of the positive coefficients of dummy variables for Zones 3 through 5 suggest that suburban living is positively correlated with weekday, home-based shopping trips. A comparison of observed and fitted values of aggregate probabilities of making 0,1,2,3, and 4 trips for households in all five zones indicates that the model has good predictive ability and that the inclusion of zonal dummy variables contributes significantly to that ability.

Two implications can be identified from this analysis. First, the ordered response model provides a viable methodology for trip generation. The trip-making decision should no longer be treated as a continuous variable or as a dichotomous response but as a multiple response with a natural order. The other implication is that trip-making behavior appears to be sensitive to location within the metropolitan area, even after controlling for spatial variations in observed household characteristics.

There is a need for further refinements in the application of the ordered response model to discretionary trip generation. The most important is probably the use of accessibility measures in place of spatial dummy variables. Accessibility indexes, which take account of travel costs in time, money, and human effort and of the spatial distribution of opportunities offer transportation planners a more direct way to measure the effect of location on trip-making behavior.

ACKNOWLEDGMENTS

This paper represents the results of the early stages of Kwaku Agyemang-Duah's Ph.D. thesis. Essential assistance in making the data available for this research came from Loy-Sai Cheah at the Joint Program in Transportation, University of Toronto, Canada. The financial support of the Natural Science and Engineering Research Council of Canada is gratefully acknowledged.

REFERENCES

1. Hutchinson, B. G. *Principles of Urban Transport System.* McGraw-Hill, Toronto, Ontario, Canada, 1974.
2. Gordon, P. A., A. Kumar, and H. W. Richardson. Beyond the Journey to Work. *Transportation Research A,* Vol. 22A, No. 6, 1988, pp. 419–426.
3. *The Regional Municipality of Ottawa-Carleton: Environment Review of the Regional Official Plan—Background Paper on Environment, Transportation and Land Use.* IBI Group, Feb. 1993.
4. Prevedouros, P. D., and J. L. Schoffer. Trip Characteristics and Travel Patterns of Suburban Residents. In *Transportation Research Record 1328,* TRB, National Research Council, Washington, D.C., 1991, pp. 49–57.
5. Greene, W. H. *Econometric Analysis,* 2nd ed. Maxwell Macmillan, Toronto, Ontario, Canada, 1993, Chapter 21.
6. Guy, C. M. Recent Advances in Spatial Interaction Modelling: An Application to the Forecasting of Shopping Travel. *Environment and Planning A,* Vol. 19, 1987, pp. 173–186.
7. Ortuzar, J. de D., and L. G. Willumsen. *Modelling Transport.* Wiley, London, 1990, Chapter 4.
8. *Trip Generation Rates,* 4th ed. ITE, Washington, D.C., 1987.
9. Friedman, B., S. P. Gordon, and J. B. Peers. The Effect of Neotraditional Neighborhood Design on Travel Characteristics. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1994.
10. McKelvey, R. D., and W. Zavoina. A Statistical Model for the Analysis of Ordinal Level Dependent Variables. *Journal of Mathematical Sociology,* Vol. 4, 1975, pp. 103–120.
11. Bhat, C. R., and F. S. Koppelman. An Endogenous Switching Simultaneous Equation System of Employment, Income, and Car Ownership. *Transportation Research A,* Vol. 27A, No. 6, 1993, pp. 447–459.
12. Train, K. *Qualitative Choice Analysis: Theory, Econometrics and Application to Automobile Demand.* MIT Press, Cambridge, Mass., 1986.
13. Ben-Akiva, M., and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Predict Travel Demand.* MIT Press, Cambridge, Mass., 1985.
14. Small, K. A. A Discrete Choice Model for Ordered Alternatives. *Econometrica,* Vol. 55, No. 2, 1987, pp. 409–424.
15. Sheffi, Y. Estimating Choice Probabilities Among Nested Alternatives. *Transportation Research B,* Vol. 13B, 1979, pp. 189–205.
16. *Census of Canada 1986: Population and Dwelling Characteristics Profile,* Series 94-112. Statistics Canada, 1986.

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

Alternative Methods To Iterate a Regional Travel Simulation Model: Computational Practicality and Accuracy

W. THOMAS WALKER AND HAIOU PENG

The results of full-scale testing of the available methods for implementing an iterative travel simulation process are presented. These methods include simple iteration of the simulation model chain, weighting iterative model outputs by the method of successive averages, and the Evans equilibrium algorithm. Simulated travel demands for each version of the model are compared with regional highway performance monitoring system data, 1990 highway traffic counts summarized by screenline, and public transit ridership data. These accuracy checks, in concert with estimates of the computational effort needed to execute each model variation, provide a useful insight into the costs and benefits associated with implementing an iterative travel simulation model. These comparisons also give guidance with respect to the relative efficacy of each iterative approach.

The Clean Air Act Amendments of 1990 and Intermodal Surface Transportation Efficiency Act of 1991 have significantly expanded the role of travel simulation models in evaluating the efficacy of proposed transportation improvements and in projecting the impact of these improvements on progress toward achieving mandated air quality standards. To adequately fulfill this expanded role in determining conformity, the federal legislation also requires that existing travel simulation models be validated with ground counts and upgraded to reflect an acceptable level of modeling practice. Perhaps the most significant of these requirements involves starting the simulation process with observed free-flow speeds and then iterating the entire simulation until a "reasonable agreement" is achieved between the travel speeds assumed for trip distribution and modal split and the resulting restrained speeds output by the highway assignment model.

The purpose of this paper is to conduct a full-scale exploration of the available methods for implementing this required iterative simulation process using the existing regional travel simulation model for the Delaware Valley Region as a test system. Socioeconomic data based on the 1990 Census, together with highway and public transit networks that reflect the facilities open to traffic in 1990, are used as inputs to Delaware Valley Regional Planning Commission's (DVRPC) existing simulation and selected iterative reconfigurations of this model. Simulated travel demands are then compared with regional highway performance monitoring system data, 1990 highway traffic counts summarized by screenline, and public transit ridership data. These accuracy checks, in concert with the relative computational effort needed to execute the model, provide a useful insight into the costs and benefits associated with adopting an iterative travel simulation model. These accuracy comparisons also provide guidance with respect to the efficacy of alterative iterative approaches. This analysis is limited to accuracy comparisons that are based on the existing noniterative calibrations of the travel demand models—the starting point for most state and regional attempts to implement an iterative simulation procedure. The results for all iterative approaches analyzed might be improved by recalibrating the models to better replicate actual travel data within an iterative formulation.

Delaware Valley Regional Planning Commission, Bourse Building, 111 South Independence Mall East, Philadelphia, Pa. 19106-2515.

METHODS FOR ITERATING TRAVEL SIMULATION MODEL

The methods that have been proposed to iterate the travel simulation model to some degree parallel the evolution of the highway assignment from simple iteration to weighted average to equilibrium, although discussions of iterative methods to date have focused on model convergence properties rather than model accuracy in a calibration sense. Levinson and Kumar (*1*) opened the current round of discussion by proposing that the modeling chain be simply iterated, starting with free-flow speeds, until a reasonable degree of convergence is obtained between the times used as input to the gravity and modal split models, and the congested times resulting from the subsequent highway assignment. The assignment results from the last iteration of this process form the basis for plan evaluation, conformity determination, and so on. Failure to iterate was found to overestimate congestion levels resulting from long-range socioeconomic and land use forecasts.

Although the convergence of travel times is monitored in the simple iteration method, it is not clear whether the simulated link volumes converge to a stable solution. For this reason Putman (*2, 3*), in work done in association with the Southern California Council of Governments, weighted together the highway link volumes from each simple iteration of the model chain using the method of successive averages (MSA). This successive weighting technique, proposed by Sheffi (*4*), uses a fixed weighting sequence, where the weight given to the link volume difference between the current iteration (n) and the weighted average resulting from the previous iterations is $1/(n + 1)$. The link volumes resulting from this method are easily shown to converge for any pattern of highway assignments. As the overwhelming proportion of the overall weight is given to the first few iterations, this method is often started with congested rather than free-flow speeds. This algorithm is usually run for a fixed number of iterations because the degree of convergence is directly determined by the progression of the weighting sequence.

The so-called Evans algorithm also uses a successive averaging technique to weight together the results of subsequent iterations of the modeling chain. However, instead of using $1/(n + 1)$ as the weighting sequence, weights are calculated by a gradient search within a Frank-Wolf decomposition. This weighting takes the following form:

$$P'_{ijm} = (1-\lambda)P_{ijm} + \lambda Q_{ijm}$$

and

$$v'_a = (1-\lambda)v_a + \lambda w_a$$

where P_{ijm} and v_a represent the results of successive weighted averages over the previous iteration estimates of trips between zones i and j by mode m and travel over link a, respectively. Q_{ijm} and w_a represent corresponding results for the current iteration run of the simulation models.

This method is related to the well-known equilibrium assignment method; however, the Evans algorithm incorporates the results of the entire simulation process from trip distribution to transit and highway assignment within the gradient search. The method is known to converge rapidly. It requires that only one iteration of highway assignment be conducted between successive runs of the simulation model chain. These two factors result in large potential savings in computation vis-à-vis simple iteration of the traditional modeling chain. This algorithm is based on work done by Evans as part of her Ph.D. dissertation in the early 1970s (*5*). Recently, Boyce et al. suggested using this method to satisfy the federal iterative modeling requirements (*6,7*).

Convergence criteria can be rigorously defined for this algorithm using the difference at a given iteration between the numerical value associated with the primal and dual of the underlying nonlinear impedance minimization problem (primal and dual are equal at convergence). Because these criteria are difficult to calculate, a convergence criterion similar to the one applied in most implementations of the equilibrium assignment is used. After weighting the current iteration trip interchanges and link volumes together with the composite results from previous iterations, the new capacity restrained link times together with transit travel times, fares, and parking charges are used to project system total impedance (S1). This value is then compared with the total impedance resulting from the next iteration of the simulation models (S2). The difference (error) between these two estimates expressed as a fraction of current impedance (S2) is taken as a measure of convergence. This assumes that the impact of reiterating the travel simulation becomes progressively smaller as convergence nears. This definition of convergence has proved to be adequate in practice.

ITERATIVE FORMULATION

The existing DVRPC travel simulation model is a classic implementation of the four-step process. All aspects of the model produce estimates of daily travel. Trip generation is based on constant trip rates imbedded in a cross-classification structure. The trip distribution, modal split, and highway assignment models are based on average daily highway travel speeds. Bus speeds are taken from the existing a.m. peak transit operating schedule and held constant throughout the simulation. Trip distribution uses a doubly constrained gravity model, stratified into three-person (home-based work, home-based nonwork, and non-home–based) and four-vehicle trip purposes. The person-trip gravity models utilize a combined highway/transit network interzonal impedance measure based on a relative highway/transit service level bias adjustment (*8*). Modal split utilizes a binary probit-like formulation stratified by trip purpose, transit submode, and automobile ownership. The highway assignment is based on the equilibrium method using minimum travel time paths. Initial highway speeds are input through a table lookup stratified by functional class and density of development (area type). The transit assignment is unrestrained. It uses minimum paths that are based on the modal split model definition of impedance.

The DVRPC model is among the largest in existence, covering a densely developed area of about 10 400 km^2 (4,000 mi^2) that is subdivided into 1,449 traffic zones. The highway network contains about 34,000 one-way links, and the transit network contains about 360 routes, including commuter rail, rapid transit, light rail, and bus facilities. Overall, the model has been stable over time, achieving validation with counts for 1960, 1970, 1980, and 1990 with minimal structural or parameter changes. A more detailed description of this model is given in Walker (*9*). The DVRPC model was originally developed on a mainframe using the PLANPAC/UTPS packages but recently has been converted to a microcomputer environment using TRANPLAN. All sensitivity tests reported in this paper were done using the TRANPLAN microcomputer system.

Incorporating Actual Highway Speeds into Simulation Model

The DVRPC model has a fundamental problem that prevents it from being used directly in an iterative framework. Input highway speeds are unrealistically low, particularly on freeways. Furthermore, the output speeds from the assignment (via the BPR restraining curve) are even more unrealistic, perhaps half the true average daily highway operating speeds. This is common in simulation models developed during the 1970s. Although these speeds cannot be used for emissions calculations, they generally improve the accuracy of the highway assignment, which responds favorably to a bias against freeways and severe capacity restraint. A postprocessor methodology is used to reestimate highway operating speeds on the basis of assigned volumes before it estimates emissions.

The most straightforward way to correct this problem is to insert "actual" congested speeds into the highway network through a revised speed lookup table. However, this substitution increased the simulation error to an unacceptable level. Clearly, a more sophisticated method is needed to incorporate actual operating speeds into the travel simulation model. It was always obvious that some of the values in the original highway speed lookup table were not real speeds but rather a crude form of impedance. The phenomenon being addressed was that drivers consider distance (or operating cost) as well as travel time when choosing routes. Freeways move faster than arterials, but there is a limit to the route circuity that drivers will accept to achieve a savings in travel time.

The modal split model already had a highway impedance measure that considered both highway time and operating cost. A theoretically appealing way to incorporate actual congested speeds is to extend this impedance measure to the gravity model and highway assignment as well. The entire simulation model would then be based on a uniform definition of impedance. This impedance definition is similar to the one found in most disutility-based modal split models:

$$Z = k_1\,ET + k_2\,RT + k_3\,C$$

where

Z = impedance for given travel mode;
ET = excess or out-of-vehicle time (i.e., terminal time for highway, sum of walk and wait time for transit; transit impedance also includes a supplemental transfer penalty);
RT = running or in-vehicle time;
C = monetary cost (i.e., fare for transit; out-of-pocket operating cost plus tolls and parking for highway); and
k_1, k_2, k_3 = calibration constants.

To test this approach, highway trees were built using the modal split impedance definition with actual congested times in the lookup table. The resulting impedance skims were found to be perfectly collinear with the minimum time skims from the original speed lookup table. Only a simple-scale factor was required to make these impedance skims usable with the original gravity model friction factor curves and terminal and intrazonal times, and so on. Highway assignment path building also was based on this impedance definition. However, the capacity restraint calculation was limited to the travel time portion of the impedance. To improve the highway travel speeds produced by the model, the exponent of the Bureau of Public Roads (BPR) restraint curve was reduced from 4.0 to 3.0.

The accuracy of the resulting assignment was checked on the basis of 1990 traffic counts summarized through a series of 14 screenlines. These screenlines form the basis for FHWA model validation within the DVRPC region. Included are circumferential central business district and intermediate suburban cordon lines, all crossings of the Schuylkill and Delaware rivers, and a series of radial cutlines.

The use of a highway impedance model increased the total volume error for all screenline counts from 2.2 percent for the original model to 3.8 percent. The number of screenlines with volume errors greater than 10 percent increased from two, with the worst screenline having 12 percent error, to four, with the worst being 13 percent. The R^2 between predicted and actual traffic volumes for all screenline crossings was reduced from 0.89 to 0.85 by the highway impedance model. The simulated highway speeds produced by the impedance-based model, although almost 10 percent low on average, were judged to be sufficiently accurate to test iterative simulation methods.

Implementing Simple Iteration and MSA Approaches

The simple and MSA methods of iterating the travel simulation model are straightforward to incorporate into an existing travel simulation model. The simple iterative method requires only that a feedback loop be inserted into the model that inputs the highway link speeds output from highway assignment of the current iteration into the network before rebuilding and reskimming the minimum impedance paths, so that trip distribution and modal split model runs of the next iteration step will be based on the current iteration's congested link travel times. Most, if not all, travel simulation model software packages incorporate link travel times. Most, if not all, travel simulation model software packages incorporate provisions for this feedback loop. The estimates of link volumes produced by the simple iterative approach are taken directly from the highway and transit assignments of the last iteration that is executed.

The MSA approach builds on these simple simulation model iterations by combining the link volumes of each simple iteration into a composite volume, using the weighting scheme outlined earlier. This composite link volume is calculated by a postprocessor computer program that processes the output of each successive model iteration. The MSA software used in this analysis was prepared by the Urban Analysis Group as part of the TRANPLAN package.

Implementing Evans Algorithm

The Evans algorithm is not difficult to implement in a four-step travel simulation model that includes a highway assignment model based on the equilibrium method, although some extension of the modal split and highway assignment models, as well as the associated computer code is required (Figure 1). Evans reexecutes the gravity and modal split models after each iteration of highway assignment. Therefore, a restart procedure must be available in the highway assignment program to access the weighted average highway link volumes from the previous iteration, load the network for the current iteration, calculate the weight for the current iteration (λ), and prepare a convex combination of the link volumes for the current iteration and previous weighted average. This is not a fundamental departure from the way things are normally done in the equilibrium assignment, and the restart option already exists in TRANPLAN and perhaps other packages.

The second required extension is to include the impedance implications of the highway and transit trip tables into the gradient calculation that is used to determine λ. This requires an estimate of transit impedance and off-network highway impedance (terminal

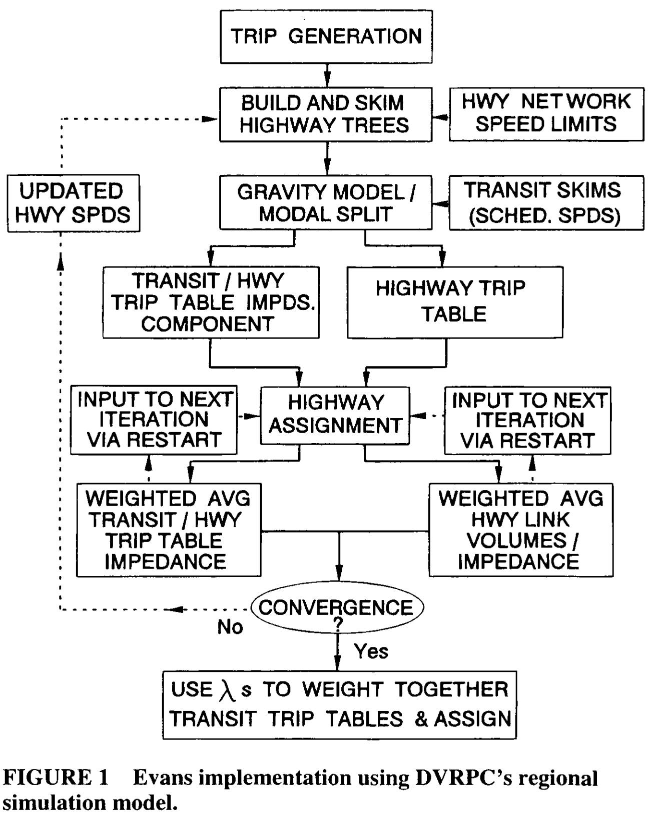


FIGURE 1 Evans implementation using DVRPC's regional simulation model.

times and parking charges) for the trip tables of the current iteration and the weighted average of the previous iterations. Transit impedance is assumed to be independent of the highway link restraining process and is calculated as the sum of the products of interzonal transit impedances and transit volumes. It may be theoretically desirable to also include the effect of highway congestion on bus and trolley travel times. However, this enhancement requires massive changes to the highway assignment computer program and is beyond the scope of this study. In any case, only about 4 percent of the region's total travel is made by transit.

In this implementation, it is assumed that weighted average totals of transit and off-network highway impedance are linear in λ and can be calculated directly from the system totals for the current and weighted average of the previous iterations. The alternative would be to calculate a new λ-weighted trip table and multiply this new table by the interzonal impedance matrix. This simplification has little effect on the accuracy of the calculation. It greatly reduces the computational effort in the search routine that is used to determine λ and the complexity of the required program code changes. For the current iteration, the system total for both the off-network highway and transit impedance are calculated in the modal split model and passed in a scratch file to the highway assignment for inclusion in the gradient calculation. Similarly, the weighted transit and off-network highway impedance calculated in the highway assignment is passed from iteration to iteration in a scratch file.

In the Evans algorithm, trip tables are weighted together from iteration to iteration using λ-based successive averages in exactly the same way as highway link volumes. Thus, the transit trip table must be calculated with this method before assignment to the transit network.

COMPARISON OF RESULTS

This section compares the results of the impedance version of the DVRPC simulation model under three alternative methods for iterating the model: simple iteration, MSA, and the Evans algorithm. All iterative simulation model runs were started with highway speed limits, which are assumed to represent the "free-flow speeds" recommended in the federal guidance. Congested speeds are unacceptable as a starting point in iterative processes using the BPR restraining curve shown later because the restrained link times can never be lower than the input times T_0.

$$T = T_0\left[1.0 + 0.15\left(\frac{V}{C}\right)^f\right]$$

where

T = adjusted link time,
T_0 = initial input link time,
V/C = ratio of volume to capacity in current assignment, and
f = exponent on V/C; 3.0 in these runs, 4.0 default.

The fact that T_0 is not increased leads to errors in mobile source emissions because speed increases above current congested speeds, perhaps resulting from highway capacity improvements or land use changes, cannot be modeled. To standardize the comparisons, all three methods were iterated five times after an initial iteration (0) execution of the travel simulation models.

Convergence and Computational Efficiency

Table 1 compares the systemwide convergence criteria for the simple and Evans method. Both the simple and Evans algorithms converged to the neighborhood of 0.01 error after five iterations. The error statistic in this table refers to highway link impedance only in the simple model but also reflects the trip table impedance components in the Evans results. For this reason the Evans model estimates of S1 and S2 are somewhat larger. The difference between S2 in the simple and Evans cases gives an indication of the relative impact of the trip table impedance component within the Evans gradient calculation. Overall, the highway links provide about 90 percent of the influence in the determination of λ. The effect of the trip table gradient component is usually to reduce the weight given to the first two or three iterations.

Although not quite reaching the 0.01 criteria, the Evans convergence rate was particularly impressive because it is based on only six executions of the highway assignment. The DVRPC network is slow to converge in equilibrium assignment, requiring 12 to 15 iterations to reach this level of error. For this reason, the simple iteration results required a total of 90 executions of the highway assign-

TABLE 1 Convergence Statistics After Five Iterations from Speed Limits

ITERATIVE METHOD	PROJECTED TOTAL IMPEDANCE $\times 10^4$ (S1)	ACTUAL TOTAL IMPEDANCE $\times 10^4$ (S2)	ERROR (S1-S2)/S1	HIGHWAY ASSIGNMENT ITERATIONS	APPROXIMATE COMPUTATION TIME [a]
SIMPLE ITERATION	99,508	98,498	0.010	90	78 HRS.
MSA	NA	NA	NA	90	79 HRS.
EVANS ALGORITHM	111,120	109,636	0.014	6	15 HRS.
EVANS ALGORITHM FULL RESTRAINT ITERATION 0	109,486	109,484	0.000	20	26 HRS.

[a] 66 MHZ 486 UNDER OS/2.

ment. The total computation time for the simple method is about 78 hr on a 66-MHZ 486 microcomputer under OS/2. This is an impractical running time for most planning applications. For example, current federal guidance for the DVRPC region requires a total of seven simulations (1990 base year plus build and no-build alternatives for 1996, 2005, and 2015) to demonstrate conformity. Including the time needed for program setups and output checking, the simple method would require somewhere between 25 and 35 days to complete the computation. Five iterations of the Evans approach will run overnight (15 hr per alternative), or about 7 to 10 days to complete the required conformity simulations. Despite the ongoing advances in microprocessing speed, this is an overwhelming computational advantage.

The MSA method weights together the results of the five simple method iterations with a special postprocessor program. This MSA weighting operation requires something less than 1 additional hr to complete (79 hr total for five iterations). The MSA approach does not lend itself to the calculation of S1 and S2 parameters. Furthermore, these parameters reflect only the system total of impedance and do not directly measure the variation in link volumes from iteration to iteration. To directly measure link level convergence, the percent root mean square (RMS) difference in link volumes, from iteration to iteration, was also calculated for each of the three iterative methods. The results of the calculation are shown graphically in Figure 2. As one might expect, the MSA approach had the fastest rate of link volume convergence. It significantly improved the convergence rate of the simple method, which tended to level out at about 5 percent RMS difference per iteration. With the MSA method it would seem that it is possible to terminate computation after Iteration 2, a savings of 50 percent (40 hr of computation per simulation). The Evans approach demonstrated a high rate of convergence, closing all the way from 78 percent RMS difference between Iterations 0 and 1 to 18 percent between Iterations 4 and 5. However, it is clear from Table 1 and Figure 2 that additional iterations of the simulation model are required for the Evans algorithm to reach the level of convergence of the simple approach. This lack of link-level convergence is also reflected in the error statistics.

To achieve complete convergence, the Evans algorithm was restarted and run for five additional iterations. Convergence to the 0.01 criteria was achieved on Iteration 7 (20 hr of computation time), although link-level convergence to 5 percent RMS difference was not achieved. On Iteration 7, this measure dropped to just below 10 percent, and stayed around 10 percent throughout Iterations 8 to 10. It seems that the 5 percent RMS difference level of link convergence requires multiple iterations of the highway assignment within each Evans iteration to smooth out the highway assignment though traditional capacity restraint. DVRPC's highway network is dense in terms of link topology, and the Evans results might also be improved by the creation of additional traffic zones.

When starting from speed limits, a particularly critical point in terms of the smoothness of the traffic assignment was Iteration 0. For this reason, it seemed probable that the convergence properties of the Evans algorithm could be improved by executing a full traditional capacity restraint (15 iterations) in Evans Iteration 0, thence continuing with the standard single iteration of restraint within each Evans iteration. The results of the test are also reported in Table 1. In terms of total impedance, this variation on the Evans model significantly improved the rate of convergence. After five iterations of Evans, the error term was reduced to less than 0.001, although the link-level convergence did not go below 10 percent RMS difference. As the 0.01 level of convergence was achieved in Iteration 3 the last two iterations could be eliminated saving about 5 hr of computation time over the 26-hr required. An alternative to a full restraint may be to use congested speeds as the BPR curve T_0 value in Iteration 0 and then switch to speed limits in subsequent Evans iterations.

Accuracy and Usability for Emissions Calculations

The effect of iterating the travel simulation models on assignment accuracy is indicated in Table 2. Although the total of predicted and counted volumes for all screenline links remains well below a 5 percent difference, individual screenline accuracy is degraded versus the noniterated travel simulation under all three iterative approaches. The R^2 between predicted and counted volumes for all screenline links is also significantly reduced by the iterative simulations. The biggest factor in this error increase is the use of speed limits rather than congested speeds as the starting point for the assignment. The

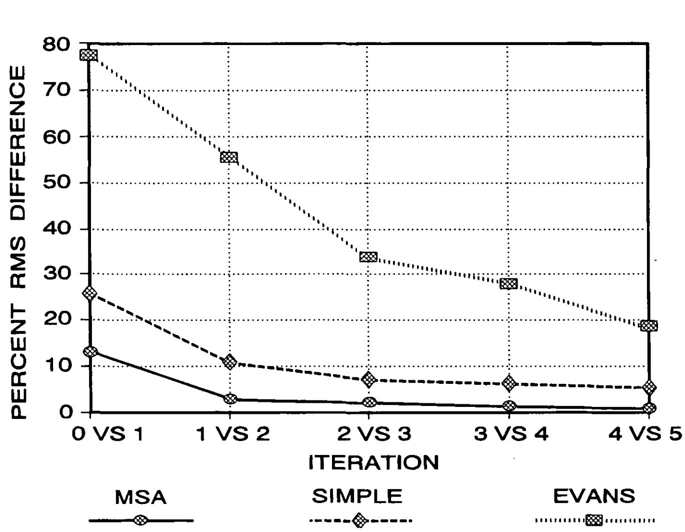


FIGURE 2 Comparative rate of convergence of highway link volumes.

TABLE 2 Highway Screenline Error Statistics for Simple Iteration, MSA, and Evans Algorithm from Speed Limits

ITERATIVE METHOD	OVERALL ERROR	# OF SCREEN LINES > 10% ERROR (WORST)	AVG. ABS. SCREEN LINE ERROR	R^2 ALL LINKS
SIMPLE ITERATION	2.2%	4 (19%)	7.17%	0.75
MSA	3.6%	5 (18%)	7.54%	0.75
EVANS ALGORITHM FIVE ITERATIONS	-3.5%	6 (23%)	8.57%	0.66
EVANS ALGORITHM COMPLETE CONVERGENCE	-3.6%	4 (24%)	7.61%	0.67
EVANS ALGORITHM FULL RESTRAINT ITERATION 0	-2.9%	3 (22%)	7.38%	0.67

results for the MSA approach are slightly worse than for the simple iteration method but comparable overall.

The Evans method showed a somewhat larger reduction in accuracy. In part, this resulted from executing the highway assignment only six times. Restarting Evans for two additional iterations resulted in some improvement in accuracy of the screenline volumes, but the full-capacity restraint in Evans Iteration 0 almost achieved screenline validation in terms of volume totals. Only 1 of the 14 screenlines and cutlines checked had a total traffic volume error greater than 11 percent, with the worst (22 percent) being a small suburban circumferential cutline. However, this variation of the Evans algorithm continued to have a significantly smaller link-level R^2 than either simple iteration or MSA. The trip table and restrained link volumes rapidly converge to a hand-in-glove fit in the Evans approach. This tends to magnify the effect of network topological and model calibration/specification deficiencies. All three modeling approaches will require some degree of simulation model enhancement and recalibration to achieve screenline validation. This recalibration is beyond the scope of this investigation.

As indicated in Table 3, all iterative approaches produced acceptable estimates of regional highway vehicle kilometers of travel (VKMT) and transit ridership; however, all significantly overestimated highway operating speed (by 12.4 to 17.6 percent). None of these methods can be used to estimate mobile source emissions without first reestimating congested speeds with a postprocessor.

Alternative Capacity Restraining Functions

In an attempt to improve the accuracy of the speed estimates produced by the iterative simulations, four variations of the capacity restraining function were tried. Because the computation associated

TABLE 3 Selected Regional Travel Statistics for Simple Iteration, MSA, and Evans Algorithm from Speed Limits

ITERATIVE METHOD	HWY. VKMT [a] $\times 10^6$ (% DIFF. FROM HPMS)	HIGHWAY AVG. SPEED KM/H (% ERROR)	TRANSIT BOARDING $\times 10^6$ (% ERROR)
SIMPLE ITERATION	143.2 (-2.5%)	50.5 (17.6%)	1.26 (7.7%)
MSA	145.1 (-1.2%)	48.3 (12.4%)	1.26 (7.7%)
EVANS ALGORITHM FIVE ITERATIONS	142.2 (-3.2%)	48.8 (13.5%)	1.24 (5.9%)
EVANS ALGORITHM COMPLETE CONVERGENCE	141.4 (-3.7%)	48.9(13.9%)	1.24 (5.9%)
EVANS ALGORITHM FULL RESTRAINT ITERATION 0	140.8 (-4.2%)	49.1 (14.2%)	1.26 (7.7%)

[a] VKMT = VEHICLE KILOMETERS OF TRAVEL;
VEHICLE MILES OF TRAVEL = VKMT ÷ 1.6093

with the simple iteration method is excessive, these tests were limited to the Evans algorithm. All three iterative methods produce similar estimates of regional VKMT and operating speed in the earlier comparisons.

The first and second alternative restraining function involved resetting the exponent of the BPR curve *V/C* ratio to 4.0 and 7.0, respectively, and then running the iterative simulation from speed limits. The standard value of the V/C exponent is 4.0, but recent research has suggested that larger values, perhaps 7.0, may produce better results. The third and fourth variations involved direct use of the speed curves from DVRPC's emissions postprocessor methodology as the restraining function. These speed curves are much more complex than the BPR function, being related to the methods contained in the Highway Capacity Manual. The exact formulation of these curves may be found in Walker (*9*). Because the times output by these curves are not limited by the input T_0 values, the postprocessor speed curves were used in two ways: one using speed limits as the starting point of the simulation process and the other using congested speeds.

The results produced by these tests are presented in Tables 4 and 5. Resetting the BPR exponent to 4.0 significantly improved the screenline accuracy of the Evans algorithm, although the results were still not as good as the simple or MSA results shown earlier. The exponent value of 7.0 improved the screenline results even further, being comparable with those of simple iteration and MSA shown earlier. The regional VKMT and transit ridership estimates for both exponent values were comparable with those produced by the 3.0 case, but average speed estimates produced by the 7.0 exponent value had less than 1 percent error, raising the possibility of eliminating the speed estimation postprocessor. This version of the Evans model seems to be able to produce reasonably accurate estimates of both VKMT and speed. However, the 7.0 BPR exponent slowed down the rate of algorithm convergence. Ten Evans iterations were required to achieve 0.01 convergence.

Use of the postprocessor speed curves generally degraded the accuracy of travel volumes produced by the Evans algorithm. This occurred in part because the modal split model went out of calibration, leading to severe overestimation of center-city transit ridership and corresponding underestimation of some highway screenline totals and of regional VKMT. This restraining function did produce significantly more accurate estimates of simulated highway speeds, however (about 3.3 percent overestimated).

The postprocessor curves produced about the same error statistics, whether the simulation was iterated from speed limits or congested speeds. The highway link volumes produced by these alternative starting points had about a 13.5 percent RMS difference after five iterations. This version of the Evans algorithm seemed to produce relatively unique results at both the regional and link level, regardless of the initial speeds, although as one might expect, convergence was significantly faster when the algorithm was started from congested speeds.

CONCLUSIONS

It is clear from the results presented in this paper that converting the DVRPC travel simulation model to an iterative formulation on the basis of initial free-flow speeds is not a trivial undertaking. Simple iteration of the modeling chain requires days of computation to complete the simulation for a single alternative. The draft federal guidance also requires disaggregating the simulation process into separate peak and off-peak models. Implementing this requirement would effectively double all computing times reported in this paper. Furthermore, the off-peak time period is far from homogeneous in terms of congestion. Midday congestion resembles the peak period in many suburban areas, whereas evening travel in these areas in virtually free flow. Three or four time periods may be required. For this reason the computational efficiencies resulting from the MSA and Evans algorithms are essential to the continued computational practicality of the travel simulation process.

The Evans algorithm required the least amount of computer time to achieve convergence in terms of systemwide total impedance, reducing the time required by 80 percent versus simple iteration. This time savings is dependent on the number of iterations of restraint that are required for the highway assignment in the simple method. DVRPC's network requires 15 iterations in a normal assignment. Other regions whose network converges faster may receive a smaller time savings from the Evans algorithm.

The MSA procedure allows the number of iterations (and associated computation) required to achieve link-level convergence to be

TABLE 4 Highway Screenline Error Statistics for Alternative Restraining Functions

ITERATIVE METHOD	OVERALL ERROR	# OF SCREEN LINES > 10% ERROR (WORST)	AVG. ABS. SCREEN LINE ERROR	R^2 ALL LINKS
BPR EXP. 4.0 FROM SPEED LIMITS	-3.4%	3 (21%)	8.16%	0.70
BPR EXP. 7.0 FROM SPEED LIMITS	-6.6%	4 (19%)	7.20%	0.74
POST-PROCESSOR CURVES FROM SPEED LIMITS	-7.1%	4 (26%)	10.53%	0.74
POST-PROCESSOR CURVES FROM CONGESTED SPEEDS	-7.3%	4 (21%)	10.89%	0.73

TABLE 5 Selected Regional Travel Statistics for Alternative Restraining Functions

ITERATIVE METHOD	HWY. VKMT [a] $\times 10^6$ (% DIFF. FROM HPMS)	HIGHWAY AVG. SPEED KM/H (% ERROR)	TRANSIT BOARDING $\times 10^6$ (% ERROR)
BPR EXP. 4.0 FROM SPEED LIMITS	141.4 (-3.7%)	47.1 (9.7%)	1.25 (6.8%)
BPR EXP. 7.0 FROM SPEED LIMITS	140.8 (-4.2%)	42.6 (-0.7%)	1.24 (6.0%)
POST-PROCESSOR CURVES FROM SPEED LIMITS	134.4 (-8.5%)	44.4 (3.3%)	1.34 (14.5%)
POST-PROCESSOR CURVES FROM CONGESTED SPEEDS	134.2 (-8.7%)	44.4 (3.3%)	1.34 (14.5%)

[a] VKMT = VEHICLE KILOMETERS OF TRAVEL; VEHICLE MILES OF TRAVEL = VKMT ÷ 1.6093

reduced by one-half. Although converging very rapidly, the Evans algorithm did not achieve the degree of link-level convergence of the simple iteration or MSA approach in the test applications. Running Evans for two additional iterations improved the link and system level convergence (and accuracy) but reduced the computer time savings versus MSA somewhat. However, the Evans algorithm has considerable theoretical appeal, in that the weights on successive simulation model iterations are based on a Frank-Wolf decomposition rather than the arbitrary sequence used by MSA.

All three iterative approaches significantly degraded the accuracy of the travel simulation model, making validation of screenline volumes and congested speed much more difficult to achieve. The use of speed limits rather than congested speeds as a starting point for the iterative process was a major factor in this accuracy loss. The Evans approach was somewhat less accurate in part because of the drastic reduction in the number of iterations of the highway assignment required for five iterations. However, the rapid convergence between trip table and congested link volumes in this approach may also magnify the effect of certain deficiencies in the travel simulation model. Simulation model enhancement or recalibration may be necessary to optimize the accuracy of the results from any of the three iterative approaches.

Almost all iterative formulations tested tended to significantly overestimate congested highway link speeds and will require post-processor-based reestimation of speeds before mobile source emissions calculation. Only the Evans algorithm with a BPR restraint curve exponent of 7.0 seems to produce estimates of both highway VKMT and congested operating speed when starting the iterative process from highway speed limits.

The motivation for implementing an iterative simulation is to be able to accurately assess the impact of future land use patterns and proposed transportation facilities. It is interesting to note that the highway travel speed lookup table and other model parameters in the existing DVRPC model have remained almost unchanged for the last 30 years, despite repeated intervening forecasts of increased highway congestion. Furthermore, budget constraints and intense citizen opposition have limited the region's ability to build new freeways and improve existing roadways. Potential excessive congestion resulting from population and employment growth and increasing dependance on automobiles has been controlled by highway peak spreading and decentralization of urban activity into suburban and rural areas of the region. From this perspective, it would seem more likely that a significant projected imbalance between input and output speeds in the simulation model would be caused by an underestimate of decentralization and peak spreading than a failure to iterate. Iterative travel simulation models should include a feedback loop that incorporates the impact of localized projected congestion levels on the underlying land use and socioeconomic forecast. This feedback could utilize formal land use models, if sufficiently sensitive to localized congestion conditions, or might be accomplished through ad hoc adjustments.

REFERENCES

1. Levinson, D., and A. Kumar. Integrating Feedback into the Transportation Planning Model: Structure and Application. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., 1993, pp. 70–77.
2. Putman, S. H. *Integrated Urban Models 2: New Research and Applications of Optimization and Dynamics.* Pion Limited, London, 1991.
3. *MSACOM Computer Program Documentation.* Southern California Association of Governments, 1994.
4. Sheffi, Y. *Urban Transportation Networks.* Prentice-Hall, Englewood Cliffs, N.J., 1985.
5. Evans, S. P. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research,* Vol. 10, 1976, pp. 37–57.
6. Boyce, D. E., M. Tatineni, and Y. Zhang. *Scenario Analysis for the Chicago Region with a Sketch Planning Model of Origin-Destination Model and Route Choice.* Urban Transportation Center, University of Illinois at Chicago, 1992.
7. Boyce, D. E., M. R. Lupa, and Y. Zhang. Introducing Feedback into the Four-Step Travel Forecasting Procedure versus the Equilibrium Solution of a Combined Model. In *Transportation Research Record 1443,* TRB, National Research Council, Washington, D.C., 1994.
8. Walker, W. T. Testing and Adjusting Regional Travel Simulation Models with 1980 Census Data. *Transportation Quarterly,* Jan. 1988, pp. 63–88.
9. Walker, W. T. The Impact of Preaggregation of Highway Network Travel Data on the Accuracy of Mobile4-Based Emissions. In *Transportation Research Record 1366,* TRB, National Research Council, Washington, D.C., 1992, pp. 51–59.

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

Nationwide Recreation Travel Survey in Japan: Outline and Modeling Applicability

TETSUO YAI, HARUTOSHI YAMADA, AND NAOHISA OKAMOTO

The Nationwide Recreation Travel Survey (NRTS) was conducted by the Ministry of Construction in Japan in 1992. It covered the nine regions of Japan and collected more than 30,000 samples through home-based surveys and nearly 13,000 samples from recreation site surveys. Before the survey, recreational activities had been investigated by smaller-scale surveys (one-tenth the size of NRTS) every 2 years. The survey is expected to provide fundamental and useful information for suburban highway planning. Whereas road investment plans conventionally have been based on future weekday traffic volumes, several roads in suburban areas have become heavily congested on weekends. The prediction of weekend travel will have a more important role in road planning. The characteristics of recreational travel by car should be examined to gain valuable insight into highway planning in recreational areas. After survey profiles were summarized recreational travel demand models for trip generation and trip distribution were developed using an aggregated regression model and a disaggregate model. Finally, fruitful data sources and sufficient modeling applicabilities are provided.

The first large-scale survey for recreational travel in Japan is introduced in this paper. The Nationwide Recreation Travel Survey (NRTS) was conducted by the Ministry of Construction (MOC) in 1992 to understand the characteristics of recreation travel and the applicability of the survey data to demand modeling.

Every 5 years, MOC collects weekday vehicle trip data by roadside and car owner surveys entitled Road Traffic Census (RTC). Nearly 3 percent of car owners in Japan have been sampled in the RTC, and all vehicles that pass through interregional arterial roads have been intercepted by roadside surveyors. The RTC in 1990 also included a survey for weekend car trips using the same questionnaire sheet with the weekday survey. In both surveys, the trips that respondents were required to fill out were restricted to those within a specific 24-hr period.

Most weekday activities in a city, such as commuting or shopping, are completed within a day. However, recreational travel often exceeds 24 hr and characteristics of recreation travel, such as destinations, activities, and durations change by season. Additionally, a route choice behavior between two recreational facilities, which is not included in RTC questionnaires, is determined probably not only by minimum travel time but also by attraction of the route itself (e.g., road quality or scenic beauty) (*1*).

Although road investment plans in Japan have been based conventionally on future weekday traffic volumes, several roads in suburban areas are more heavily congested on weekends than on weekdays. This is because the low density of road networks radiating to recreational areas outside of cities provides fewer alternative routes for travelers. Additionally, substantial volumes of traffic by passenger vehicles often are higher than road capacities on weekends. The potential demand of recreational car travel is expected to be high in metropolitan regions because residents evidently have experiences of giving up a weekend drive because of heavily congested roads to or in suburban recreation areas.

Actually, the volumes of passenger vehicles for recreation traffic on weekends are equivalent to those of weekday commuter traffic. For example, according to RTC data, car trip generations in Kanto region amount to 3.1 million commuter vehicles on a weekday and 2.6 million recreational trip vehicles on a weekend. However, the average distance of recreational trips is twice as long as that of commuter trips, and total vehicle kilometers traveled per day of weekend recreational trips is much higher than that of commuter trips.

Consequently, after one understands recreational travel behavior, the prediction of weekend traffic, which is principally composed of recreational trips, and revision of highway planning in suburban areas are expected to resolve the traffic and environmental problems in those areas. A new survey was designed to collect individual histories of recreational activities during a year because characteristics of recreational trips change seasonally, and recreational trips often are not completed within a day. After several properties of recreational travel from home-based surveys are briefly summarized, trip generation models and destination choice models also are examined in this paper (*2*).

SURVEY SYSTEM AND QUESTIONNAIRES

Survey Method for Recreational Travel

Fundamental characteristics of recreational travel, which are concerned with survey system selections, are explained as follows: (*a*) recreation travel, in particular overnight travel, is generated infrequently for each household; (*b*) individual recreational activities differ with the seasons; and (*c*) a trip route depends on the attractiveness of the route, such as the landscaping as seen from the road. These cause the inefficiency of origin-destination trip surveys conducted on a specified day of a specific season. Considering these properties, home-based and choice-based surveys were conducted. To examine characteristics of recreation travel and to demonstrate the modeling applicability of trip generation and trip distribution, data of personal records of travel, which depend on memory, were collected using home-based surveys. Telephone surveys such as the National Personal Transportation Survey in the United States provided 1-year period data from 24-hr individual samples; however, the applicability of large-scale telephone surveys is still uncertain in Japan.

Choice-based sampling in the specific recreation sites can target a specified group in the whole population but cannot collect random

T. Yai, Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo #152, Japan. H. Yamada, Public Works Research Institute, 1 Banchi, Asahi, Tsukuba #305, Japan. N. Okamoto, Department of Civil Engineering, Tokyo Institute of Technology, 4259, Nagatsuda-cho, Midori-ku, Yokohama #226, Japan.

samples from the population. Choice-based surveys were utilized; these were conducted in recreational sites to estimate trip chaining behaviors in recreational areas and combine the data with personal experience data of home-based surveys. Additionally, to estimate trip attraction volumes in the recreational areas, number plate surveys were conducted at several sites and access roads to the areas.

These surveys were conducted in nine regions by eight regional construction bureaus of MOC and the Hokkaido Development Agency in 1992 or 1993. Home-based surveys were conducted between July and October 1992 by random sampling of households in the selected areas. The areas cover 22 cities in 19 prefectures that belong to nine regions in Japan, as indicated in Figure 1. The surveyed areas are concentrated in large cities, including cities in every metropolitan region and several central cities in local areas such as Sapporo in the Hokkaido region, Sendai in the Tohoku region, Hiroshima in the Chugoku region, and Fukuoka in the Kyushu region. These indicate that home-based surveys in NRTS evidently represent characteristics of recreation activities of urban residents. The respondent of a survey is required to be more than 18 years old. Total individual samples exceed 30,000, encompassing 13,600 households. All except for three cities have more than 1,000 samples.

Choice-based surveys were conducted in nine specific popular sites corresponding to the nine regions in Japan. Most of the surveys were conducted on weekends in August during summer vacation, with the exception of Bandai in Tohoku region in October and Yuzawa in Hokuriku region in February 1993. These schedules were determined by the top reasons for and seasons of recreation: sightseeing of autumn foliage in Tohoku and skiing in Hokuriku. The sample size for all sites is nearly 13,000, and the response ratio is about 10 percent. The survey was distributed by hand and returned by mail.

Structure of Questionnaires in Both Surveys

The personal questionnaire for home-based surveys inquires of the respondents their annual travel records by mode, overseas travel experiences, and travel activities that correspond to the specific date of the choice-based survey in the region.

Because respondents for choice-based surveys are persons who drove by private car to a recreational region, most questionnaires are concerned with car travel. The major difference between the home-based and choice-based surveys is that in the choice-based survey questionnaire respondents are requested to explain their route patterns on the map of the recreational area. As mentioned earlier, this is distinctly a different questionnaire from that for urban transportation surveys, and such information should be collected for road planning in recreational areas.

Figure 2 presents a brief structure of surveyed items and primary goals in both surveys. The primary objective of home-based surveys

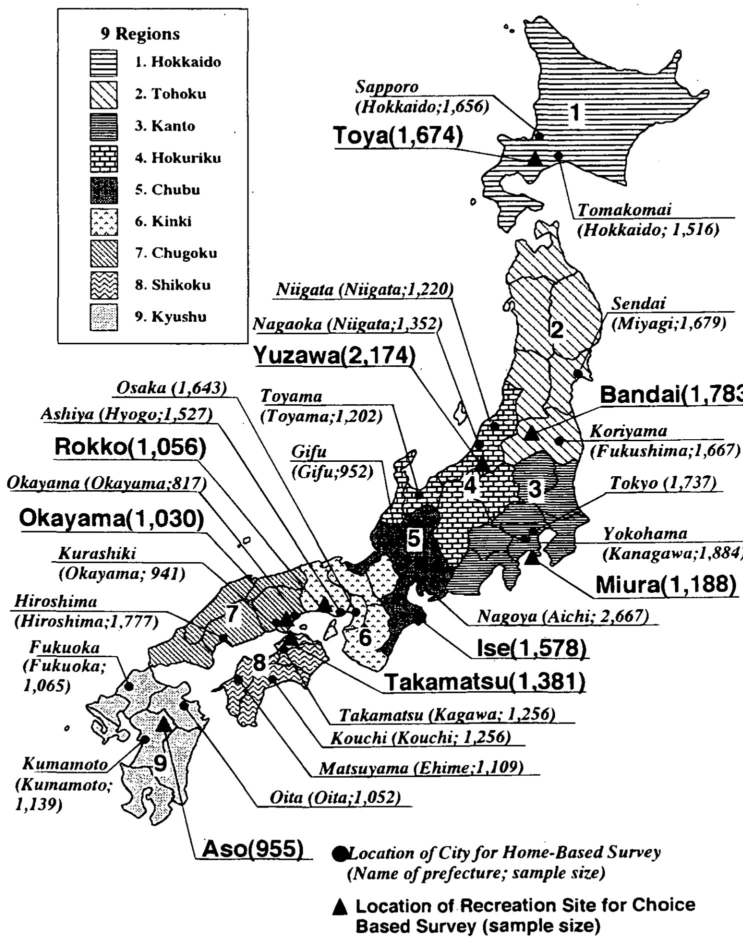


FIGURE 1 Locations of home-based and recreational site surveys.

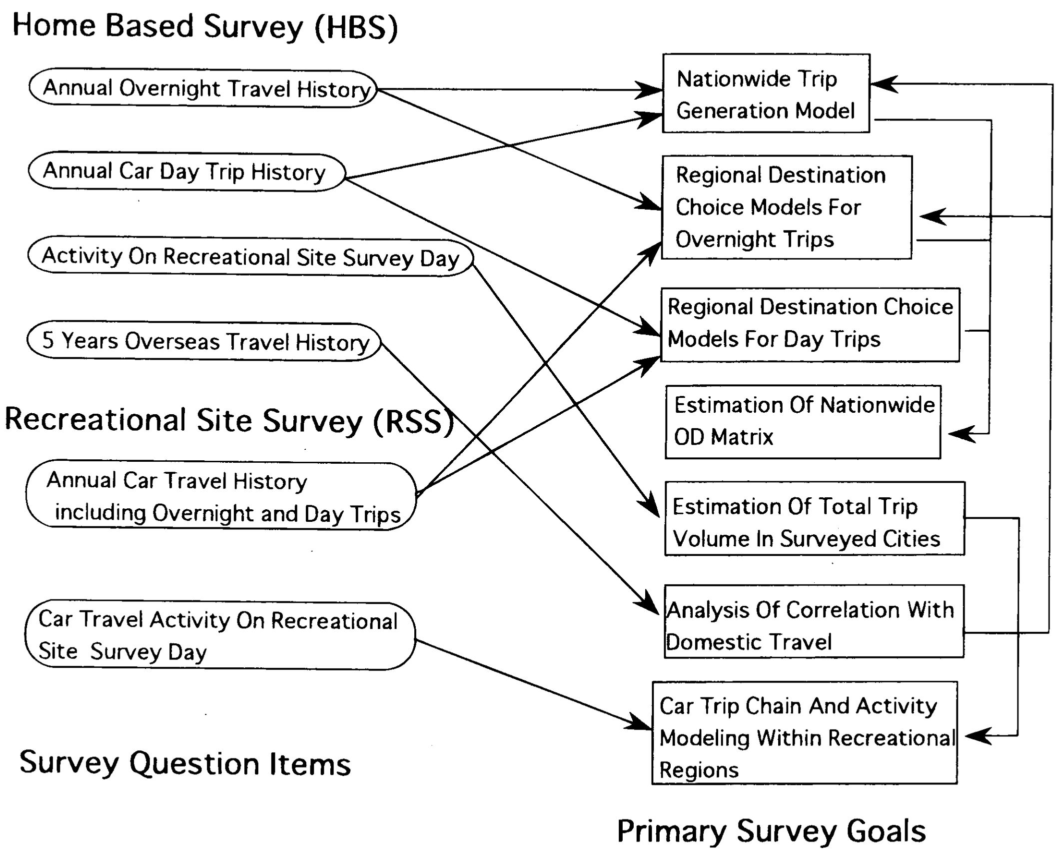


FIGURE 2 Surveyed items and primary goals in home-based and recreational site surveys.

is to collect data for the development of a nationwide trip generation model and regional destination choice models. On the other hand, the main purpose of choice-based recreation site surveys is to collect data for modeling of trip and activity chain behaviors within a recreational region. However, the questionnaires for both surveys supplement each other. Travel activity on the designated day of the recreational site survey was obtained from the home-based survey to examine the total travel volume from the city and portion of the total volume headed into the recreational region on that day. The annual car travel record was surveyed for recreational site survey respondents to combine them with home-based survey data to develop car travel destination choice models.

PROFILES OF RECREATIONAL TRAVEL IN JAPAN

Using survey results, various profiles of recreational travel have been examined. Particularly overnight travel, day trips, route choice, and trip chaining behaviors in recreation areas, and the correlation between domestic and overseas travel, are briefly investigated.

Because of lack of space, only the profile of overnight travel from home-based surveys is introduced in this paper. Although the recreational day trips exceed the overnight trips in volume in metropolitan suburban areas (the percentage of persons on day trips in the recreation site survey is 67 percent in Miura, Kanto region, and 74 percent in Rokko, Kinki region), principal recreational areas and most spa resorts in Japan often attract a larger percentage of overnight trips. (The percentage of overnight travelers in the recreation site survey is approximately 65 percent in Yuzawa in the Hokuriku region, Ise in the Chubu region, Okayama in the Chugoku region, and Aso in the Kyushu region.) An increase in overnight travel is expected in the future because of a long series of holidays and the growing need for recreational activities. As a result, an understanding of the mechanisms of overnight recreation travel and a comparison of them with day trips are required before examining the total weekend traffic and the system planning of suburban road networks (*3*).

PROFILES OF OVERNIGHT TRAVEL IN JAPAN

Average frequencies of overnight travel by age categories are shown in Figure 3. The frequency of those aged 70 or older is no more than half that of those aged 30 or older. Because this includes business travel, the categories for the 30 and 40 year olds have a somewhat larger frequency in total. Although recreation travel is the lowest in the 40s category among the working ages between 20 and 60, their business travel is possibly combined with recreation activities.

Average trip frequencies for individual income levels are shown in Figure 4. The total frequency increases in accordance with an increase in income level. The increasing trend is moderate for recreation travel. This is because business travel is highly generated in higher-income categories.

The modal choices for recreational trips are summarized in Figure 5. Car usage occupies the largest share in most of the categories of vacation period, and this share depends on the departure date and season. The reason why the share is largest during the summer vaca-

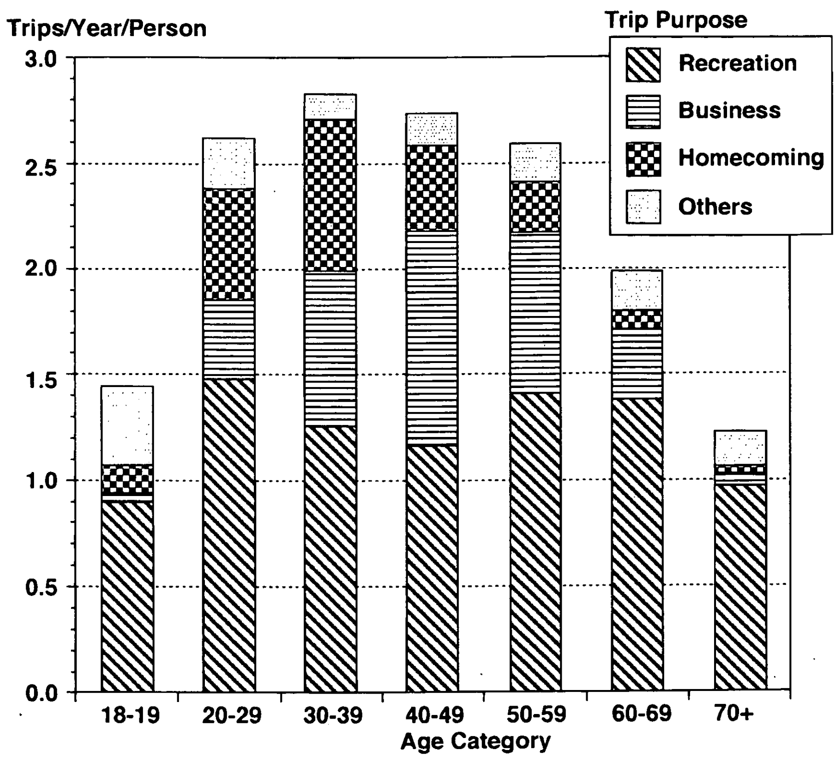


FIGURE 3 Annual overnight trip frequency per person by age category and trip purpose.

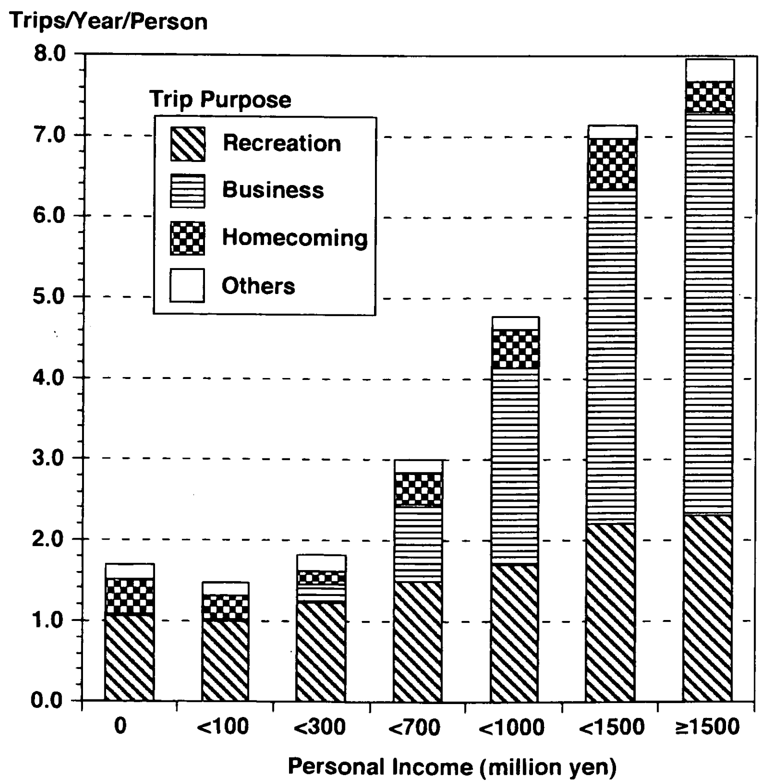


FIGURE 4 Annual overnight trip frequency per person by income category (personal) and trip purpose.

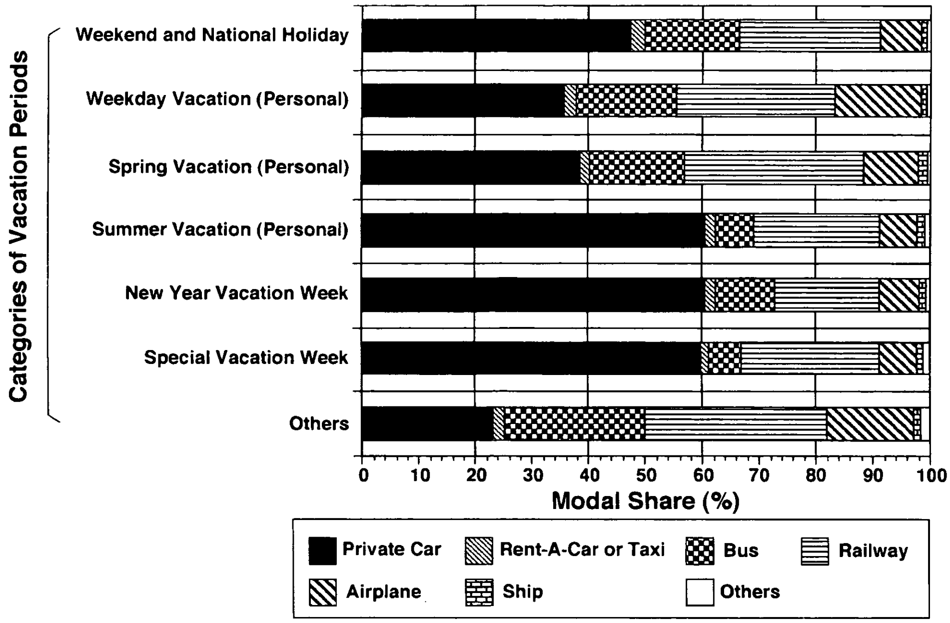


FIGURE 5 Modal shares of overnight recreational trips by vacation type.

tion in August or during the New Year holidays in late December and early January is that travelers prefer to drive cars with their families during those periods.

TRIP GENERATION MODELS

Regional Characteristics of Trip Generation

Modeling of trip frequency of recreational activities in Japan has been achieved by a disaggregated linear regression method, in which the objective variable is the individual travel frequency and the explanatory variables are personal demographic data, such as age and gender. Samples to estimate the linear function came from the former survey data by the Japan Tourist Association (JTA). Aggregate models using zonal average data could not be calibrated from JTA data because of the restriction in sample size. However, the coefficients of determination in the disaggregated linear functions were insufficient for the prediction of future trip generation (*2*).

Figure 6 shows regional differences in the recreation trip frequencies by prefecture. Japan is composed of 47 prefectures, in which the largest prefecture, Tokyo Metropolis, has a population of

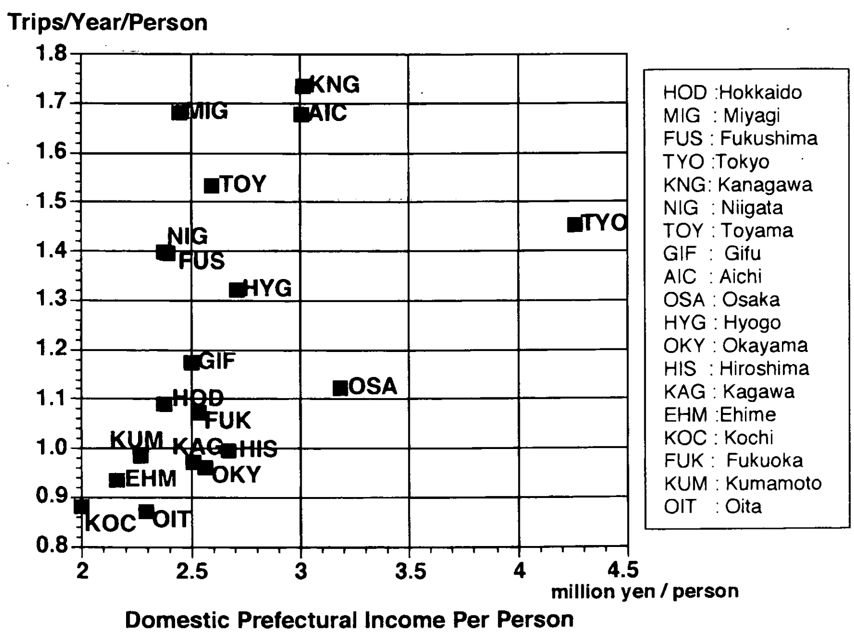


FIGURE 6 Correlation between overnight recreation trip frequency and domestic prefectural income for surveyed cities.

nearly 12 million and the smallest prefecture has a population of only 0.6 million population. Every surveyed city indicated in the figure corresponds to a prefecture described in Figure 1. Prefectural average frequencies from samples and income levels measured by domestic prefectural income statistics (thousand yen per capita) have a fairly strong correlation. This causes one to consider not only individual factors but also regional effects in trip frequency, such as regional income levels.

Regional differences in trip generation may depend on the difference of transportation facility conditions, such as the accessibility to expressways. Figure 7 shows the relationship between trip frequencies and general road densities (kilometers per square kilometers) in prefectures. A positive relation between these quantities is observable for prefectures whose trip frequencies exceed 1.0. However, the prefectures with a trip frequency lower than 1.0 are concentrated at the position of relatively higher road densities. Accordingly, the prefectures with lower frequency should be explained by factors other than road density. Because local prefectures in Hokkaido, Shikoku, and Kyushu have some attractive recreation sites in day trip areas, residents may have little motivation to take overnight trips to other regions.

These results imply that trip generation models should include regional factors in addition to individual characteristics. Although disaggregate models estimated by samples from several regions could contain regional variables, calibrating aggregate models to include regional factors and improve models' predictability is now possible with NRTS data.

Cross-Categorical Aggregated Model

To introduce both personal characteristics and regional factors into the model, a cross-categorical aggregated (CCA) model formulation was developed. Aggregate generation model is usually described by

$$Y_r = \sum_{j=1}^{J} \beta^j x_r^j \tag{1}$$

where

β^j = coefficient,
x^j = explanatory variable aggregated by zone (here by prefecture),
J = number of variables, and
Y_r = average trip frequency for every prefecture: r.

On the other hand, for categorical data classified according to demographic attributes, the volume in a cell of a multiple cross table is regarded as a sample in the estimation of models. The volume in a cell corresponds to the average trip frequency of multiple categories, such as the combination of age and income. Using such samples,

$$y_g = \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha^{mk} \delta_g^{mk} \tag{2}$$

will be a model formulation of trip generation y_g of multiple categories, g. In Equation 2, g is identical to a cell; M is number of factors; K is the number of categories in each factor; δ_g^{mk} is 1 if k of factor m corresponds to the category of a cell g, and otherwise it is 0; α^{mk} is the parameter of a category in a factor.

Therefore, integration of regional and aggregated demographic data is performed by the following joint equation:

$$y_{rg} = \sum_{j=1}^{J} \beta^j x_r^j + \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha^{mk} \delta_g^{mk} \tag{3}$$

where y_{rg} is an average trip frequency in a cell g of prefecture r. As categorical demographic data and regional factors are combined in the estimation of an aggregate model, the model is the CCA model. The parameters β and α in CCA models are estimated by the weighted least-squares method. Sample sizes in cells are used as weights in the estimation.

For example, if two categories are considered by gender and five categories are considered by age for 20 regions, 10 (2 × 5) cells are produced and samples in the estimation will be 200 (10 × 20). Therefore, the sample size in the estimation depends on the size of

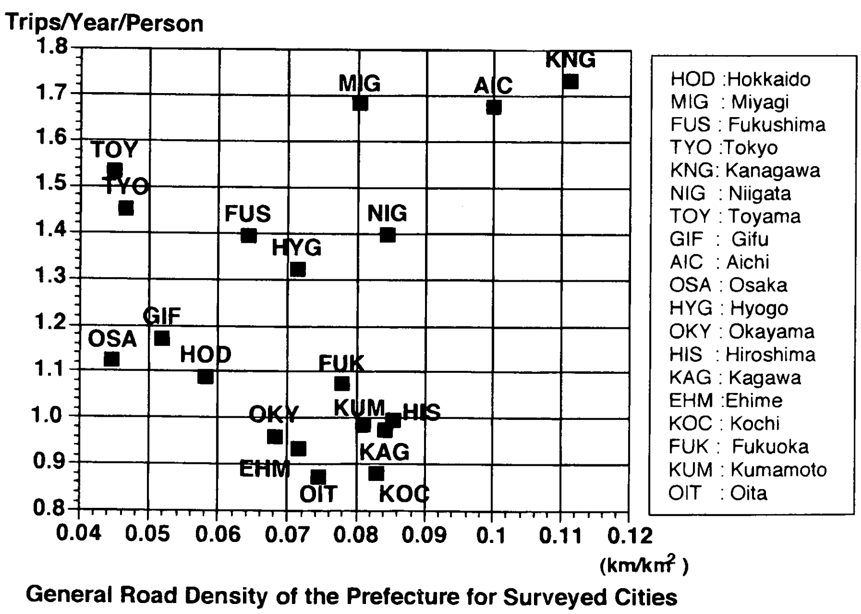


FIGURE 7 Correlation between overnight recreation trip frequency and general road density in prefecture for surveyed cities.

cells. The parameters that represent the correlation among factors, such as age and income, are also available in the models.

Estimation Results of CCA Models

The estimation results of CCA models are presented in Table 1. Three kinds of regional variables—the prefectural income per capita, the car ownership ratio, and the general road density—were introduced in every model to explain regional differences of trip generation. Regional variables differ by prefecture, not by city or household. The individual income level was incorporated in every model because of its evident importance to the models. One additional factor was selected to increase the explanatory power because introducing three or more personal factors increases the cell size and decreases the model's reliability. The adjusted correlation coefficients are fairly high compared with those of ordinary models for recreational trip generation (*2*). The model using gender or car ownership as a personal variable has a higher correlation coefficient. The parameters of the individual income variable increase in accor-

TABLE 1 Estimation Results of CCA Models for Overnight Recreational Trip Frequency

variables		model 1	model 2	model 3	model 4	model 5
Regional characteristics						
regional income per person		0.3913	0.3348	0.3352	0.3288	0.2713
[million yen]		(5.63)	(5.13)	(6.52)	(3.44)	(4.21)
share of car-ownership		0.02267	0.02590	0.02686	0.02706	0.02588
[%=vehicle/ 100person]		(2.55)	(3.07)	(4.05)	(2.16)	(3.12)
index of road development		4.995	5.090	5.304	4.818	4.501
		(2.87)	(3.07)	(4.08)	(2.00)	(2.76)
Demographic characteristics						
car-ownership	no					
	yes	0.3571				
		(4.56)				
sex	male		-			
	female		0.29921			
			(4.03)			
age	≤ 29			-		
	30- 49			-0.22073		
				(3.26)		
	50- 64			-0.068590		
				(0.94)		
	≥65			-0.25887		
				(3.30)		
frequency of holiday						
	≤6days per month				-	
	≥7days per month				0.2806	
					(2.39)	
	housewife & student				0.5695	
					(2.02)	
passport possession						
	no					-
	yes					0.5686
						(9.38)
personal income	<100	-	-	-	-	-
[million yen]	<700	0.2935	0.4670	0.3008	0.6354	0.2393
		(4.55)	(6.37)	(6.26)	(2.28)	(3.95)
	<1000	0.5303	0.8492	0.5910	0.9036	0.44607
		(3.94)	(5.89)	(5.68)	(2.78)	(3.55)
	≥1000	1.042	1.358	1.018	1.476	0.7766
		(5.55)	(7.43)	(6.24)	(3.70)	(3.93)
const		-1.3097	-1.2552	-0.88534	-1.4481	-0.95945
		(3.42)	(3.43)	(3.07)	(2.46)	(2.71)
sample size		117	112	202	118	123
multiple correlation coefficient adjusted for the degrees of freedom		0.8004	0.8098	0.7474	0.6286	0.7934

dance with the increase in income level. Three regional variables also are adequately significant, and increases in these variables enlarge the trip frequency. Most of these results correspond with aggregation results of survey data and experimental facts. The estimation of these models was successful because of NRTS, the first large-scale survey.

DESTINATION CHOICE MODELS

Destination choice depends on activity, season, travel time, travel cost, attractiveness of destination, and so forth. Modeling trip distribution was traditionally conducted by aggregate models such as the gravity model or the present pattern method used in the planning of metropolitan regions. Selection of model forms sometimes depends on the sample size for the estimation.

NRTS has a large sample size in total. If total samples are pooled, as in trip generation modeling, an aggregated distribution model of the whole country is applicable. However, before pooling total samples, comparing regional differences in destination choice behavior and examining the applicability of the distribution model to recreational travel are required. This is because alternative destinations in a region differ from those in other regions and the parameters of the models may be different among regions.

In this section, the applicability of destination choice models was examined using a disaggregate approach after some previous studies (*4,5*). The approach has the advantage of modeling individual behaviors with relatively small samples and also has a form similar to that of the aggregated distribution model.

Destination Choice Behaviors of Recreational Travel

Figure 8 indicates an example of the distribution of travel destinations for overnight and day trips from Tokyo and Yokohama home-based surveys. The circle indicates the sample size for destination. Overnight and day trip percentages are indicated within the circle. Trip destinations beyond the described area are not illustrated here, in spite of the existence of a few long-distance trips. Distributions of day trips are in accordance with distance from the origin, whereas distributions of overnight trips for nearby destinations represent a small portion of the total.

Data for Destination Choice Models

Car travel data from a personal record of home-based and recreational site surveys were pooled for each origin region to estimate destination choice models for overnight and day trips. A multinomial logit (MNL) model was employed, and travel record data from July through December were selected. The sample size for each model is presented in Table 2 and is discussed later. Not including the trips on the surveyed day, travel record data from recreational site surveys are assumed to be data from exogenous sampling for the corresponding destination choice model. Estimation of MNL model parameters with such samples is possible because of the former works on the estimation theory of discrete choice models.

With chosen destinations arranged in order of decreasing percentage of samples for each region, those within the top 90 percent (cumulative) were selected for the choice set. This 90 percent threshold provides a maximum choice set size for each region. The largest regional choice set size is 30 in Chubu and Kinki for

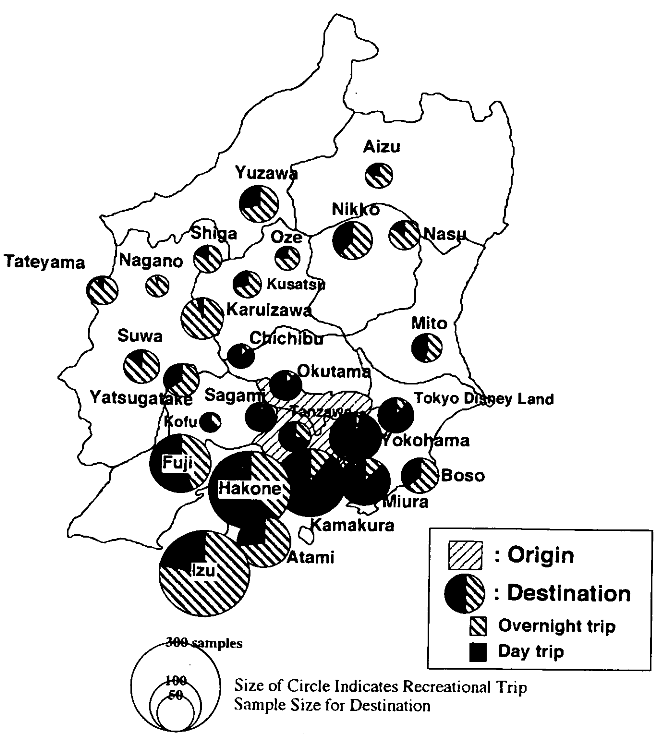


FIGURE 8 Sample size for recreational destinations from Tokyo and Yokohama home-based survey.

overnight travel destinations, and the smallest size is 11 in Shikoku for day trips. These data are also given in Table 2.

Because a destination choice set depends on personal activity interests, the destination alternatives that have no recreational resources corresponding to an individual's activity were excluded. (It is obvious that no one goes to the seaside to climb a mountain.)

Regional utility functions are composed of six variables: travel time, travel cost, and four attraction variables. Travel time between origin and destination was calculated using road network data and shortest-path algorithm. The travel cost variable was transformed by dividing out-of-pocket costs by the logarithm of personal income. Attraction variables are combined with attractiveness of destination and a personal activity dummy. The attraction variable is expressed as

$$\text{Activity attraction of } k = (k \text{ activity dummy}) * [\ln (k \text{ attraction resources})] \tag{4}$$

$$k \text{ activity dummy} = \begin{cases} 1 & \text{if traveler's activity is } k \\ 0 & \text{if traveler's activity is not } k \end{cases}$$

where k attraction resources is the number of attraction resources corresponding to activity k in a destination. For example, only travelers who participate in seaside or marine activity have an attraction variable of "seaside and marine activities." The attraction resources of each destination were obtained by summing up the number of recreational spots recognized by the Japan Travel Bureau.

Estimation Results of Destination Choice Models

Comparisons of model parameters among regions are discussed here, with consideration given to future integration to a nationwide common model. Previous survey samples never enable the re-

TABLE 2 Estimation Results of Destination Choice Models for Overnight and Day Recreational Trips

(a) Overnight Trip Destination Choice Models

	Surveyed Area								
Variables	Hokkaido	Tohoku	Kanto	Hokuriku	Chubu	Kinki	Chugoku	Shikoku	Kyushu
Travel time	-0.00178	-0.00454	-0.003	-0.0038	-0.000024	-0.00203	-0.00108	-0.00458	-0.00266
t-statistics	-8.58	-8.0	-5.8	-8.46	-0.05	-4.6	-1.87	-8.48	-2.8
Travel cost / ln(Personal Income)	-0.000287	-0.000455	-0.000246	-0.000406	-0.00105	-0.000689	-0.00106	-0.000055	-0.00086
t-statistics	-2.12	-4.11	-2.48	-4.21	-11.5	-7.25	-9.4	-0.28	-4.37
Seaside & Marine Activity Attraction	0.326	-0.107	0.319	-0.123	0.144	-0.235	0.736	0.0336	0.123
t-statistics	6.04	-1.8	4.78	-1.96	2.75	-3.54	8.14	0.31	0.8
Field Activity Attraction	0.0682	0.552	0.596	0.682	0.828	1.12	0.595	0.305	0.762
t-statistics	0.73	5.75	8.39	3.65	10.0	12.23	4.32	2.83	3.61
Spa Visit Attraction	0.392	0.545	0.435	0.73	1.3	0.0303	0.0611	0.612	1.32
t-statistics	5.99	4.48	4.48	5.11	12.6	0.26	0.35	3.24	5.78
Sightseeing Attraction	0.706	0.303	0.923	0.510	0.722	0.422	1.35	1.45	1.39
t-statistics	11.5	3.26	14.7	4.46	11.6	4.84	10.1	8.60	13.9
Log-Likelihood at zero	-2594.7	-2271.9	-3071.8	-1898.9	-3192.1	-2471.8	-1372.0	-1025.5	-1303.2
Log-Likelihood at convergence	-2400.1	-2040.5	-2786.0	-1737.0	-2897.1	-2277.6	-1183.6	-906.5	-1051.2
Sample Size	1000	802	1000	614	1000	765	442	353	469
Choice Set Size	15	21	27	27	30	30	26	21	19

(b) Day Trip Destination Choice Models

	Surveyed Area								
Variables	Hokkaido	Tohoku	Kanto	Hokuriku	Chubu	Kinki	Chugoku	Shikoku	Kyushu
Travel time	-0.00815	-0.0156	-0.0164	-0.0125	-0.00602	-0.00464	-0.00876	-0.00899	-0.00971
t-statistics	-22.9	-11.1	-17.7	-23.7	-7.64	-5.26	-13.4	-16.2	-11.9
Travel cost / ln(Personal Income)	-0.000237	-0.00143	0.000276	-0.00128	-0.00275	-0.00148	-0.000856	-0.00215	-0.00273
t-statistics	-1.57	-6.04	1.48	-9.07	-15.4	-6.78	-4.65	-4.92	-11.3
Seaside & Marine Activity Attraction	0.134	0.494	0.543	0.128	0.316	0.0936	0.296	0.434	0.494
t-statistics	2.6	8.14	8.54	1.36	5.81	1.63	2.79	4.67	4.1
Field Activity Attraction	0.196	0.898	1.01	0.572	0.485	0.779	1.62	0.495	1.12
t-statistics	1.83	7.1	12.5	4.49	5.36	6.48	6.05	5.19	5.39
Spa Visit Attraction	0.815	0.83	0.617	0.517	0.844	0.917	0.276	1.27	2.67
t-statistics	4.47	5.52	6.84	3.55	6.52	3.48	1.4	4.81	8.57
Sightseeing Attraction	0.674	0.0918	0.788	0.28	1.06	-0.585	0.845	1.63	1.23
t-statistics	5.68	0.38	6.4	1.64	5.1	-1.05	3.77	7.21	6.98
Log-Likelihood at zero	-2558.8	-2624.5	-2598.1	-2445.0	-2598.7	-1553.3	-1494.5	-1667.8	-1754.0
Log Likelihood at convergence	-1988.2	-1629.8	-2252.9	-1451.9	-2163.8	-1387.4	-1129.4	-1225.5	-1154.1
Sample Size	1000	1000	922	939	933	577	626	744	660
Choice Set Size	15	16	23	17	18	18	13	11	17

searchers to examine such a comparison for recreational trip distributions, in spite of unique recreation sites in each region.

Using samples from home-based and recreation site surveys, regional destination choice models were estimated for overnight and day trips (Table 2).

Estimation results indicate that most regions have reasonable and expected signs of parameters. Parameters of travel time in Chubu and travel cost in Shikoku are insignificant for overnight trips. The attraction variable for sightseeing in Tohoku, Hokuriku, and Kinki has an unexpected sign for overnight trip models. For day trip models, the travel cost variable in Kanto and an attraction variable for a spa visit in Kinki had unacceptable parameters. However, most parameters had reasonable results, which is useful for the future integration of models.

The sample size and log likelihoods are given in Table 2. Sample sizes in most regions exceeded 1,000 and log likelihood ratios stand between nearly 0.1 and 0.4, permitting a comparison of the models.

Figure 9 shows regional differences of two important parameters of overnight trips. The intersection of two lines in the figure indicates expected values of parameters, and the line length indicates the standard deviation of the parameter. It seems that there are two significantly different groups of parameters, except for the parameters for the Shikoku and Chubu region. One is formed by the parameters for East Japan: Hokkaido, Tohoku, Kanto, and Hokuriku regions. The other is formed by the parameters for West Japan: Kinki, Chugoku, and Kyushu regions. West Japan has larger cost and smaller time parameters than East Japan. Relatively speaking, this implies that the West is cost conscious and the East is time conscious. Results of day trip models were similar, although the positions of Chugoku and Shikoku changed. As a result, different regions may have different parameters of time and cost, and the regional combination to make a few model segmentations, such as east and west Japan, is possible to represent trip distributions in Japan. Furthermore, the fact that parameter trade-off ratios between time and cost variables, in regions in which the *t*-statistics of both parameters exceed 2.0 (Tohoku, Hokuriku, Kinki, and Kyusyu), were nearly identical among overnight and day trip models suggests the possibility of integrating overnight and day trip models.

CONCLUSIONS

The outline of the first large-scale survey of recreation travel in Japan is summarized and the applicability of survey data to trip generation and distribution models is briefly examined. Recreational travel volume is definitely increasing in Japan, and the improvement of transportation facilities is expected in several recreational areas. Understanding recreational trip behavior is essential to revising the road planning in suburban areas.

The results of NRTS provided fundamental characteristics of recreational travel, which differ by demographic attributes and regional factors. The modeling abilities from NRTS were also

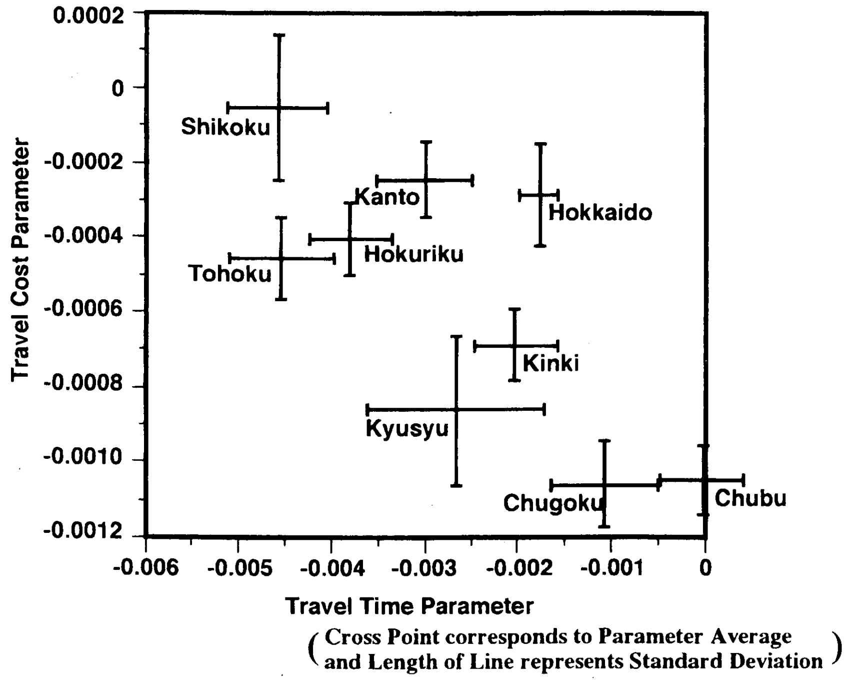


FIGURE 9 Regional difference of time and cost parameters of destination choice models for overnight recreational trips.

conformed by successful results of trip generation and destination choice models. The trip generation models had much higher correlation coefficients than other models from previous studies in Japan. Destination choice models in most regions provided significant and reasonable parameters and the possibility of regional data pooling.

However, several fruitful researches using NRTS data are still unexplored. Ongoing and additional research on the following aspects will activate NRTS potentials for planning fields: (*a*) establishment of a survey method using home-based and choice-based recreation site samples to estimate a nationwide origin destination matrix; (*b*) integration of destination choice models to improve their statistical accuracy and stability; (*c*) consideration of travel duration and interval in the trip generation process to improve the model's predictability; (*d*) modeling of intraregional travel behaviors in several regions to establish the prediction methods of recreational traffic volume in recreational areas.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the valuable discussions with Hideki Nakamura, Tetsuro Hyodo, Hiroshi Wakui, Toru Nishimura, and Michelle Franciose in the calibration of the models and their contributions to this paper. The authors also would like to thank the anonymous referees for their useful suggestions about this paper.

REFERENCES

1. Okamoto, N., S. Morichi, and T. Yai. A Study on Road Network from the Environmental Viewpoint. Presented at World Conference on Transport Research, Lyons, France, 1992.
2. Yamada, H., T. Yai, H. Nakamura, and T. Hyodo. A Trip Generation Model Using Nationwide Survey for Recreation Trips (in Japanese). *Traffic Engineering,* Vol. 29, No. 2, 1994, pp. 19–28.
3. Yai, T., and N. Okamoto. Fundamentals of Transport Planning Based on the Characteristics of Infrequent Trips in Recreational Regions (in Japanese). *City Planning,* Vol. 25, No. 6, 1994, pp. 28–34.
4. Morichi, S., and T. Yai. Application of Disaggregate Travel Demand Model and Choice Based Sampling for Infrequent Trips (in Japanese). *Proc., JSCE,* No. 343, 1984, pp. 161–170.
5. Yai, T. Disaggregate Behavioral Models and Their Applications in Japan. *Transportation Research A,* Vol. 23A, No. 1, 1989, pp. 45–51.

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

Investigating Effect of Travel Time Variability on Route Choice Using Repeated-Measurement Stated Preference Data

MOHAMED A. ABDEL-ATY, RYUICHI KITAMURA, AND PAUL P. JOVANIS

A study was conducted to determine ways in which travel time variation affects route choice behavior and the potential interplay among travel time variation, traffic information acquisition, and route choice. In a computer-aided telephone interview, a stated preference section was included to investigate this issue, and 564 respondents in the Los Angeles area gave their choices to five hypothetical binary choice sets. The repeated measurement issue is addressed with individual-specific random error components in a binary logit model with normal mixing distribution. The results indicate the significance of both the degree of travel time variation and traffic information on route choice and illustrate the viability of the survey methodology used. The study also underscores the need for a statistical correction to account for the correlation among error components in repeated-measurement data.

In recent years, with an increased desire for better urban transportation systems arising from environmental and increased levels of traffic congestion concerns, there has been an increased need for better modeling in the transportation planning process. Much of the emphasis has been on gaining a better understanding of an individual's route choice behavior. It is in the area of traffic assignment that a better understanding of that behavior would be beneficial.

TRAVEL TIME UNCERTAINTY AS CONTRIBUTING FACTOR TO ROUTE CHOICE

Several empirical studies have examined the factors affecting drivers' route choice. In the urban context the governing relationship is not clear; some researchers have concluded that time minimization is the dominant criterion, whereas others have noted the importance of other factors, such as road type (*1,2*); avoidance of congestion (*1*); and avoidance of stops and traffic signals (*3*).

The reliability of a particular route can be expected to play an important role in the traveler's route choice behavior. In several attitudinal studies, reliability-related attributes have been found among the most important service attributes in a variety of situations (*4*). Black and Towriss (*5*) indicated that travelers are likely to suffer disutility because of the uncertainty or unreliability in travel times. However, the effect of travel time variation has been rarely investigated in route choice studies. In an empirical study by the authors (*6*), travel time reliability was found as one of the most important factors for route choice, with about 54 percent of respondents in a route choice survey indicating that travel time reliability is either the most important or second most important reason for choosing their primary commute routes.

The Wardrop user equilibrium model states that travelers choose the fastest available route; it implies that they always choose the same route on repeated trips (*7*). However, travelers are not always capable of identifying the fastest route, and if travel time is uncertain, they may wish to acquire additional information that helps to select a better route. Therefore, investigating the effect of travel time reliability is significant in understanding the impact of traffic information on route choice.

Several studies by the authors have investigated the effect of numerous criteria on route choice behavior (*8–10*). The main objective of this study is to explore one measure of reliability—travel time variability—and assess its importance on route choice. The possible interplay between traffic information, travel time variability, and route choice will also be addressed. Five stated preference choice sets were included in a route choice survey to investigate the effect of travel time variation on route choice. This repeated measurement data set is used in the modeling effort presented in this paper.

REVIEW OF DISCRETE CHOICE MODELS WITH REPEATED-MEASUREMENT DATA

Discrete choice models typically are estimated on the basis of revealed preferences, with a single choice made by each respondent in the sample. Under these conditions, the disturbance term (ε) accounts for the taste variation from one decision maker to another. In contrast to the revealed preference approach, repeated hypothetical choice sets are often presented to the decision makers in the stated preference approach.

The estimation of a discrete choice model with repeated observations for each respondent gives rise to an obvious correlation of disturbances, or heterogeneity, which refers to variations in unobserved contributing factors across behavioral units. If behavioral differences are largely caused by unobserved factors, and if unobserved factors are correlated with the measured explanatory variables, then estimates of model coefficients will be biased if this heterogeneity is not taken into account. The problem may be more pronounced in repeated measurement data because such unobserved factors may be invariant across the repeated measurements. In this paper, an error component method is used to account for unobserved heterogeneity and correct for potential bias that would otherwise arise.

Many studies, for example, Bunch et al. (*11*), ignored the effect of heterogeneity by indicating that in a small number of repeated observations by each individual the properties of parameter estimates themselves do not rely on the strict independence assump-

M. A. Abdel-Aty, Department of Civil and Environmental Engineering, University of Central Florida, P.O Box 162450, Orlando, Fla. 32816. R. Kitamura and P. P. Jovanis, Institute of Transportation Studies, 2028 Academic Surge, University of California at Davis, Davis, Calif. 95616.

tion, and the benefits of using a much larger pooled data set more than outweigh this concern.

Mannering and Winston (*12*) presented a dynamic model composed of nested-logit models of car ownership level and vehicle type choice, combined with linear utilization models. The paper emphasizes dynamic aspects of car ownership and utilization behavior, for example, stationarity and state dependence. However, it neglects completely the possible intertemporal correlation in the error terms. Mannering (*13*) discusses the same model system but assumes that disturbances are serially independent because of the difficulty in accounting for serial correlation in the presence of lagged endogenous variables in discrete choice models. Hocherman et al. (*14*) estimated a nested logit dynamic household vehicle transaction model assuming that serial correlation is not present.

Louviere and Woodworth (*15*) corrected the standard errors produced by a repeated responses regression model by multiplying the standard errors by the square root of the number of repeated observations. Mannering (*16*) estimated a vehicle choice logit model with repeated observations and also used the same correction procedure. However, this method is said to be a conservative approach and tends to overcorrect the value of the standard errors (*15*) (or *t*-statistics when divided by the square root of the number of observations for each respondent).

A number of other discrete choice panel data models have been discussed in the literature, usually limited to the dichotomous case. One of the oldest models is the beta-logistic model proposed by Heckman and Willis (*17*). In this model heterogeneity is introduced by specifying the beta distribution as a mixing distribution on the outcomes. The exogenous variables are assumed to be time invariant. The presence of heterogeneity in mode choice models is shown in Uncles (*18*) also using a beta-logistic model.

Kitamura and Bunch (*19*) used a dynamic ordered-response probit model with error components of car ownership. This approach allows more flexible formulation of the error terms and thus offers a better accounting of heterogeneity than do the beta-logistic models suggested by Heckman and Willis (*17*) and Uncles (*18*). Morikawa (*20*) also used logit models with error components to treat serial correlation (heterogeneity) between the error terms of revealed and stated preference models.

Incorporating the effect of the correlation of disturbances into repeated observations, discrete choice models must be addressed explicitly if unbiased estimates of the structural parameters are to be obtained. This paper is concerned primarily with the empirical results investigating the effect of travel time variation on route choice. However, heterogeneity will be accounted for by using a parametric functional form (normal mixing distribution and Gaussian quadratures estimation). Comparative analysis will be performed using the pooled data and applying the heuristic correction procedure suggested in other studies (*15,16*) and using one observation randomly drawn from each respondent. A subsequent paper will concentrate on different specifications of the error components, that is, parametric estimation with different distributions and nonparametric estimation.

ROUTE CHOICE SURVEY

An ongoing effort for the Partners for Advanced Transit and Highways at University of California, Davis, is to investigate the actual route choices of drivers, with the objective of developing refined route choice models that can include the effect of traveler information.

To probe into drivers' route choice behavior, a computer-aided telephone interview (CATI) of Los Angeles-area morning commuters was conducted. The survey, undertaken in May and June 1992, was designed to investigate how much information drivers have about their routes; their awareness of alternate routes; their awareness of traffic conditions, which could affect their route choices; and their use of available traffic information either en route or pretrip, or both. A detailed description of the survey design and descriptive statistics are included in a research report by the authors (*8*), and models of information use and route choice and of commuters' frequency in changing routes are reported in Abdel-Aty et al. (*9*).

A second CATI survey was designed and conducted in May 1993. Its objectives were to probe further into drivers' route choice behavior, to measure any changes within the last year, to investigate commuters' attitudes and perceptions about several commute characteristics, and to understand the effect of travel time variation on route choice. The survey targeted the same sample interviewed in May and June 1992. A maximum of 10 callbacks were attempted before abandoning a respondent's number, which yielded 564 (about a 60 percent response rate) completed interviews (1 year after the first survey of May and early June 1992). Abdel-Aty et al. (*21*) describe the survey design, and introduce general descriptive statistics that show commuters' perceptions, preferences, and decisions in route choice. Factor analysis was performed to investigate the commuters' perceptions of several commute route characteristics.

This paper is concerned with the last objective of the survey, which is to measure and investigate whether commuters choose a route that is longer but more reliable or a route that is shorter but has uncertain travel times and to what extent uncertainty affects route choice. The paper presents models of the effect of travel time variation on route choice and the possible interplay between travel time variation, traffic information, and route choice.

DESCRIPTION OF HYPOTHETICAL CHOICE SETS

The advantage of using revealed preference (RP) data is that resulting models are based on the observation of actual behavior, not on respondents' responses to questions about their intentions. However, the family of market research survey techniques, termed stated preference (SP) methods, has been used often in transportation planning over the past decade [e.g., Morikawa (*20*) and Khattak et al. (*22*)]. Such methods are now becoming seen as a complement to the more traditional RP survey methods in cases where the latter cannot provide the full information needed for analysis. Investigating the effect of travel time variation on commuters' route choice would be difficult largely because it is time consuming to collect data that support the analysis. Therefore, in the context of this study there is no alternative but to solicit preferences in hypothetical settings, as is often done in many marketing research contexts.

It was therefore decided to include repeated hypothetical choice sets in the CATI survey. A major concern was that the design of SP choices could be complicated because the intention was to quantify the tradeoffs between a reliable but slow route versus an unreliable but fast route. It was intended also to make the design of the choice sets as easy as possible to be understood on the telephone, which was the medium chosen for the survey (in mail questionnaires more complete and complicated SP choice scenarios can be formulated, whereas in telephone interviews there is a limitation to what a respondent can comprehend and visualize). Another concern was that the degree of travel time variation needed to be as realistic as possible. The SP choices were thus designed to be as simple as possible so that respondents can comprehend and present their choices on the telephone.

Five SP choices are included in the survey. In each choice the respondent is asked to choose between two hypothetical routes. The first route has a fixed travel time every day (5 days a week), whereas the second route has the possibility that the travel time increases on some day (s). For example, Route 1 has a travel time of 30 min every day, whereas Route 2 takes 20 min 4 days per week and 40 min 1 day per week. In this case respondents are informed that if they choose Route 1, they are certain that travel time will be 30 min every day, but if they choose Route 2 they must expect that it is possible that on any one day of the week travel time could be 40 min and on the other 4 days it could be 20 min.

The choices are designed such that the travel time on the first route is always longer and certain, whereas that of the second route is shorter but uncertain. The mean travel time on the second route changes and reaches in some choices the mean of Route 1. The sequence of the choices is randomized across respondents to avoid any ordering biases.

Table 1 presents the five stated preference questions, and the mean and standard deviation of travel times and observed frequency of choices for each case (only Columns 2 and 3 were presented to the respondent in the interview). The average travel time on Route 2 ranged between 24 and 30 min, whereas the mean travel time on Route 1 was 30 min. The standard deviation ranged between about 5 min (Case 3) and about 33 min (Case 5).

Turning to the frequency of choices for each case, it is clear that (*a*) in Cases 2, 4, and 5 the majority of the respondents had chosen Route 1; (*b*) these cases have the largest standard deviations on Route 2 (>10 min); and (*c*) the mean travel time on Route 2 is either 28 or 30 min. In Case 1 both routes were almost equally chosen, the mean and standard deviation on Route 2 are 24 and about 9 min, respectively. In Case 3, where the standard deviation is the least and the mean is 24 min, Route 2 was chosen by the majority of the respondents.

Figure 1 depicts the relationship between the standard deviation of travel times on Route 2 and the frequency of each alternative being chosen. Figure 1 and Table 1 illustrate that the respondents correctly recognize the time savings and degree of variation and are willing to tolerate travel time variation to a certain limit, after which they are more likely to use the certain (although slightly longer) route.

ROUTE CHOICE MODELING

In this section, two sets of models are estimated. The first uses the pooled data set that contains all repeated choices, and the second is based on a randomly drawn observation for each respondent.

Route Choice Models Using Repeated Observations

In developing statistical models of repeated discrete choice, a central concern is the identification of the structural parameters of exogenous determinants of choice behavior, while controlling for other influences on behavior. These other influences include such effects as state dependence, initial conditions, nonstationarity, and omitted variables and unobservable variables such as taste and motivation. In the context of the data from the short-term repeated-choice sets analyzed in this paper, it is possible to argue that the values of most exogenous determinants of choice behavior remain constant over time and that the assumptions of stationarity and the lack of state dependence are reasonable. The lack of state depen-

TABLE 1 Stated Preference Choices

Case	Route	Route Description	Travel Time Mean (min/day)	Travel Time Standard Deviation (min)	Delay per Day (min)	Expected Travel Time Saving of Route 2 per week (min)	Stated Choices
1	1	30 min every day	30	0	0	—	310
	2	20 min 4 days/week 40 min 1 day/week	24	8.94	4	30	254
2	1	30 min every day	30	0	0	—	476
	2	20 min 4 days/week 60 min 1 day/week	28	17.89	8	10	88
3	1	30 min every day	30	0	0	—	159
	2	20 min 3 days/week 30 min 2 days/week	24	5.48	4	30	405
4	1	30 min every day	30	0	0	—	454
	2	20 min 3 days/week 45 min 2 days/week	30	13.69	10	0	110
5	1	30 min every day	30	0	0	—	496
	2	20 min every day 120 min 1 day/2 week	30	33.54	10	0	68

NOTE: delay/day = mean − usual travel time (most frequent) expected saving in travel time of Route 2 = travel time on Route 1/week-travel time on Route 2/week = difference in expected travel time between Routes 1 and 2/week

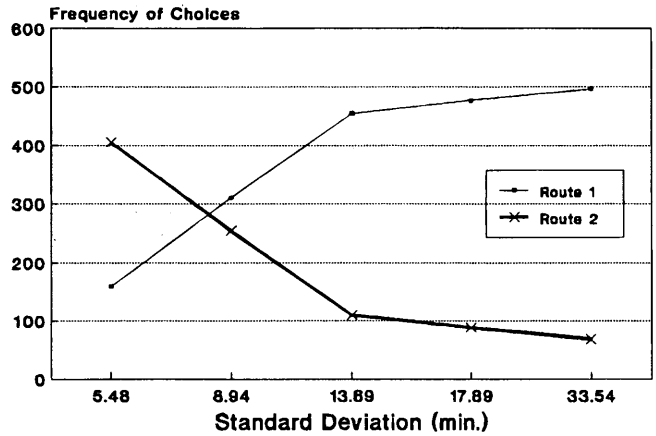


FIGURE 1 Relationship between frequency of observed choices and standard deviation.

dence in its turn implies exclusion of initial conditions (*23*). Empirical evidence for the assumption of a lack of state dependence in this context is available in the work of Uncles (*24*).

Given the simplifying assumptions of stationarity and lack of state dependence, one is left with the problem of controlling for omitted and unobserved variables whose influences are defined collectively as unobserved heterogeneity. If the possible existence of unobserved heterogeneity is not recognized and accounted for in the model's econometric structure, the model will have biased coefficients. Also, because the pooled data set includes repeated observations from each individual, the strict independence among choices and the asymptotic standard errors would be understated.

Methodological Approach

The approach taken in this paper to account for unobserved heterogeneity is to assume a parametric functional form for the pattern of the heterogeneity. The vector of observed choices or responses for individual i is defined as y_i. Each element of y_i is written as y_{it}:$t = 1, \ldots, T_i$, each of which is a repeated binary choice, expressed as the integers 0 and 1. The length of y_i is T_i, which may vary between individuals. The sample size is written as I, so $i = 1, \ldots, I$.

The assumptions of lack of state dependence, stationarity, constant exogenous variables, and constant probabilities over the repeated choices facilitate the writing of the probabilities that individual i chooses alternatives 0 or 1, P_{0it} and P_{1it}, respectively, in the standard logistic regression form

$$p_{0it} = P(y_{it} = 0|\alpha, \beta, x_{it}) = 1/[1 + \exp(x_{it}'\beta + \alpha)]$$

$$p_{1it} = P(y_{it} = 1 \mid \alpha, \beta, x_{it}) = \exp(x_{it}'\beta + \alpha)/[1 + \exp(x_{it}'\beta + \alpha)] \quad (1)$$

where

α = constant;
β = vector of parameters, and
x_{it} = vector of exogenous variables.

The influence of the unobserved variables in Equation 1 is represented by the constant term α; that is, the influence is assumed constant across individuals. The probability of observing y_i given T_i in this specification is

$$P(y_{it} \mid \alpha, \beta, T_i, x_{it}) = \prod_{t=1}^{T_i} \left[\frac{\exp(x_{it}'\beta + \alpha)}{1 + \exp(x_{it}'\beta + \alpha)} \right]^{D_{it}} \quad (2)$$

$$D_{it} = \begin{cases} 1 & \text{if } Y_{it} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Heterogeneity is introduced into the model by assuming that the probabilities p_{0it} and p_{1it} are conditional on both x_{it} and an individual specific error term, ξ_i, which represents all the other influences. Equation 1 becomes

$$p_{0it} = P(y_{it} = 0 \mid \beta, x_{it}, \xi_i) = 1/[1 + \exp(x_{it}'\beta + \alpha + \xi_i)]$$

$$\begin{aligned} p_{1it} &= P(y_{it} = 1 \mid \beta, x_{it}, \xi_i) \\ &= \exp(x_{it}'\beta + \alpha + \xi_i)/[1 + \exp(x_{it}'\beta + \alpha + \xi_i)] \end{aligned} \quad (3)$$

The ξ_i: $i = 1, \ldots, I$ are assumed to be identically distributed with density function $f(\xi)$ independent of the x_i, so that Equation 2 becomes

$$\begin{aligned} &P[y_{it} \mid \beta, T_i, x_{it}, f(\xi_i)] \\ &\quad = \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} \left[\frac{\exp(x_{it}'\beta + \alpha + \xi_i)}{1 + \exp(x_{it}'\beta + \alpha + \xi_i)} \right]^{D_{it}} f(\xi_i)\, d(\xi_i) \end{aligned} \quad (4)$$

This yields a marginal likelihood function. The unknown variables ξ are integrated out. Equation 4 is based on the assumption that ξ has a continuous distribution function. The distribution of ξ_i is called a mixing distribution. The log likelihood function is

$$L = \sum_{i=1}^{I} \ln \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} \left[\frac{\exp(x_{it}'\beta + \alpha + \xi_i)}{1 + \exp(x_{it}'\beta + \alpha + \xi_i)} \right]^{D_{it}} f(\xi_i)\, d(\xi_i) \quad (5)$$

A parametric form and $\xi_i \sim N(0, \sigma^2)$ are assumed. The integral is evaluated using Gaussian quadratures. General MLE packages, such as the one provided with GAUSS statistical software (*25*) can be used for this problem. The Broyden, Fletcher, Goldfarb, and Shanno (BFGS) optimization method is used (*26*). The BFGS method is similar to the Newton method in that it uses both first and second derivative information . However, in BFGS the Hessian is approximated, reducing considerably the computational requirements, and although it takes more iterations than Newton it converges in less overall time.

Estimation Results

A binary logit model is developed using the methodology presented above. The model is developed to estimate the commuters' choice between Route 1 (longer with reliable travel time) and Route 2 (shorter with uncertain travel time). The overall observations are used to estimate the models, which give a total of 2,820 observations (i.e., 564 respondents, each making five choices). The data used for estimating the model came from the two CATI surveys (e.g., perception of shorter distance) and Table 1 (e.g., standard deviation of the travel time).

The model is presented in the first part of Table 2 and shows that commuters' perceptions and attitudes have important effects on their choice; that is, if respondents perceive shorter travel distances as being extremely or very important, then they are likely to choose Route 2 in trying to minimize their travel time.

The standard deviation of the travel time on Route 2 has a negative coefficient, indicating that the more the variation in travel time on Route 2, the less likely this route is to be chosen. This result shows that commuters realize travel time and its variability on alternative routes and try to minimize them. Also, the larger the difference in the expected travel time between Routes 1 and 2, the more likely the respondent chooses Route 2, indicating that commuters realize the savings in travel time and choose the route that achieves a minimum travel time. These two variables show clearly that commuters try to minimize their travel time but only if travel time variation is acceptable. If travel time varies significantly on a particular route then commuters will choose the longer certain route.

Receiving traffic information is a very significant variable in this model. Information is more likely to affect the degree of uncertainty and hence influences the commuter's route choice. Acquiring traffic information could be treated as either an endogenous or an exogenous variable. Commuters receive information because of personal reasons (e.g., to reduce their degree of uncertainty) or because of their commute characteristics (e.g., long commute trip). Therefore receiving pretrip traffic information is most likely to be an endogenous variable; thus the variable was instrumented using a binary logit model estimated in Abdel-Aty et al. (*8*)—the data used in the instrument come from the first CATI survey, i.e., commute distance, gender, traffic conditions on the regular commute route, and perception of the uncertainty of travel time. The variable has a significant positive coefficient that indicates that commuters who listen to pretrip information are more likely to choose the uncertain route, possibly because they are confident that they can know if there is a delay on a particular day and avoid this route. The significance of the information variable validates the SP choice sets used in this study because people do acquire information in the real world to reduce their uncertainty.

Gender also had a significant effect on route choice. Males are found to be more likely to choose Route 2. This indicates that males are more risk prone and are ready to choose uncertain routes in trying to minimize their travel time.

Finally, the significance of σ illustrates that the unobserved influences affecting a specific individual's choice are correlated from one of his or her selections to the next.

A second model that describes the route choice with normal mixing distribution was estimated. This model is similar to the model presented in Table 2, but receiving pretrip traffic information is substituted by receiving en route information. This model is similar to a large extent to that shown in Table 2. However the overall fit of the first model (including the effect of pretrip information) is slightly better (log likelihood of = -1133.804 versus -1135.078). Also the *t*-statistics of receiving pretrip information are significant at the 95 percent level of significance, whereas receiving en route information was significant only at the 90 percent confidence level. As found in a previous study (*10*), commuters might value and use pretrip information more than en route information because it notifies them of the status of their routes in advance, which enables them to change route or departure time, or both. In the context of this study, traffic information, particularly pretrip, will help reduce the degree of uncertainty when commuters encounter a variation in travel time on their routes.

The inclusion of both types of information in a model was attempted, but this caused problems in the model estimation because of multicollinearity. A possible extension of this work is to estimate a similar model that considers whether the respondent receives pretrip or en route traffic information.

TABLE 2 Estimates Describing Route Choice with Normal Mixing Distribution and Gaussian Quadrature Estimation Using Pooled Data and Randomly Drawn Observation, Including Effect of Pretrip Traffic Information

	Normal mixing distribution		Pooled repeated measurement			Randomly drawn observation	
	Coef.	t-stat.	Coef.	t-stat.	correc. t-stat.	Coef.	t-stat.
Constant	-2.394	-5.89	-1.655	-6.45	-2.88	-1.199	-2.08
X_1 Attitude toward shorter distance dummy (1 if extremely or very important, 0 otherwise)	0.550	3.26	0.391	4.22	1.89	0.212	1.06
X_2 Standard deviation of travel time on Route 2 (min.)	-0.067	-6.32	-0.052	-5.67	-2.54	-0.065	-2.99
X_3 Difference in expected travel time between Route 1 & 2 /week	0.067	10.31	0.048	8.99	4.02	0.031	2.61
X_4 Receive pre-trip information - instrumented	0.416	2.54	0.294	3.84	1.72	0.276	1.71
X_5 Male dummy variable	0.548	3.25	0.372	4.08	1.82	0.432	2.18
σ Standard Deviation of ξ_t	1.462	13.51					
Summary Statistics							
Log Likelihood at zero	-1954.675		-1954.65			-390.93	
Log Likelihood at market share	-1784.392		-1784.39			-351.66	
Log Likelihood at convergence	-1133.804		-1477.12			-307.86	
Likelihood ratio index	0.419		0.244			0.213	
Number of observations = 2820 (564 respondents)	2820		2820			564	

Models Using Pooled and Randomly Drawn Data

The same model presented earlier is estimated using the same model specifications such as (*a*) pooled repeated measurement data and correcting the *t*-statistic by dividing it by the square root of the number of observations for each respondent [this heuristic method was used in Louviere and Woodworth (*15*) and Mannering (*16*)] and (*b*) one observation randomly drawn from each respondent. The models are also presented in Table 2 to facilitate comparisons among the three models.

A comparison of the three models presented in Table 2 indicates that the results of the pooled data model, after correcting the *t*-statistics value, and the randomly drawn observation model are to a great extent close. The *t*-statistics of the model estimated using the mixing distribution and that of the uncorrected pooled data model are comparable to a large extent (mixing distribution produced the largest *t*-statistics for route-specific attributes, whereas pooled data tended to give the highest *t*-statistics for individual attributes). It is apparent that the corrected pooled data model produces a conservative estimate of the *t*-statistic values, which might have overcorrected these values. On the other hand, the model with the randomly drawn observation lacks the benefits of using additional information in the much larger pooled data set. Figure 2 compares the coefficient estimates of the three models and shows that the coefficients are similar for some of the variables, that is, standard deviation and difference in expected travel time (route-specific attributes). On the other hand, the coefficients are different for other variables.

These comparisons illustrate the need for a method to account for heterogeneity. The use of normal mixing distribution is used in this paper. However, extending this effort to include different mixing distributions and nonparametric distributions remains as a future task.

CONCLUSIONS

The primary conclusion of this research is that a specific measure of travel time reliability, variability of travel time, has an important impact on the route choice behavior. It is clear that the choice sets could not be posed to the respondents in formal statistical terms, such as mean and standard deviation. However, the results showed that using repeated hypothetical choice sets while varying travel time on one of the routes is a viable method. This method achieved travel time dimensions that are easily convertible to the more formal statistical measures, which are desirable from the modeler's standpoint. More impressively, the respondents understood the degree of variation and responded rationally.

The results of the models estimated using the stated preference route choices yielded important insights into the commuters' route choice in general and the tradeoffs involved in the choice between a route that is longer but has reliable travel time versus another route that is shorter but has an uncertain travel time. The models that are estimated either by using single or repeated observations for each respondent show that both expected travel time and variation in travel time influence route choice; commuters' attitudes toward several commute characteristics (e.g., distance and traffic safety) influence route choice; and, among the socioeconomic factors, gender has a significant effect on route choice.

Receiving traffic information is found to have a significant effect in the models. Information might be used by the commuters to reduce the degree of travel time uncertainty and enables them to choose routes adaptively.

The data also suggest that the impact of travel time variability varies substantially across individuals, ranging from those who will choose routes that are significantly longer to avoid the possibility of delay to those who are essentially expected value decision makers with respect to commute alternatives. A possible extension to this work is to introduce the idea of risk aversion and being risk prone in the route choice models and measure the bound of risk aversion.

The error components account for unobserved heterogeneity and correct for potential bias that would otherwise arise from the use of repeated measurement data. The repeated measurement issue is addressed in this study with individual-specific random error components in a series of binary logit models with normal mixing distribution. The significance of the standard deviation of the error components shows clearly the need for some formal statistical

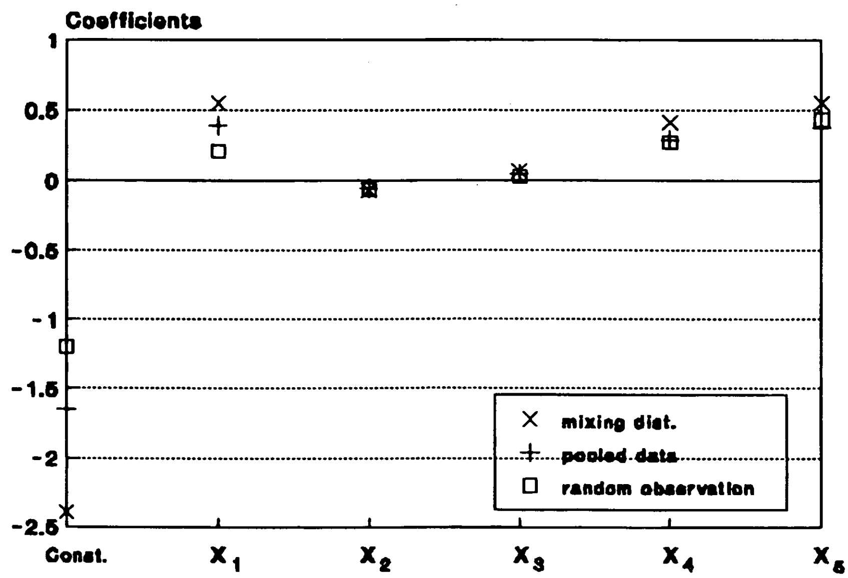


FIGURE 2 Comparison coefficient estimates of three models.

correction to account for heterogeneity. A methodological future direction is to attempt models with the same specifications using the nonparametric approach and compare them with the models presented in this paper to reach conclusions about the best way to account for heterogeneity in route choice models.

ACKNOWLEDGMENTS

The authors wish to acknowledge the California Department of Transportation and the Partners for Advanced Transit and Highways for funding this research.

REFERENCES

1. Wachs, M. Relationships Between Drivers' Attitudes Toward Alternate Routes and Driver and Route Characteristics. In *Highway Research Record 197,* TRB, National Research Council, Washington, D.C., 1967, pp. 70–87.
2. Ben-Akiva, M., M. Bergman, A. Daly, and R. Ramaswamy. Modelling Inter Urban Route Choice Behavior. *Proc., 9th International Symposium on Transportation and Traffic Theory,* Delft, The Netherlands, 1984.
3. Huchingson, R., R. McNees, and C. Dudek. Survey of Motorist Route-Selection Criteria. In *Transportation Research Record 643,* TRB, National Research Council, Washington, D.C., 1977, pp. 45–48.
4. Jackson, W., and J. Jucker. An Empirical Study of Travel Time Variability and Travel Choice Behavior. In *Transportation Science,* Vol. 16, No. 4, 1981, pp. 460–475.
5. Black, I., and J. Towriss. Quantifying the Value of Uncertainty in Travel Time. *Proc., 18th PTRC,* Seminar H, 1991.
6. Abdel-Aty, M., R. Kitamura, P. Jovanis, and K. Vaughn. *Investigation of Criteria Influencing Route Choice: Initial Analysis Using Revealed and Stated Preference Data.* Research Report UCD-ITS-RR-94-12. Institute of Transportation Studies, University of California, Davis, 1994.
7. Wardrop, J. Some Theoretical Aspects of Road Traffic Research. *Proc., Institution of Civil Engineers,* Part II, Vol. 1, 1952, pp. 325–378.
8. Abdel-Aty, M., K. Vaughn, R. Kitamura, and P. Jovanis. *Survey of Route Choice Behavior: Empirical Results from Southern California and Their Implications for Advanced Information Systems.* Research Report UCD-ITS-RR-93-12. Institute of Transportation Studies, University of California, Davis, 1993.
9. Abdel-Aty, M., K. Vaughn, R., Kitamura, P. Jovanis, and F. Mannering. Models of Commuters' Information Use and Route Choice: Initial Results Based on a Southern California Commuter Route Choice Survey. In *Transportation Research Record 1453,* TRB, National Research Council, Washington, D.C., 1994.
10. Abdel-Aty, M., K. Vaughn, P. Jovanis, R. Kitamura, and F. Mannering. Impact of Traffic Information on Commuters' Behavior: Empirical Results from Southern California and Their Implications for ATIS. *Proc., 4th Annual Meeting of IVHS America,* Atlanta, Ga., 1994.
11. Bunch, D., M. Bradley, T. Golob, R. Kitamura, and G. Occhiuzzo. Demand for Clean-Fuel Vehicles in California: A Discrete-Choice Stated Preference Pilot Project. *Transportation Research A,* Vol. 27A, No. 3, 1993, pp. 237–253.
12. Mannering, F., and C. Winston. A Dynamic Empirical Analysis of Household Vehicle Ownership and Utilization. *Rand Journal of Economics,* Vol. 16, No. 2, 1985, pp. 215–236.
13. Mannering, F. A Note on Endogenous Variables in Household Vehicle Utilization Equations. *Transportation Research B,* Vol. 20B, No. 1, 1986, pp. 1–6.
14. Hocherman, I., J. Prashker, and M. Ben-Akiva. Estimation and Use of Dynamic Transaction Models of Automobile Ownership. In *Transportation Research Record 944,* TRB, National Research Council, Washington, D.C., 1983, pp. 134–141.
15. Louviere, J., and G. Woodworth. Design and Analysis of Simulated Consumer Choice or Allocation Experiments: An Approach Based on Aggregate Data. *Journal of Marketing Research,* Vol. 20, 1983, pp. 350–367.
16. Mannering, F. Analysis of the Impact of Interest Rates on Automobile Demand. In *Transportation Research Record 1116,* TRB, National Research Council, Washington, D.C., 1987, pp. 10–14.
17. Heckman, J., and R. Willis. A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women. *Journal of Political Economy,* Vol. 85, 1977, pp. 27–58.
18. Uncles, M. A Beta-Logistic Model of Mode Choice: Goodness of Fit and Intertemporal Dependence. *Transportation Research B,* Vol. 21B, No. 3, 1987, pp. 195–205.
19. Kitamura, R., and D. Bunch. Heterogeneity and State Dependence in Household Car Ownership: A Panel Analysis Using Ordered-Response Probit Models with Error Components. *Transportation and Traffic Theory.* Elsevier Science Publishing Co., Amsterdam, The Netherlands, 1990, pp. 477–496.
20. Morikawa, T. Correcting State Dependence and Serial Correlation in the RP/SP Combined Estimation Method. *Transportation* (in press).
21. Abdel-Aty, M., R. Kitamura, P. Jovanis, and K. Vaughn. *Understanding Commuters' Attitudes, Uncertainties, and Decision-Making and Their Implications for Route Choice.* Research Report UCD-ITS-RR-94-5. Institute of Transportation Studies, University of California, Davis, 1994.
22. Khattak, A., A. Kanafani, and E. Colletter. Stated and Reported Route Diversion Behavior: Implications on the Benefits of ATIS. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1994.
23. Dunn, R., S. Reader, and N. Wrigley. A Nonparametric Approach to the Incorporation of Heterogeneity into Repeated Polytomous Choice Models of Urban Shopping Behavior. *Transportation Research A,* Vol. 21A, No. 4/5, 1987, pp. 327–343.
24. Uncles, M. *Models of Consumer Shopping Behavior in Urban Areas.* Ph.D. dissertation. Department of Geography, University of Bristol, United Kingdom, 1985.
25. *GAUSS—Version 3.0 User's Manual.* Aptech Systems, Inc., Kent, Wash., 1992.
26. Luenberger, D. *Linear and Nonlinear Programming.* Addison-Wesley, Cambridge, Mass., 1984.

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

Evolution of Network Flows Under Real-Time Information: Day-to-Day Dynamic Simulation Assignment Framework

TA-YIN HU AND HANI S. MAHMASSANI

A day-to-day dynamic framework, in which the DYNASMART simulation assignment model was applied to evaluate the performance of traffic networks, was developed to study network dynamics under different information systems. Two levels of tripmaker decision-making processes are identified: (*a*) day-to-day dynamics and (*b*) real-time dynamics. Day-to-day dynamics consider the choices of departure time and route according to indifference bands of tolerable "schedule delay" defined as the difference between the user's actual and preferred arrival times. Real-time dynamics consider en route switching decisions. Numerical experiments were conducted to investigate the day-to-day evolution of network flows under real-time information and assess the effectiveness of such information in a proper dynamic perspective.

Advanced traveler information systems (ATIS) and advanced traffic management systems provide a variety of capabilities to alleviate traffic congestion in urban networks by strengthening the connection between traffic control and available information (*1*). The evaluation of such information-based systems has been concerned primarily with the potential of this information to redistribute flows spatially over the network during the peak period on a given day (*2–4*). However, real-time information can also induce changes in time of departure, leading to temporal redistribution of the flows. Such effects tend to take place over several days. In other words, although the ability of real-time information to affect en route switching is well recognized, its potential effect on the day-to-day decisions of departure time and route remains to be investigated systematically. A key question is how tripmakers make decisions on the basis of experienced or received information, or both. Although the importance of learning processes in such systems has been recognized (*5–8*), consideration of such processes needs to be incorporated into the effectiveness of analysis and evaluation of information systems.

This paper describes a day-to-day dynamic simulation assignment framework to study the interaction among individual decisions, traffic control strategies, and network flow patterns under real-time information systems. The framework integrates two previous lines of investigation, namely (*a*) day-to-day forecasting methods for commuter systems, previously considered only in a corridor context and without en route real-time information (*9*), and (*b*) time-dependent assignment-simulation modeling for networks with general topology under real-time ATIS in the form of DYNASMART (*10*). The resulting methodology is applicable to general networks with detailed representation of traffic processes, including traffic control actions, and provides a tool for forecasting the day-to-day evolution of the system under various information policies, network supply actions or control strategies. System users are represented individually in the model, and their daily decisions of route and departure time (and possibly mode) provide the principal mechanism governing day-to-day evolution. Similarly, user decisions in response to information, both en route and pretrip, are also represented individually. As such, this framework provides an illustration of an operational dynamic demand forecasting tool on the basis of microsimulation of individual tripmaking decisions (although traffic interactions are modeled using macroscopic relations).

The next section presents the day-to-day dynamic simulation assignment model framework and DYNASMART. The algorithmic procedure and experimental design and numerical results are discussed, and concluding comments follow.

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

DAY-TO-DAY DYNAMIC SIMULATION ASSIGNMENT FRAMEWORK

Given the focus on peak-period network flows, the framework considers primarily the variation in route and departure time in the context of commuting trips to work, for which tripmaker behavior rules for day-to-day decisions have been calibrated in previous work. Extensions to consider noncommuters and nonwork trips are conceptually straightforward in terms of overall framework, although appropriate individual decision rules for these situations remain to be developed.

Consider a network $G(N,A)$ consisting of a set of nodes N connected by the set of directed arcs A. Suppose user i intends to go from origin r to destination s and arrives at his or her preferred arrival time (PAT_i), $\forall i \in D$, the set of all drivers. PAT_i reflects inherent preferences and risk attitudes of commuter i, as well as the characteristics of the work place. In this paper, PAT_i is assumed fixed for a given tripmaker; however, it could be generalized and varied through appropriate behavior models to reflect flexible work schedules. The selected departure time $j_{i,t+1}$ and route $k_{i,t+1}$ for driver i on Day $t + 1$ are the outcomes of its decision-making process, described as

$$k_{i,t+1} = f_r(X_i, Z_{i,t}, Y_{i,t} \mid \theta_r) \quad (1)$$

$$j_{i,t+1} = f_{dt}(X_i, Z_{i,t}, Y_{i,t} \mid \theta_{dt}) \quad (2)$$

where

$k_{i,t+1}$ = selected route for driver i on day $t + 1$
$j_{i,t+1}$ = selected departure time for driver i on day $t + 1$,
$f_r(.)$ = route choice decision-making process function,

$f_{dt}(.)$ = decision-making process function for departure time,
X_i = vector of driver characteristics,
$Z_{i,t}$ = vector of endogenous information characteristics for driver i up to day t,
$Y_{i,t}$ = vector of exogenous information characteristics for driver i up to day t, and
θ_r, θ_{dt} = parameter vectors to be calibrated.

The choices of departure time and route of tripmaker i on day $t + 1$ depend on individual tripmaker characteristics, endogenous information from personal experience, and exogenous information from traffic control centers.

The aggregated departure time decisions of all users determine a three-dimensional time-dependent origin-destination (OD) matrix; the route choices determine the spatial distribution of flows over the peak period. The time-dependent OD matrix and the initial route assignment form the major input for DYNASMART, in which individual en route decisions are represented. Within the simulation period, tripmaker i equipped to receive in-vehicle information makes en route decisions according to his or her own behavioral characteristics and information received about prevailing traffic conditions in the network. Let $\delta_{i,n,1}$ denote a binary indicator that is 1 when driver i switches to a new path 1 at node n from the current path and 0 otherwise; $\delta_{i,n,1}$ can be determined by the user's characteristics, knowledge of the paths at node n, $Z_{i,t}(n)$ and new information about path from node n to his or her destination and is expressed as

$$\delta_{i,n,1} = f_s[X_i, Z_{i,t}(n), Y_{i,t}(n)|\theta_s] \tag{3}$$

where

$Z_{i,t}(n)$ = endogenous knowledge of driver i at node n on day t,
$Y_{i,t}(n)$ = exogenous information for driver i at node n on day t,
$f_s(.)$ = en route path-switching function, and
θ_s = parameter vector, to be calibrated.

As a consequence, the flow pattern in the network on day t, F_t, resulting from a time-dependent OD, initial path selections for day t, and en route path-switching decisions can be expressed as

$$F_t = \text{flow}_t\,(k_{i,t}, j_{i,t}, \delta_{i,n,1}, \forall_i \in D \text{ and } n \in N) \tag{4}$$

Endogenous and exogenous information $Z_{i,t}$ and $Y_{i,t}$ can be written as

$$Z_{i,t} = f_n\,(X_i, F_t) \tag{5}$$

$$Y_{i,t} = f_x\,(C_{r,t+1}, C_{s,t+1}, F_t) \tag{6}$$

where

$f_n(.)$ = endogenous information acquisition function,
$f_x(.)$ = exogenous information provision function, and
$C_{r,t+1}$ and $C_{s,t+1}$ = route control and signal control on day $t + 1$.

$Z_{i,t}$ and $Y_{i,t}$ are then used in Equations 1 and 2 to determine the departure time and initial route on day $t + 1$. Note that the control actions $C_{r,t+1}$ and $C_{s,t+1}$ on day $(t + 1)$ are generated with knowledge by the controller on traffic conditions associated with flow pattern F_t on day t. The whole process takes place in a recursive form. Naturally, the complexity of the interactions depicted earlier precludes analytic solution of system performance descriptors.

Information Systems

Information types and flow for different types of user classes within this framework are defined to illustrate the possible interaction between them. Vehicles (i.e., users) are differentiated into equipped and nonequipped classes on the basis of their ability to communicate in real time with a central controller. Nonequipped vehicles do not receive real-time information and are assumed to follow the initial path selected before their departure. Although users in this class do not make decisions on the basis of in-vehicle real-time information, they can still respond to exogenous information supplied through variable message signs. Equipped vehicles communicate with the controller, and their drivers can therefore make decisions on path selection en route.

Information strategies can be categorized into two general types: descriptive and normative. Descriptive information, currently the most common type used or proposed, provides tripmakers with current traffic conditions through different communication channels. Tripmakers can use this information to make their own travel decisions, independently of other users' decisions. On the other hand, normative information delivers instructions aimed at achieving some systemwide objectives. Information can be experienced by travelers or collected by control centers by probes, detectors, or equipped vehicles, or all of these.

A fundamental problem is what actions drivers might make on the basis of different information types. In the day-to-day dynamics context, studies that have explicitly dealt with this aspect have relied on a convenient Markovian assumption, whereby the anticipated travel time on a given day is assumed to be equal to its actual value on the preceding day only (*11–13*.) Horowitz (*14*) proposed to model the predicted trip time on day t as a weighted sum of all previous days' trip times. Empirical investigation of this issue is limited. Mahmassani and Chang (*15*) and Tong et al. (*16*) have calibrated departure time adjustment rules in which the predicted travel time is based on the driver's own previous experience as well as exogenous information. The calibrated models show that the influence of travel time on the immediately preceding day, $TR_{i,t-1}$, is much greater than that of $TR_{i,t-2}$ (experienced 2 days previously). Functional forms of how information is processed can thus be generalized as the weighted sum of all previous days' information and different assumptions on tripmaker behavior can be reflected by varying the relative weights.

Day-to-Day Dynamic Choice Behavior

The behavior component within the day-to-day framework addresses the selection of route and departure time in accordance with individual attributes and received information. The theoretical underpinnings of the model are grounded in Simon's well-known notion of bounded rationality, applied to commuter day-to-day decisions of departure time and route in work by Mahmassani and Chang (*17,18*). Essentially, the model is founded on the simple notion that if tripmakers are not satisfied with their previous selections, they will seek to select a new route or adjust their departure time, or both. Satisfaction is implemented on the basis of "indifference bands" of tolerable schedule delay (relative to one's preferred arrival time).

This decision process consists of two levels, as indicated in Figure 1. The first level is concerned with acceptability of the conse-

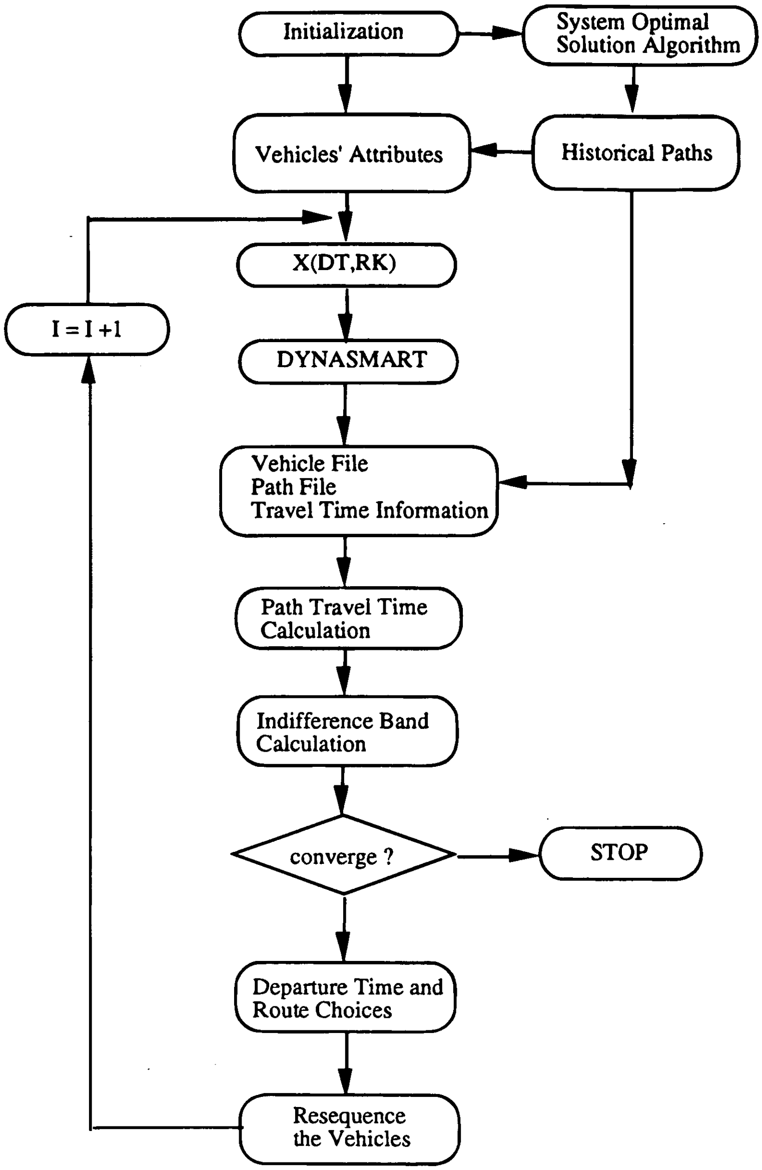


FIGURE 1 Day-to-day dynamic analysis procedure.

quences of the latest choices, vis-à-vis the indifference bands; the second level is used to select an alternative conditional on the decision to switch taken at the first level. Previous studies have shown that arrival time is of major concern to commuters and have suggested that an indifference band of tolerable "schedule delay," defined as the difference between the actual arrival time (AT) and the preferred arrival time (PAT) for a given tripmaker, is the primary mechanism governing the day-to-day responses of commuters to congestion. In their daily commute, tripmakers are assumed to maintain the choice as long as they can tolerate the associated earliness or lateness relative to PAT.

$$\gamma_{i,t} = \begin{cases} 0 & \text{if } 0 \leq \mathrm{ESD}_{it} \leq \mathrm{EBD}_{it} \text{ or } -\mathrm{LBD}_{it} \leq \mathrm{LSD}_{it} \leq 0 \\ 1 & \text{otherwise} \end{cases} \tag{7}$$

$$\lambda_{i,t} = \begin{cases} 0 & \text{if } 0 \leq \mathrm{ESD}_{it} \leq \mathrm{EBR}_{it} \text{ or } -\mathrm{LBR}_{it} \leq \mathrm{LSD}_{it} \leq 0 \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

where

$\gamma_{i,t}$ = departure-time switching binary indicator, equal to 1 if switch, 0 otherwise;

$\lambda_{i,t}$ = route choice indicator, equal to 1 if switch, 0 otherwise;

$\mathrm{ESD}_{i,t}$ = early schedule delay, equal to $\mathrm{Max}(\mathrm{PAT}_{i,t-1} - \mathrm{AT}_{i,t-1}, 0)$; and

$\mathrm{LSD}_{i,t}$ = late schedule delay, equal to $\mathrm{Max}(\mathrm{AT}_{i,t-1} - \mathrm{PAT}_{i,t-1}, 0)$.

There are four possible combinations of departure time and route-choice switching decisions, corresponding to the combinations of values for the pair ($\gamma_{i,t}$, $\lambda_{i,t}$). Note that EBD and LBD are the respective departure time indifference bands of tolerable schedule delay corresponding to early and late arrivals for day *t*, and EBR and LBR denote the early and late indifference bands governing route switching. Because the indifference bands are latent terms, internal to each individual, and therefore can be neither observed nor measured directly, the indifference bands are treated as random variables, distributed over days and across commuters with systematically varying mean values (*9*).

The second level in Figure 2 is the selection of an alternative, which could be a new departure time, a new route, or both, conditional on the decision to switch. Several rules, based on different behavioral assumptions, can be applied in the individual selection process. In this study, alternative selection is based on a simple utility maximization process. Two particular models, proposed by Small (*19*) and Hendrickson and Plank (*20*), respectively, are used in the numerical experiments.

DYNASMART Simulation Assignment Model

DYNASMART is a descriptive analysis tool for the evaluation of information supply strategies, traffic control measures, and route assignment rules at the network level (*2,4,21,22*) The model is designed around a flexible structure that provides sensitivity to a wide range of traffic control measures for both intersections and freeways, capability to model traffic disruptions as a result of incidents and other occurrences, and representation of several user classes corresponding to different vehicle performance characteristics (e.g., cars verses trucks), access to physical facilities (e.g., high occupancy vehicle lanes), different information availability status, and different behavioral rules.

The framework of DYNASMART is shown in Figure 2. The approach integrates traffic flow models, path processing methodologies, behavioral rules, and information supply strategies into a single simulation assignment framework. The input data include a time-dependent OD matrix (or a schedule of individual departures) and network data. Given the network representation, the simulation component will take a time-dependent loading pattern and process the movement of vehicles on links and the transfers between links according to specified control parameters. These transfers, which are determined by path processing and path selection rules, require instructions that direct vehicles approaching the downstream node of a link to the desired outgoing link. The user behavior component is the source of these instructions.

DYNASMART uses established macroscopic traffic flow models and relationships to model the flow of vehicles through a network. Whereas macroscopic simulation models do not keep track of individual vehicles, DYNASMART moves vehicles individually or in packets, thereby keeping a record of the locations and itineraries of the individual particles. This level of representation also has been referred to as "mesoscopic." Multiple user classes of different vehicle performance characteristics are modeled as packets, consisting of one or more passenger car units; for instance, a bus is represented by a packet with two (or other user-specified values) passenger car units. The traffic simulation consists of two principal modules: link movement and node transfer, as described previously (*4,22*).

One of the principal features of DYNASMART that allows it to interface with activity-based behavioral models is its explicit representation of individual tripmaking decisions, particularly for path selection decisions, both at the trip origin and en route. Behavioral rules governing route choice decisions are incorporated, including the special case in which drivers are assumed (required) to follow specific route guidance instructions. Experimental evidence presented by Mahmassani and Stephan (*23*) suggested that commuter route choice behavior exhibits a boundedly rational character. This means that drivers look for gains only outside a threshold, within which the results are satisfying and sufficing for them. This can be translated to the following route switching model (*2*):

$$\delta_{i,n,1} = \begin{cases} 1 & \text{if } \mathrm{TTC}_i(n) - \mathrm{TTB}_i(n) > \max[\eta_i \cdot \mathrm{TTC}_i(n), \tau_i] \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

where

$\delta_{i,n,1}$ = binary indicator variable of 1 when user i switches from current path to best alternate 1 and 0 if current path is maintained;
$\mathrm{TTC}_i(n)$, $\mathrm{TTB}_i(n)$ = trip times along current path and along best path from node k to destination on current path, respectively;
η_i = relative indifference threshold; and
τ_i = absolute minimum travel time improvement needed for a switch.

The threshold level may reflect perceptual factors, preferential indifference, or persistence and aversion to switching. The quantity η_i governs users' responses to the supplied information and their propensity to switch. The minimum improvement τ_i is currently taken to be identical across users. Efforts are under way to calibrate these parameters from the results of laboratory experiments.

ALGORITHMIC STEPS OF DAY-TO-DAY DYNAMIC MODEL

Day-to-Day Dynamic Algorithm

The conceptual framework of day-to-day dynamics was discussed in the previous section. The procedure, as shown in Figure 3, can be summarized as follows:

- Step 0: *Initialization.* Generate vehicles' attributes and historical paths. Obtain a set of paths from origin r to destination s for each discrete departure time interval, denoted as $P^t_{r,s}$. Also, each driver i will be assigned a set of simulation attributes, S_i, and a set of behavior attributes, B_i. Set iteration counter $I = 1$.
- Step 1: *Network loading.* For each driver i, assign a path p from r to s, $p_i \in P^t_{r,s}$, an initial departure time, and a loading location, i.e., a generation link. For each day, the number of vehicles for each time interval *DT* and for each path *RK*, denoted *X(DT,RK)*, is generated to form a three-dimensional matrix over both space and time.
- Step 2: *Traffic simulation.* Simulate network performance during peak period under given demand pattern using DYNASMART. Obtain an updated vehicle file, additional path files (if any diversion rule is applied), and time-dependent travel time information for links and movements.
- Step 3: *Information update.* Update the historical path information in terms of travel time, add new paths, or delete obsolete paths from the historical path file.
- Step 4: *Day-to-day behavior: indifference bands.* Calculate the departure time and route choice indifference bands for the driver i according to B_i. Determine values of the switching indexes $\gamma_{i,t}$ and $\lambda_{i,t}$ $\forall_i$ for all given t.
- Step 5: *Convergence test.* If convergence criterion is satisfied (the current flow pattern is stable), stop. Otherwise, continue.
- Step 6: *Selection of departure time and route.* If the outcomes of $\gamma_{i,t}$ and $\lambda_{i,t}$ are (1,0), (0,1), or (1,1), update departure time and route choice according to B_i.

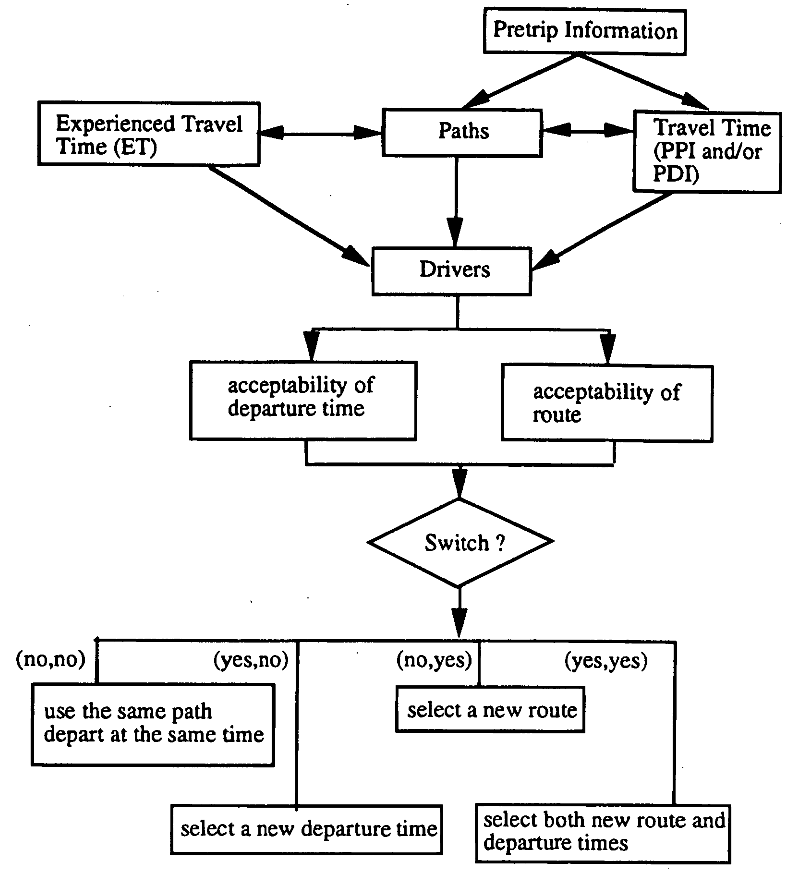


FIGURE 2 Pretrip decision-making process.

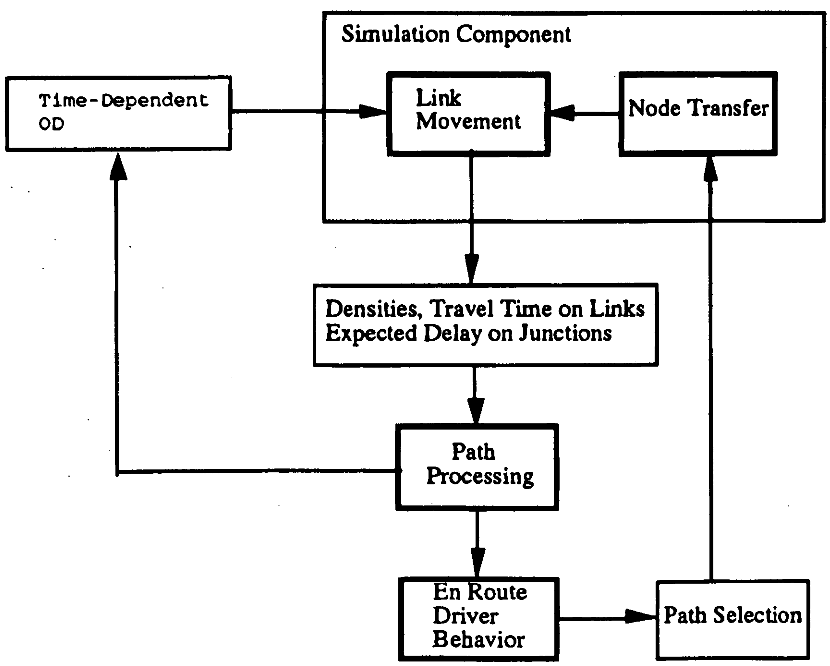


FIGURE 3 Framework of simulation assignment model with real-time information.

• Step 7: *Resequence and feedback.* Resequence vehicles according to their departure time. Obtain a time-dependent OD matrix. Set $I = I + 1$ and go to Step 1.

To overcome the problem of an arbitrary starting point, the initial set of paths is system optimal in terms of minimizing total trip time and is obtained using an algorithm recently developed by Mahmassani and Peeta (*24*), for the given time-dependent demand pattern. The vehicle file and the historical path file are used and updated through the whole simulation period. Currently, for each discrete departure time for each OD pair, up to 10 paths are stored and dynamically updated in terms of travel time for each path. All the path travel times are updated by combining recent travel time information with "historical" information, as follows:

$$\mathrm{PT}(T,r,s,j,k) = \sum_{t=1}^{T-1} w(t) \cdot \mathrm{PT}(t,r,s,j,k) \tag{10}$$

where $\mathrm{PT}(t,r,s,j,k)$ is the path travel time for day t on route j at departure time k, and $\Sigma w(t)$ is 1 and can be used to express the relative importance of historical travel time. Currently, the particular values used for $w(T-1) = 1$, and $w(T-2) = 0$.

Convergence Concept: BRUE

The boundedly rational user equilibrium (BRUE) concept proposed by Mahmassani and Chang (*15*) was applied in this study as the convergence concept. A BRUE arises in a system when no user is compelled to change his or her current selection, which he or she considers satisfactory in a boundedly rational sense. In this context, this corresponded to all users' arrival times falling within their respective departure time and route indifference bands. The particular operational definition adopted in the simulation experiments required at least a certain fraction, say 90 percent, of tripmakers to be satisfied with their current decisions.

EXPERIMENTAL DESIGN

Numerical experiments were performed to illustrate the day-to-day dynamic framework and to explore the evolution of a traffic system in response to different information supply strategies under different assumptions. The primary concerns of these experiments were (*a*) the dynamic evolution of the system, (*b*) congestion formation and dissipation, and (*c*) effectiveness of real-time information.

Traffic Characteristics

The network structure indicated in Figure 4 was used in these experiments. It consists of 50 nodes and 168 links and includes 10 demand zones with 32 origins and 10 destinations. Each link is 0.25 mi (0.4 km) long. The freeway links have a free-flow speed of 55 mph and all other links have a 30-mph (48-kph) mean free speed. The maximum bumper-to-bumper and jam densities are assumed to be 260 and 160 vehicles per mile (approximately 152 and 100 vehicles per kilometer), respectively, for all links of the network. With regard to intersection signal control, 26 nodes have pretimed signalization, 8 have actuated signal control, and the rest have no signal control. The pretimed signals have a 60-sec cycle length with two phases, each with 26 sec of green time and 4 sec of amber time. The actuated signals have 10 sec of minimum green time and 26 sec of maximum green time for each phase. In these experiments, signal control parameters are assumed fixed. The OD matrix D has a total number of 9,634 vehicles for a period of 25 min (8:05 to 8:30 a.m.) in the first day from 32 origins to 10 destinations. Time of departure is discretized into 40 intervals of 1 min between 8:00 and 8:40 a.m.

Models of Departure Time and Route Switching

The particular models applied in this dynamic analysis were calibrated by Jou et al. (*25*) using survey data from the Dallas, Texas, area. Tripmakers in that survey had an average travel time of 23.5 min. Because the average trip time in the simulation experiments is much smaller, the indifference bands given by the models are adjusted by the average travel time in the simulation experiments. The indifference band for departure time selection is as follows:

$$\begin{aligned}
\mathrm{IBDT}it = {} & \beta_1 && \text{[initial bands]} \\
& + \beta_2\mathrm{AGE}_i + \beta_3\mathrm{GENDER}_i && \text{[socioeconomic component]} \\
& + \beta_4\mathrm{NFAIL}_{it}^{\beta_5} && \text{[dynamic component]} \\
& + \beta_6\delta_{i,t}\left(\frac{\Delta\mathrm{TR}_{it}}{\Delta\mathrm{DT}_{it}}\right) && \text{[myopic component]} \\
& + \epsilon_{it} && \text{[unobserved component]}
\end{aligned} \tag{11}$$

where

$\beta_1, \ldots, \beta_6$ = estimated parameters;
AGE, GENDER = individual's characteristics;
NFAIL_{it} = number of unacceptable early and late arrivals until day t;
$\Delta\mathrm{TR}_{it}$ = difference between travel times of commuter i on day t and $t-1$;
$\Delta\mathrm{DT}_{it}$ = departure time that commuter i has adjusted between day t and $t-1$;
δ_{it} = binary indicator variable equal to 0 if $DT_{it} = DT_{it-1}$; otherwise 1; and
ϵ_{it} = error term for commuter i on day t.

The values of the estimated parameters are indicated as follows:

	Early	*Late*
β_1	23.26	17.82
β_2	7.61	4.51
β_3	−5.59	−6.57
β_4	5.49	4.36
β_5	1.16	0.78
β_6	4.17	2.98

The calibrated indifference band for route choice is as follows:

$$\begin{aligned}
\mathrm{IBRC}_{it} = {} & \beta_1 && \text{[initial bands]} \\
& + \beta_2\mathrm{STDTR}_{it} && \text{[dynamic component]} \\
& + \beta_3\mathrm{NFAIL}_{it} && \text{[myopic component]} \\
& + \tau_{it} && \text{[unobserved component]}
\end{aligned} \tag{12}$$

where

$\beta_1, \ldots, \beta_3$ = estimated parameters,
STDTR_{it} = standard deviation of travel time up to day t, and
τ_{it} = error term for commuter i on day t.

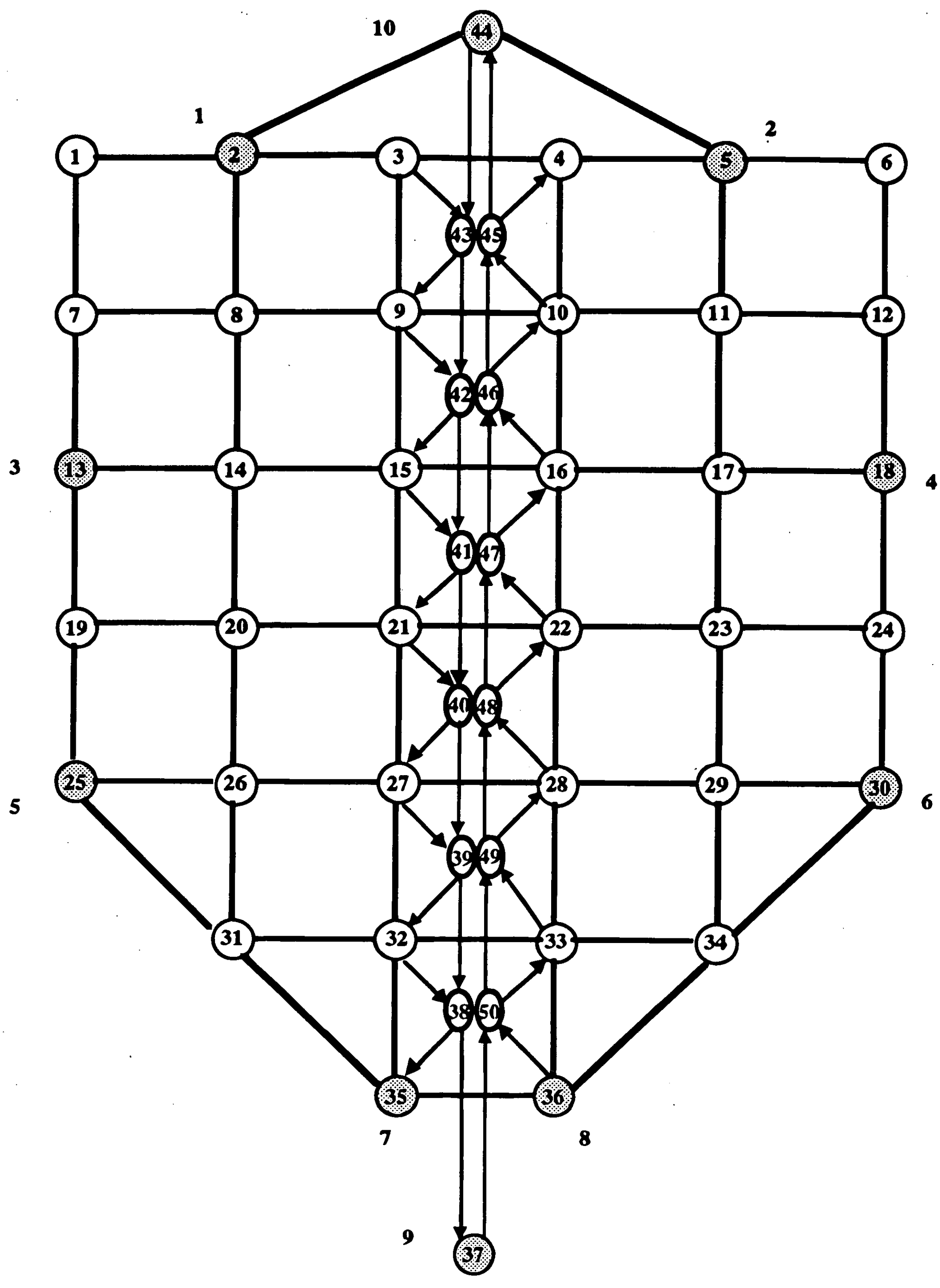


FIGURE 4 Network structure.

The values of the estimated parameters are indicated as follows:

	Early	*Late*
β_1	27.22	18.76
β_2	8.87	4.37
β_3	8.95	9.13

Models of Departure Time and Route Selection

Two particular models, proposed by Small (*19*) and Hendrickson and Plank (*20*), are used. The specification of the functional form proposed by Small can be summarized in the following equation:

$$U_{ij} = -0.106\text{TR}_{ij} - 0.065\text{SDE}_{ij} - 0.254\ \text{SDL}_{ij} - 0.58D1L_{ij} + \epsilon_{ij} \quad (13)$$

where

U_{ij} = measure of utility or "attractiveness" of trip characteristics for individual *i* and alternative *j;*
SDE = max {− SD,0}, early schedule delay for individual *i* under alternative *j;*
SDL = max {SD,0}, late schedule delay,
$D1$L = late dummy variable of 1 if SD $\geq$ 0, and 0 otherwise;

SD = schedule delay, arrival time minus official work start time (min); and

TR = travel time (min).

The originally calibrated utility function was based on 363 observations from four suburban areas and included constant terms for mode, such as drive alone, shared ride, and transit automobile. Some terms in the function are not applicable in this study; therefore, a modified utility function without those terms is used, as follows:

$$U_{ij} = -0.021\mathrm{TR}_{ij} - 0.00042\mathrm{SDE}_{ij} - 0.148\,\mathrm{SDL}_{ij} + 0.0014\,\mathrm{SDL}^2 + \epsilon_{ij} \quad (14)$$

All the variables are the same as those listed earlier. In this particular expression, late arrival incurs a high penalty.

NUMERICAL RESULTS

The numerical results are discussed in three parts. The first part describes the evolution of daily flows in the base case. The results of two random utility maximization models are discussed in the second part. The last part discusses the impact of real-time information in the day-to-day dynamic flow patterns, followed by a brief discussion of computational results.

Base Case

In the base case, all vehicles are assumed to be nonequipped (to receive real-time information), but to have access to path information from the preceding day's experience. Starting with a uniform loading pattern, the day-to-day dynamic flow patterns of Days 1,2, and 14 are indicated in Figure 5. The temporal loading pattern on the first day begins with a uniform profile, starting from 8:05 to 8:30 a.m. (Note: time 0 in the figure corresponds to 8:00 a.m.; the work start time is 8:30 a.m. or Time 30). However, a peak develops from day to day. On Day 14 (final state), fewer than 10 percent of vehicles are still not satisfied with their current selection; the associated pattern indicates that most drivers want to arrive at their preferred arrival time instead of being uniformly distributed along the whole time span. The fact that the dynamic flow pattern shifts dramatically from Day 1 to Day 2 indicates the unreasonableness of the initial uniform load spreading assumption. As expected, peak-period congestion forms because most tripmakers do not wish to arrive too early or too late in relation to their scheduled work time. Although the dynamic flow pattern tends to shift to a higher peak in this case from Day 2 to Day 14, this does not mean that all vehicles will select the same departure time in the final steady state. In the base case, the number of vehicles departing at the peak 5-min interval is about 700 vehicles for Day 2 and 1,010 vehicles for Day 14, an increase of about 50 percent.

The peaks shift from Time 28 of Day 2 to Time 22 of Day 14. Experiencing congestion, most of the drivers choose to leave earlier, although a few of them choose to leave later to avoid the congestion. In the process of adjusting to satisfy the schedule delay constraint, drivers collectively generate more serious congestion, as implied by the higher peak. Although demand managers and traffic control centers seek to spread the demand in a smoother pattern, drivers have a tendency to collectively create a peak-period flow pattern. If this is representative of what happens in actual systems, in-vehicle information systems probably can only shift or raise the peak instead of eliminating it altogether.

Average travel time (ATT) and average stopped time (AST) from day to day are indicated in Figure 6. While starting from a system-optimal solution point, drivers experience longer travel time and greater stopped time from day to day to arrive at their preferred arrival time. The overall average travel time doubles, from about 2.5 to 5.0 min. However, the travel time after Day 11 tends to reach a maximum limit.

Variation of daily time-dependent concentration is indicated in Figure 7. The figure provides a clear picture of system convergence. Although about 10 percent of vehicles are still seeking better alternatives, the system does not change because of those slight varia-

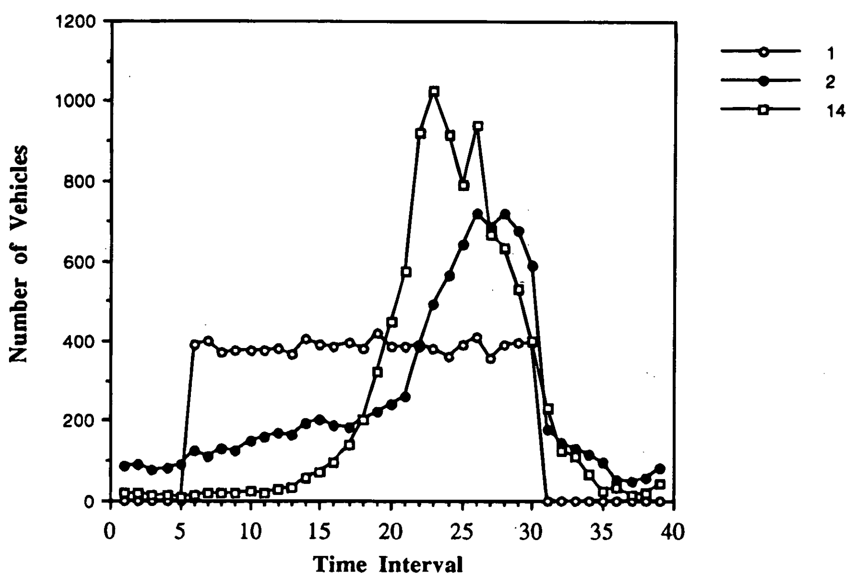


FIGURE 5 Variation of day-to-day dynamic flow patterns (Days 1, 2, and 14) for base case.

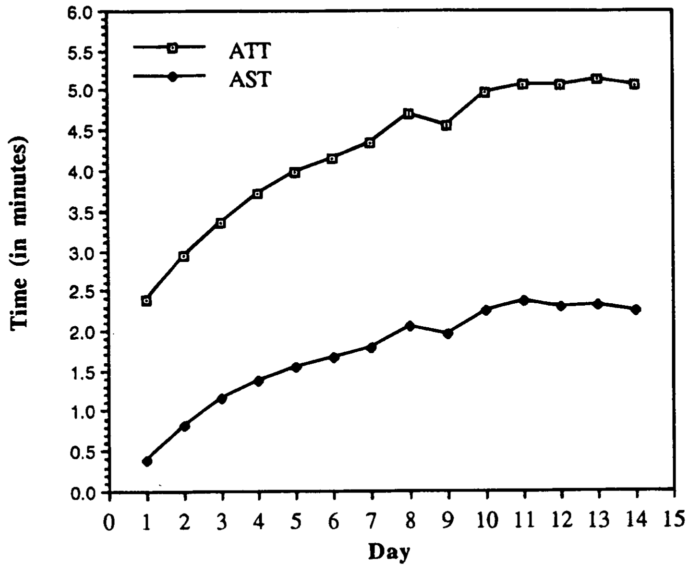


FIGURE 6 Comparison of ATT and AST of day-to-day dynamic flows for base case.

tions. It is evident that a traffic system with a fixed traffic control strategy can always absorb slight variations of demand pattern without this causing additional congestion.

Random Utility Maximization Models

The previous results were based on experiments performed with Hendrickson and Plank's modified model described earlier. Similar experiments were conducted using Small's model. The results indicated in Figure 8 depict similar patterns in terms of the evolution of dynamic flow, switching percentage, and system-wide average travel time. The results suggest that different random utility models might have a similar effect as long as they can capture the relative magnitudes of the travel time and schedule delay. In other words, the day-to-day evolution patterns appear robust vis à vis the underlying choice models.

Effectiveness of Real-Time Information

The effectiveness of real-time information is evaluated from day to day for different market penetrations of equipped vehicles (Table 1). Nonequipped vehicles must continue along their assigned initial path set. However, if equipped vehicles are satisfied with their new paths, they are assumed to use the paths as their initial paths. In this set of experiments, three levels of market penetrations, 10, 25, and 50 percent, and two real-time behavior assumptions,

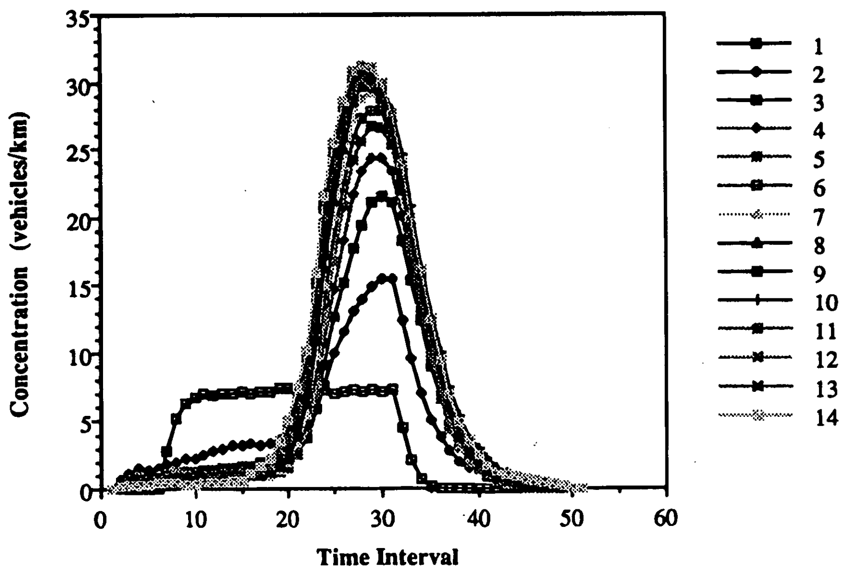


FIGURE 7 Variation of time-dependent network concentration from day to day for base case (1 km = 0.6 mi).

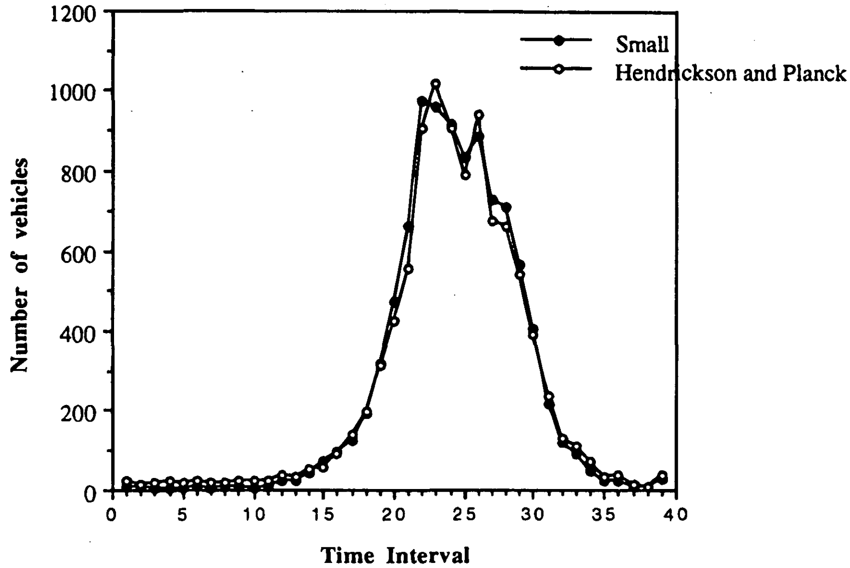


FIGURE 8 Comparison of day-to-day flow patterns with different utility models.

namely, myopic and boundedly rational behavior with a threshold of 0.2 and a minimum bound of 0.5 min, are considered. These tests are termed info-10, info-25, info-50, info-10-b, info-25-b, and info-50-b. Real-time information provides path information for equipped vehicles switching en route; in the meantime, the new experienced paths are collected and added into the path file for all vehicles to use for the next day. In brief, this case is termed "info-50-np," which means new path information is collected through equipped vehicles and distributed to all the tripmakers.

General Flow Dynamics

The evolution of day-to-day dynamic flow patterns is similar to that of the previous cases. Therefore, the results are summarized in Table 1 instead of in the figures. The results show that similar patterns are reached in the final steady state, although with different peak heights, in spite of different assumptions. The peak-period flow pattern indicates that most drivers wish to depart closer to their work schedule times in spite of the congestion. It is surprising to note that real-time information has an insignificant effect on improving the formation of the peak pattern; on the contrary, such information apparently can lead to raising the peak, reducing the travel time, and shifting the peak toward the work start time.

With real-time information, the peaks of all the info cases shift toward the work schedule time. The gap between the base case and info-10 is about 3 min, more than 50 percent of the travel time in these experiments. The info-50 case not only shifts the peak by 3 min but also raises the peak to about 1,100 vehicles. Although the increase of the peak is not quite significant, it offers insight into how drivers respond to real-time information through their day-to-day dynamic choices. These shifts imply that real-time information improves drivers' understanding of the traffic system, so tripmakers select late departure times without delaying their arrival time. In other words, the information system may lead to a reduction in travel time, but the traffic system compensates by attracting more tripmakers to use the facility and maintain the same level of service. Such phenomena are not quite clear in traffic systems and need some validation from field tests.

Real-Time Information Paths

The comparison is made for the info-50 case and the info-50-np case. The loading patterns of the final state (Day 12 for info-50 and Day

TABLE 1 Summary Statistics of Effectiveness of Real-Time Information Experiments

	Time of Peak	Height of Peak (number of vehicles)	ATT (in minutes)	AST (in minutes)	ATD (in km[1])	ATS (in km/hr)	Days for Convergence
BASE	23	1010	5.1	2.2	1.95	22.94	14
info-10	26	994	4.2	1.8	1.78	25.43	11
info-25	26	1065	4.1	1.7	1.78	26.05	11
info-50	26	1085	4.2	1.7	1.76	25.14	12
info-10-b	26	994	4.5	1.9	1.79	23.87	11
info-25-b	26	1062	4.4	1.9	1.81	24.68	11
info-50-b	25	1095	4.3	1.8	1.81	25.26	12
info-50-np	23	1061	5.4	2.3	2.08	23.11	19

[1] 1km = 0.6 mi.

19 for info-50-np) in both cases have a similar shape, but the peak in the info-50-np case is earlier than that of the info-50 case. This early peak implies that vehicles have an earlier departure time to satisfy their indifference bands in the evolution of the info-50-np case. To maintain the same level of convergence, more days are required for info-50-np, probably because of the higher level of congestion.

The execution time of the model on a CRAY YMP for the test network (50 nodes and 168 links) takes about 110 sec/day, including the input/output time from module to module.

CONCLUDING COMMENTS

The analysis of information-based traffic systems needs to consider tripmaker behavior, flow patterns, and traffic control systems. In this paper, two levels of tripmaker decision-making processes are identified: (*a*) day-to-day and (*b*) real-time dynamics. Day-to-day dynamics considers drivers' choices of departure time and route according to indifference bands of tolerable "schedule delay." Real-time dynamics is incorporated within DYNASMART to simulate driver's real-time en-route switching behavior. Flow patterns are obtained by simulating vehicle movement in the network, whereas traffic control systems update flow information or control strategies.

The day-to-day dynamic simulation-assignment framework presented in this paper provides a practical tool for the evaluation of network flows and associated performance measures in information-based traffic systems. The methodology allows investigation of a wide variety of alternatives and provides fundamental insights into the performance of traffic networks under a variety of assumptions on information availability and user behavior.

Naturally, the numerical results presented here should be interpreted with caution, given the limited set of experiments and the nature of the test network and associated conditions. Nonetheless, the results provide useful insights into actual traffic systems. It is also notable that the impact of the real-time information is manifested in several ways: reduces travel time, raises the peak, and pushes the peak toward the work schedule time.

ACKNOWLEDGMENTS

This paper is based in part on work supported by FHWA under a contract entitled Traffic Modeling to Support Advanced Driver Information Systems. Additional support was provided by the U.S. Department of Transportation through the Southwest Region University Transportation Center.

REFERENCES

1. *Strategic Plan for Intelligent Vehicle-Highway Systems.* IVHS America, Washington, D.C., 1992.
2. Mahmassani, H. S., and R. Jayakrishnan. System Performance and User Response under Real-Time Information in a Congested Traffic Corridor. *Transportation Research,* Vol. 25A, No. 5, 1991, pp. 293–307.
3. Van Vuren, T., and D. P. Watlings. A Multiple User Class Assignment Model for Route Guidance. In *Transportation Research Record 1306,* TRB, National Research Council, Washington, D.C., 1991, pp. 22–31.
4. Mahmassani, H. S., S. Peeta, T. Y. Hu, and R. Rothery. Effect of Real-Time Information on Network Performance Under Alternative Dynamic Assignment Rules. *Proc., PTRC Conference,* United Kingdom, Sept. 1993.
5. Bonsall, P. W., and T. Parry. Using an Interactive Route-Choice Simulator to Investigate Drivers' Compliance with Route-Guidance Advice. In *Transportation Research Record 1306,* TRB, National Research Council, Washington, D.C., 1991, pp. 59–68.
6. Allen, R. W., D. Ziedman, T. J., Rosenthal, A. C., Stein, J. F. Torres, and A. Halati. Laboratory Assessment of Driver Route Diversion in Response to In-Vehicle Navigation and Motorist Information Systems. In *Transportation Research Record 1306,* TRB, National Research Council, Washington, D.C., 1991, pp. 82–91.
7. Chen, P., and H. S. Mahmassani. A Dynamic Interactive Simulator for the Study of Commuter Behavior under Real-Time Traffic Information Supply Strategies. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., 1993.
8. Abdel-Aty, A., K. M. Vaughan, P. P. Jovanis, and R. Kitamura. Understanding Traveller Responses to Advanced Traveller Information Systems (ATIS). *Proc., 26th International Symposium on Automotive Technology and Automation,* 1993, pp. 245–253.
9. Mahmassani, H. S. Dynamic Models of Commuter Behavior: Experimental Investigation and Application to the Analysis of Planned Traffic Disruptions. *Transportation Research,* Vol. 24A, No. 6, 1990, pp. 465–484.
10. Mahmassani, H. S., T. Y. Hu, and R. Jayakrishnan. Dynamic Traffic Assignment and Simulation for Advanced Network Informatics (DYNASMART). *Proc., 2nd International Capri Seminar on Urban Traffic Networks,* Capri, Italy, 1992.
11. Ben-Akiva, M., A. de Palma, and P. Kanaroglou. Analysis of Peak-Period Traffic Congestion with Elastic Demand. *Transportation Science,* Vol. 20, No. 3, 1986, pp. 164–181.
12. Ben-Akiva, M., A. de Palma, and P. Kanaroglou. Effect of Capacity Constraints on Peak-Period Traffic Congestion. In *Transportation Research Record 1085,* TRB, National Research Council, Washington, D.C., 1986, pp. 16–26.
13. Cascetta, E. A Stochastic Process Approach to the Analysis of Temporal Dynamics in Transportation Networks. *Transportation Research,* Vol. 23B, 1989, pp. 1–18.
14. Horowitz, J. L. The Stability of Stochastic Equilibrium in a Two-Link Transportation Network; Transportation Research Science, Vol. 18B, No. 1, 1984, pp. 13–28.
15. Mahmassani, H. S., and G. L. Chang. On Boundedly-Rational User Equilibrium in Transportation Systems. *Transportation Science,* Vol. 21, No. 2, 1987, pp. 89–99.
16. Tong, C. C., H. S., Mahmassani, and G. L. Chang. Travel Time Prediction and Information Availability in Commuter behavior Dynamics. In *Transportation Research Record 1138,* TRB, National Research Council, Washington, D.C., 1987, pp. 1–7.
17. Mahmassani, H. S., and G. L. Chang. Dynamic Aspects of Departure Time Choice Behavior in a Commuting System: Theoretical Framework and Experimental Analysis. In *Transportation Research Record 1037,* TRB, National Research Council, Washington, D.C., 1985, pp. 88–101.
18. Chang, G. L., and H. S. Mahmassani. Travel Time Prediction and Departure Time Adjustment Behavior Dynamics in a Congested Traffic Systems. *Transportation Research,* Vol. 22B, No.3, 1988, pp. 217–232.
19. Small, K. A. The Scheduling of Consumer Activities: Work Trips. *American Economic Review,* Vol. 72, 1982, pp. 467–479.
20. Hendrickson, C., and E. Plank. The Flexibility of Departure Times for Work Trips. *Transportation Research,* Vol. 18A, No. 1, 1984, pp. 25–36.
21. Jayakrishnan, R. *In-Vehicle Information Systems for Network Traffic Control: A Simulation Framework to Study Alternative Guidance Strategies.* Ph.D. Dissertation. University of Texas, Austin, 1991.
22. Jayakrishnan, R., H. S. Mahmassani, and T. Y. Hu. An Evaluation Tool for Advanced Traffic Information and Management Systems in Urban Networks. *Transportation Research C,* Vol. 2, No. 3, 1994, pp. 129–147.
23. Mahmassani, H. S., and D. G. Stephan. Experimental Investigation of Route and Departure Time Dynamics of Urban Commuters. In *Transportation Research Record 1203,* TRB, National Research Council, Washington, D.C., 1988, pp. 69–84.
24. Mahmassani, H. S., and S. Peeta. System Optimal Dynamic Assignment for Electronic Route Guidance in a Congested Traffic Network. *Proc., 2nd International Capri Seminar on Urban Traffic Networks,* Capri, Italy, 1992.
25. Jou, R. C., H. S. Mahmassani, and T. Joseph. *Daily Variability of Commuter Decisions: Dallas Survey Results.* Research Report 1216-1. Center for Transportation Research, University of Texas, Austin, 1992.

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

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

Implementing Combined Model of Origin-Destination and Route Choice in EMME/2 System

PAUL METAXATOS, DAVID BOYCE, MICHAEL FLORIAN, AND ISABELLE CONSTANTIN

The issue of "feedback" in the traditional four-step urban travel forecasting procedure (UTFP) has reemerged recently under the pressure of the Clean Air Act Amendments of 1990 and the Intermodal Surface Transportation Efficiency Act of 1991. FHWA now requires that metropolitan planning organizations implement feedback in the UTFP. The combined origin-destination and route choice (OD-UE) model solves simultaneously the trip distribution and the user equilibrium traffic assignment models and hence provides for feedback. Computer codes for the computation and calibration of combined models are available from various researchers. However, they lack detailed documentation and, moreover, require computer programming expertise to adapt to professional practice. In view of these drawbacks, this paper documents the coding of a macro that implements the OD-UE model in EMME/2. The scope of this effort was twofold: first, to respond to certain modeling requirements arising from modern urban transportation planning practice; second, to motivate transportation professionals to use more sound planning methods. The quality of the results obtained using data from the city of Winnipeg, Manitoba, Canada, supports the use of the macro in planning applications.

The issue of "feedback" in the traditional four-step urban travel forecasting procedure (UTFP) has reemerged recently with the impetus of the Clean Air Act Amendments of 1990 and the Intermodal Surface Transportation Efficiency Act of 1991. FHWA now requires that metropolitan planning organizations implement feedback in the UTFP. A sound and mostly appealing alternative toward the solution to this problem is a model that combines the trip distribution, mode split, and assignment steps of the UTFP (*1*). This type of model is not new; its adoption, however, in transportation planning practice is slow. Transportation professionals seem to experience difficulty in understanding the solution procedure. In addition, research codes do not provide relief because they lack detailed documentation and require computer programming expertise to be adapted to professional practice. Moreover, software developers have ignored the issue simply because there has not been sufficient demand, at least until recently.

A number of algorithms that solve the combined origin-destination (O-D), mode choice, and user equilibrium traffic assignment model exist and their properties are well documented (*2,3*). Among those algorithms the Frank-Wolfe linear approximation algorithm and its variant Evans' partial linearization algorithm have been applied to large-scale urban networks. The algorithm implemented here is the one proposed by Evans (*4*); its advantages, compared with those of the Frank-Wolfe algorithm, especially for large-scale applications, have been reported elsewhere (*1,2,5–8*). In a recent study by Boyce et al. (*1*) the Evans algorithm for the combined distribution, mode split, and traffic assignment model was compared against various heuristics used in practice and found to provide superior results, as defined by its more rapid convergence to the true equilibrium solution.

Briefly speaking, four main reasons are presented as favoring the Evans algorithm. First, it is not heuristic; it is, however, a mathematical structure with well-understood properties. Second, the speed with which Evans fills the cells of an O-D matrix (all destinations are loaded from every origin at each iteration) is much superior to Frank-Wolfe (only two destinations per origin per iteration). Third, Evans' partial linearization approximation (as in all approximations of that kind) provides superior feasible directions (subproblem solutions closer to the optimum) compared with Frank-Wolfe linear approximation method. Fourth, the Evans algorithm provides an exact solution of the trip distribution model at each iteration given the current O-D travel costs, whereas the Frank-Wolfe algorithm converges only to the solution of the trip distribution model with equilibrium travel costs. The last becomes an issue in large-scale applications where a solution algorithm never reaches exact convergence because of the high computational costs involved.

In this paper the solution of a combined model of trip distribution and user-equilibrium traffic assignment (OD-UE) is discussed. In a subsequent paper the inclusion of mode choice will be discussed. The implementation of the Evans algorithm is realized by making use of the EMME/2 macro language capability to use various modules for mathematical and network operations both sequentially and iteratively.

The scope of this effort is twofold: first, to respond to the modeling demand arising from the modern urban transportation planning practice; second, to motivate transportation professionals to use more sound planning methods.

The paper is organized as follows. The implementation of the Evans algorithm for the OD-UE model in EMME/2 is documented in the next section. Immediately after, comparisons between the combined model and the sequential procedure are made. Finally, suggestions for future enhancements are made in the last section.

EVANS ALGORITHM FOR COMBINED OD-UE MODEL AND IMPLEMENTATION IN EMME/2

The combined OD-UE model formulated as an equivalent optimization problem requires one to minimize functions of the link travel costs (network term) and the costs of the O-D flows (demand term) subject to conservation of flow constraints, marginal constraints, nonnegativity constraints and definitional constraints. The

P. Metaxatos and D. Boyce, Urban Transportation Center, University of Illinois at Chicago, 1033 West Van Buren Street, Suite 700 South, Chicago, Ill. 60607. M. Florian and I. Constantin, INRO Consultants, Inc., 5160 Decarie Boulevard, Suite 610, Montreal H3X 2H9, Canada.

(Evans) partial linear approximation method linearizes only the first term (network term) of the objective function. The method finds, given a current solution (v, g), a descent direction ($z - v$, $w - g$) by solving a doubly constrained trip distribution model (whereas the Frank-Wolfe algorithm solves a transportation problem of linear programming). The algorithm involves two solutions at each iteration: the main problem solution, and the subproblem solution for determining the direction of descent for an improved main problem solution. The Evans algorithm is described by the following steps:

- Step 0: *Initialization.* Choose an initial solution for link flows, $v_l^0 = 0$, and demand, $g_{ij}^0 = 1$. Set the counter, $k := 0$.
- Step 1: *Update link cost.* s^k: $= s(v^{k-1})$, $k := k + 1$; and compute minimum cost routes c_{ij}^k, on the basis of updated link costs, for every O-D pair (i,j).
- Step 2: *Find the descent direction.*
  –For demand term: Solve a doubly constrained gravity model as a function of the shortest route costs, w_{ij}^k: $w_{ij}^k = A_i^k O_i B_j^k D_j \exp(-\beta c_{ij}^k)$, applying the two-dimensional balancing method;
  –For network term: z^k: Perform an all-or-nothing assignment of demand w_{ij}^k to the shortest routes computed with the updated link costs s^k.
- Step 3: *Compute the optimal step size.* Conduct a line search to find what linear combination of demand and link flows minimizes the objective function; that is, find λ^k, $0 \leq \lambda^k \leq 1$ that minimizes $f(\lambda) = f[v_l^{k-1} + \lambda(z_l^k - v_l^{k-1}); g_{ij}^{k-1} + \lambda(w_{ij}^k - g_{ij}^{k-1})]$.
- Step 4: *Update link flows and demand.* Update the link flows and the demand solution with the best linear combination of solutions from the current and previous iterations, that is, $v_l^k := v_l^{k-1} + \lambda(z_l^k - v_l^{k-1})$ for every link, and $g_{ij}^k := g_{ij}^{k-1} + \lambda(w_{ij}^k - g_{ij}^{k-1})$ for each pair (i, j).
- Step 5: *Convergence check.* If an appropriate convergence criterion is satisfied then stop; otherwise go to Step 1.

Preliminary Considerations

The first task is to build in the same directory as the EMME/2 system a file of the link cost functions of the network where, instead of the usual link flows, the initial solution v_l^0 is read. To be more specific, consider a typical link cost function used in EMME/2 automobile assignment. It is the usual Bureau of Public Roads function

$$s_l(v_l) = s_0 \left[1 + \alpha_1 \left(\frac{v_l}{k_l} \right)^{\alpha_2} \right] \tag{1}$$

where

α_1, α_2 = parameters calibrated from a previous study (typically the values are 0.15 and 4, respectively);
s_0 = free-flow travel time stored in EMME/2 link attribute length;
v_l = autoflow on link l stored in EMME/2 link attribute *volau*;
$s_l(v_l)$ = travel time on link l, an increasing function of autoflow on same link v_l; and
k_l = capacity on link l determined by assumed level of service and stored in EMME/2 link attribute lanes.

It is worth noting that the labels of the EMME/2 link attributes used to store the different arguments of the BPR function need not be taken literally. For example, the label length does not mean the link length in this application. The same is true for the field labeled lanes.

By default, when EMME/2 computes an automobile assignment it reads the link cost functions with flows different from 0 (from some previously performed assignment). To compute an initial solution for the link flows in the initialization step of the algorithm, however, these flows need to be replaced with zero link flows. This can be done by using a text editor to replace the attribute *volau* by *ul1* (which has been previously initialized to 0 in the volume-delay functions stored in the function file. Then by saving the edited file as a separate function file (here saved as *d411.ul1*), it can be read as needed. Thus, the link flows are first initialized to 0, whereas at subsequent iterations *ul1* always stores the current flows.

Step 0: Initialization

The iteration number, controlled by register **x,** is set to 0. Then an initial solution for the demand matrix (g_{ij}^0) and the link flow vector (v_l^0) is computed. The production and attraction vectors (O_i), (D_j), respectively, from an observed matrix are then computed (the observed automobile demand for Winnipeg in 1976 is used, in matrix **mf1**). A zero demand to be used later in the computation of the step size is also computed. Finally, the tolerance level of the secant root-finding method (explained later in Step 3) is saved in a scalar. These operations are summarized in Table 1.

Although the order of the modules employed does not matter from a modeling perspective, it is more efficient to do as many computations as possible in one module before starting to employ the next one. Because there will be many matrixes and scalars involved in the computations, it is a good idea to plan in advance where to store different results. It has been convenient to use the ability of EMME/2 to store full matrixes, O-D vectors, and scalars as **mf"name," mo"name," md"name," ms"name,"** respectively, where **name** is the name of the operand.

Step 1: Link Costs Update and Computation of Minimum Cost Routes

The iteration number x is increased by 1, the link costs vector s^k is updated as $s^k := s(v^{k-1})$, and the matrix of minimum cost routes (c_{ij}^k) is computed for every O-D pair (i, j). To update the link costs the link flows are initialized v^0 to those produced by assigning the demand (g_{ij}^0); otherwise, zero link flows would be used in the update of the link costs. The minimum cost routes (c_{ij}^k) result from an all-or-nothing assignment of the demand ms1$=(g_{ij}^0) = 1$ and are saved in matrix **mf"cijk."** These operations are summarized in Table 2.

Step 2: Computation of Descent Direction

The descent direction for the demand term (w_{ij}) is computed by balancing the matrix $\{\exp(-\beta c_{ij}^k)\}$ to the marginal constraints **mo8,**

TABLE 1 Implementation of Step 0

Module	Purpose	Saved in
3.21	initialize demand to one	ms"gij0"
2.41	initialize link flows to zero	ul1
3.21	compute productions from mf1	mo"produc"
3.21	compute attractions from mf1	md"attrac"
3.21	compute zero demand	ms"zero"
3.21	compute tolerance 10^{-3}	ms"lamacc"

TABLE 2 Implementation of Step 1

Module	Purpose	Saved in
4.11	read link costs based on zero link flows	d411.ul1
5.11	all-or-nothing assignment of ms"gij0"	
5.21	perform the assignment	
	save the shortest routes	mf"cijk"
2.41	link flows from volau	ul1

md8 computed in Step 0. This computation of the doubly constrained gravity model is done by applying the two-dimensional balancing method. To compute the direction of descent for the network term, an all-or-nothing assignment of the demand (w_{ij}) is performed.

It is important to note that the dispersion parameter β is held constant during the solution of the model. To obtain a reasonable value for it, β was set equal to the inverse of the observed mean travel time in the network. For the Winnipeg network (in the demonstration data bank), β was set equal to 0.06.

Finally and only in the first iteration, the demand was initialized (g^0_{ij}) to (w^0_{ij}) (otherwise, in each macro iteration, **ms1** = (g^0_{ij}) = 1 to compute the minimum cost routes) and the main problem link flows v_l to subproblem link flows z_l would be used. These operations are summarized in Table 3.

Step 3: Computation of Optimal Step Size

In each iteration of the algorithm the optimal step size λ^* is obtained by performing a one-dimensional search of the objective function along the feasible direction $\{(z_l - v^k_l), (w_{ij} - g^k_{ij})\}$. This is done by solving the following problem:

$$\min_{\lambda} f(\lambda) = f_l(\lambda) + f_{ij}(\lambda) = \sum_{l \in L} \int^{v_l + \lambda(z_l - v_l)} s_l(x)dx + \frac{1}{\beta} \sum_i \sum_j [g^k_{ij} + \lambda(w_{ij} - g^k_{ij})] \ln[g^k_{ij} + \lambda(w_{ij} - g^k_{ij})] \tag{2}$$

An efficient method of solving Equation 2 is to find the value of λ, which equates the gradient $f'(\lambda)$ to 0, where

$$f'(\lambda) = f'_l(\lambda) + f'_{ij}(\lambda) = \sum_{l \in L} s_l[v_l + \lambda(z_l - v_l)] (z_l - v_l) + \frac{1}{\beta} \sum_i \sum_j \ln[g^k_{ij} + \lambda(w_{ij} - g^k_{ij})] (w_{ij} - g^k_{ij}) \tag{3}$$

To find the 0 of the gradient function, a variation of the secant root-finding method is used. The secant method involves approximating the tangent by a secant through the two most recent iterates and using the 0 of this line as the next iterate. In this particular implementation, one end of the current bracketing interval remains fixed.

An efficient way to compute the gradient of the network term suggested by Heinz Spiess is now presented. The idea consists of using the EMME/2 equilibrium algorithm to compute the new link costs instead of the network calculator. The implementation is described next.

The two λ-values that bracket the search interval are initialized between 0 and 1 and saved in scalars **ms"lam1," ms"lam2"** (in this implementation, **ms61, ms62,** respectively). Note that **ms"lam2"** will also hold the current upper bound of the search interval. Using Module 2.41 the current main problem and subproblem link flows (saved in *ul*1 and *volau*, respectively) are copied to link attributes *ul*3, *ul*2, respectively, of a dummy scenario, say 3000 (created before the macro execution in this case). Their linear combination, *ul*3 + *%msy%* × (*ul*2 − *ul*3) for every λ is computed using Module 2.41 and saved in *ul*1 in the dummy scenario. The need to save them in *ul*1 comes from the definition of the link cost functions in Module 4.11 where the link flows are saved in *ul*1.

To compute the costs of those combined flows, an all-or-nothing assignment is performed in the dummy scenario with zero demand. The link costs based on the link flows in *ul*1 are saved by default in the link attribute *timau*. When the line search begins (in each iteration of the macro) register **y**, which controls which of the two λ-values is read, is set to $y = 61$, whereas register **z,** which keeps track of the respective gradient values for the network term, is set to $z = 71$.

The gradient of the network term in the dummy scenario is finally computed after the evaluation (in Module 2.41) of the sum of the expression 0.06 * *timau* * (*ul*2 − *ul*3). The sum is saved in scalar *ms%z%*. The multiplication by $\beta = 0.06$ is equivalent to multiplying the demand term of the gradient by $1/\beta$. In this manner, scalar **ms71** contains the sum of the gradient values for the network term with respect to the first λ (in **ms"lam1"**), and scalar **ms72** contains the sum of the gradient values for the network term with respect to the second λ (in **ms"lam2"**).

Back in the working scenario, what remains to be computed is the gradient for the demand term. Remembering that the main problem demand is saved in **mf"gijk"** and the subproblem demand in **mf"wij,"** the last task involves the computation of the expression

$$\sum_i \sum_j \ln(mf\text{“}gijk\text{”} + \%msy\% * put[mf\text{“}wij\text{”} - mf\text{“}gijk\text{”})] * get(1) \tag{4}$$

where the special functions **get** (.) and **put**(.) are used to save some computation time (*9*, pp. 3–67). The summation over all origins and destinations in the matrix calculations of Module 3.21 gives the gradient for the demand term that is saved in scalar **ms"gradem."** It is worth noting that care is taken to avoid evaluating the last expression for zero values (because the logarithm of 0 is not defined). This can be accomplished in module 3.21 by providing a constrained matrix and a constrained interval. In this case using the default-

TABLE 3 Implementation of Step 2

Module	Purpose	Saved in
3.21	compute a function of shortest route costs as exponential	mf"ecijk"
3.22	balance mf"ecijk" to mo"produc" and md"attrac"	mf"wij"
5.11	all-or-nothing assignment of mf"wij"	
5.21	perform the assignment	
3.21	iteration 1: $g^k_{ij} = w_{ij}$	mf"gijk"
2.41	iteration 1: $v^k_l = z_l$.	ul1

constrained interval (0, 0, exclude), only nonzero values are retained. The same is done whenever the logarithm of the demand matrix is involved in the computations.

So far, for each λ the gradients for both the network and the demand terms have been computed. The total gradient is then saved in scalar ms%z% as %msz% + **ms"gradem."** For example, for the first λ in **ms"lam1"** the total gradient is saved in **ms71** containing the sum of the contents of scalar **ms71** (the gradient for the network term) and the contents of scalar ms"gradem" (the gradient of the demand term). This procedure is repeated once more for the second λ-value. Note that if the gradient is found to be positive the optimal step size is set to 0 and the Evans algorithm terminates because the current solution is optimal; however, in real problems such a result never occurs because the optimal solution is never reached.

The next task is to compute the slope of the secant line. In particular, using Module 3.21 the slope ϕ of the line $\{[\lambda_{k-1}, \nabla(\lambda_{k-1})], [\lambda_k, \nabla(\lambda_k)]\}$ is computed and saved in scalar **ms"phil12"** as

$$ms\text{“phil2”} = (ms72 - ms71)/(ms62 - ms61) \tag{5}$$

The optimal λ, saved in scalar **ms"xlopt"** (in ms70 here), is next computed from the formula

$$ms\text{“xlopt”} = (0 - ms71)/ms\text{“phil2”} \tag{6}$$

Finally, the convergence of the secant loop is monitored as follows. First, the following expression is evaluated:

$$1 \times \left\{ \frac{abs(ms\text{“xlopt”} - ms\text{“xlamn2”})}{abs(ms\text{“xlopt”})} - ms\text{“lamacc”} \right\} \leq 10^{-6} \tag{7}$$

Equation 7 is a boolean expression with values of 1 for true and 0 for false. If the result of this evaluation is 1, then the secant loop has converged. A value of **ms"lamacc"** $= 10^{-3}$ is used as a stopping criterion. The optimal step size taken from the last secant iteration is then used in Step 4 to update the current solution of the combined model. If the result of the above expression is not 1, the secant loop is repeated. Table 4 summarizes these operations.

Step 4: Current Solution Update

The main problem demand solution (from the previous iteration), currently in matrix **mf"gijk,"** is saved in matrix **mf"gijk − 1."** The current solution for demand is then a weighted average (optimal weight) of the previous (main) problem solution and the current (subproblem) solution as follows:

$$mf\text{“gijk”} = mf\text{“gijk} - 1\text{”} + ms\text{“xlopt”} \times (mf\text{“wij”} - mf\text{“gijk} - 1\text{”}) \tag{8}$$

The same is done for the link flows. The main problem link flows solution (from the previous iteration), currently in link attribute *ul1,* is saved in link attribute *ul2.* The current solution for the link flows is then a weighted average (optimal weight) of the previous (main) problem solution and the current (subproblem) solution, as follows:

$$ul1 = ul1 + \%ms\text{“xlopt”}\% * (volau - ul1) \tag{9}$$

Table 5 summarizes the above operations. Note that in Module 2.41 the matrix operations cannot be performed. Therefore, the optimal λ has to be represented not as the scalar **ms"xlopt"** but rather as the contents of the scalar **ms"xlopt."**

Step 5: Criteria for Convergence

To monitor the convergence rate of the algorithm with respect to the solution for demand, the maximum over all terms (origins and destinations) of the absolute deviations between the current solution and the solution from the previous iteration is considered and saved in scalar ms"gdif"; that is,

$$\max_{i\in I, j\in J} \|g_{ij}^k - g_{ij}^{k-1}\| = \max_{i\in I, j\in J} \|mf\text{“}gijk - mf\text{“}gijk - 1\text{”}\| \tag{10}$$

The convergence of the link flows is monitored, similarly, by computing the maximum over all links of the absolute deviations between the current solution and the solution from the previous iteration and saved in scalar **ms"vdif"**; that is,

$$\max_{l\in L} \|v_l^k - v_l^{k-1}\| = \max_{l\in L} \|ul1 - ul2\| \tag{11}$$

Another convergence criterion that is strongly recommended is the current value of the GAP function. At each iteration of the Evans algorithm, the subproblem solution provides a lower bound for the objective function value. That is, the current GAP is the distance from the current value of the objective function to the lower bound. The current value of the GAP function for the combined (OD-UE) model at iteration *k,* which is simply the value of 3 corresponding to $\lambda = 0$, is

$$\text{GAP}^k = LB^k - f^k(v, g) = \sum_{l\in L} s_l(v_l^k)(z_l - v_l^k) + \frac{1}{\beta} \sum_i \sum_j \ln(g_{ij}^k)(w_{ij} - g_{ij}^k) \tag{12}$$

where $f^k(v, g)$ is the current value of the objective function. The GAP function converges to 0, although not monotonically.

An alternative convergence criterion that may also serve as a condition for the termination of the algorithm is to test at each iteration a "modified" relative gap for the network flows and the demand because

TABLE 4 Implementation of Step 3

Module.	Purpose	Saved in
3.21	$\lambda_1 = 0$	ms"lam1"
3.21	current upper bound of search interval: λ_2	ms"lam2"
2.41	gradient of network term for λ_1, λ_2	ms71,ms72
3.21	gradient of demand term (each λ)	ms"gradem"
3.21	total gradient for each λ	ms71,ms72
3.21	the slope of the secant line	ms"phil2"
3.21	optimal λ	ms"xlopt"

TABLE 5 Implementation of Step 4

Module	Purpose	Saved in
3.21	save demand from previous iteration	mf"gijk-1"
3.21	update demand solution	mf"gijk"
2.41	save link flows from previous iteration	ul2
2.41	update link flows solution	ul1

both of them converge to their equilibrium values. For the network flows the "modified" relative gap RG_l^k at iteration k is defined as

$$RG_l^k = \frac{\sum_{l \in L} s_l(v_l^k)\,(z_l - v_l^k)}{\sum_{l \in L} s_l(v_l^k) v_l^k} \quad (13)$$

while, for the demand terms, the "modified" relative gap RG_{ij}^k at iteration k is defined as

$$RG_{ij}^k = \frac{\sum_i \sum_j \ln(g_{ij}^k)(w_{ij} - g_{ij}^k)}{\sum_i \sum_j \ln(g_{ij}^k) g_{ij}^k} \quad (14)$$

Eventually as $\sum_{l \in L} s_l(v_l^k) z_l \rightarrow \sum_{l \in L} s_l(v_l^k) v_l^k$ and $w_{ij} \rightarrow g_{ij}^k$, and all the demand is on shortest routes, both these measures go to 0. However, they are not decreasing monotonically (which is also true for the relative gap in the fixed demand user equilibrium traffic assignment). Table 6 summarizes these operations.

COMPARISON BETWEEN THE COMBINED MODEL AND THE SEQUENTIAL PROCEDURE

The results presented below were obtained by solving the combined model for the city of Winnipeg, Manitoba, Canada. The network consists of 154-zone centroids, 903 regular nodes, and 2,535 automobile links. The computations were performed in a SUN SPARC-2 workstation with 64MB of memory; the macro needs about 46 sec (real time) or almost 19 sec (central processing unit time) time per iteration. The observed (1976) automobile demand in **mf1** was increased by 50 percent because the network was not very congested.

As presently implemented, the macro iterates until some prespecified convergence criterion for the link flows is satisfied. It is straightforward to apply any of the stopping criteria suggested earlier. The various performance measures from the application of the macro to the Winnipeg network are indicated in Figure 1. The rates of convergence of the optimal step size, demand, "modified" relative GAP for demand RG_{ij}, link flows, and the GAP function are satisfactory. Although it is not expected that more than 10 or 20 iterations are required in practice, the results from additional iterations provide information about the convergence of the Evans algorithm.

Further evidence of the quality of the results can be seen in a link scattergram in Figure 2. In the absence of observed link flow data the macro for 10 iterations has been solved and the obtained link flows in *ul3* (horizontal axis) have been saved. Then a trip distribution model was estimated on the basis of free-flow travel times and balanced to the production and attraction totals of the observed automobile demand matrix in **mf1** increased by 50 percent. The estimated trips were then assigned to the network for 10 iterations and the link flows obtained were saved in *ul1* (vertical axis). The plot shows that the link flows from the combined model are lower (better converged) than from a trip table based on free-flow travel times. If the two methods were equivalent, the points would lie on the line shown in the figure. Because the points lie above the line, the link flows from the four-step procedure are higher, which results from the longer trips on the basis of free-flow travel times.

In addition to the plot, a number of statistics for both variables in the combined model (automobile link flows and automobile O-D flows) have been computed. The purpose here is to compare two pairs of variables: first, the trip table estimated from free-flow travel times with the trip table from the combined model (after 10 iterations); and second, the link flows after the assignment of the estimated trip table for 10 iterations with those obtained from the solution of the combined model. The root mean square error (*RMSE*) and the χ^2 statistics reported in Table 7 are based on the following formulas:

$$RMSE = \left\{ \frac{\sum_{i=1}^{m} (M_i - T_i)^2}{m} \right\}^{0.5} \quad (15)$$

$$\chi^2 = \sum_{i=1}^{m} \left\{ \frac{(M_i - T_i)^2}{T_i} \right\} \quad (16)$$

where

T = solution from combined model,
M = solution from sequential procedure, and
m = number of data elements with positive values.

Zero values in the solutions were removed because these values are a property of the model formulation or the data, rather than the solution method.

FURTHER CONSIDERATIONS

Finally, a number of possible improvements and extensions of the initial formulation and implementation of the OD-UE model are considered. In this implementation of the OD-UE model, car drivers seek to minimize their travel time. Obviously, travel time is just one component of the travel cost. Other components may include monetary costs incurred by owning (for example, depreciation costs) and operating a car (insurance costs, fuel costs, parking costs, tolls,

TABLE 6 Implementation of Step 5

Module	Purpose	Saved in
3.21	maximum absolute difference of mf"gijk" and mf"gijk-1"	ms"gdif"
2.41	maximum absolute difference of ul1 and ul2	ms"vdif"
3.21	"modified" relative GAP for demand	ms"rgdem"
3.21	current GAP	ms"gap"

TABLE 7 Comparison Between Combined Model and Sequential Procedure

Variable	Positive Flows	RMSE	Desired	χ^2	Desired
Auto Link Flows	2,366	119.41	0	84054.31	0
Auto O-D Flows	18,630	2.35	0	4104.19	0

etc.). Consideration of these additional costs requires changes in the model formulation and, of course, data availability.

The model formulation considered here assumes one person per car. However, car occupancy data by origin zone are immediately available in the demonstration data bank of EMME/2 and could have been used in the macro. In addition, the model formulation can be modified to accommodate the occupancy factor endogenously. This idea may be applied when the model is used to assess air quality impacts from relevant policy interventions.

The combined model can be enhanced to include other modes of travel. This involves reformulating the combined model to account for mode choice and is being pursued in the application of the macro in a sketch planning network for Chicago [Boyce et al. (*10*)]. Finally, it is hoped that transit operations can be integrated into the macro and that the impacts of changing parameters such as waiting time, loading time, and headway can be studied.

Although all the above extensions and improvements are possible and interesting to study, it is not known how they will affect the per-

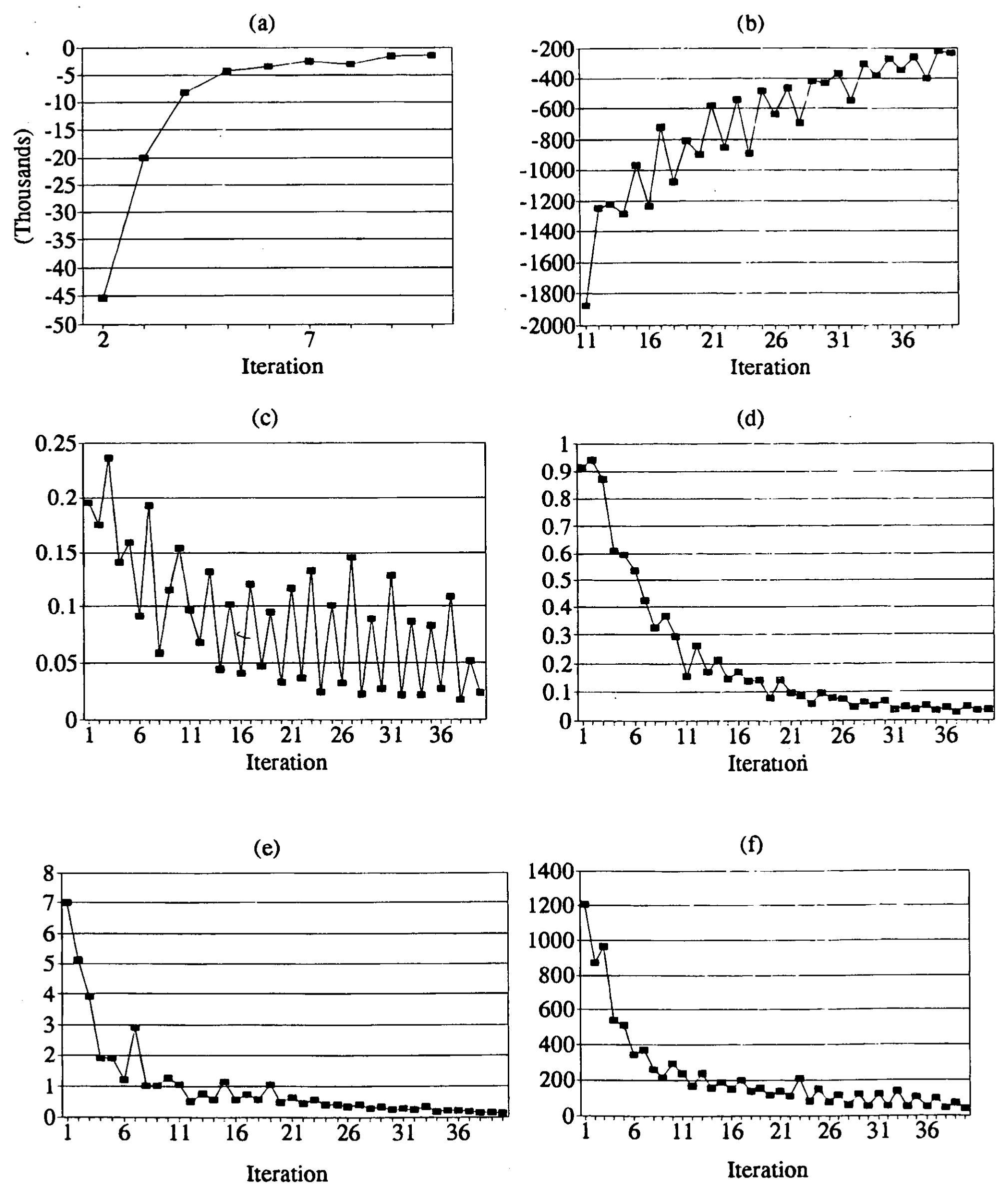


FIGURE 1 Monitoring the convergence of the Evans algorithm: (*a*) GAP function (Iterations 2–10); (*b*) GAP function (Iterations 11–40); (*c*) optimal step size; (*d*) modified relative GAP for demand; (*e*) maximum absolute deviations for demand; and (*f*) maximum absolute deviations for link flows.

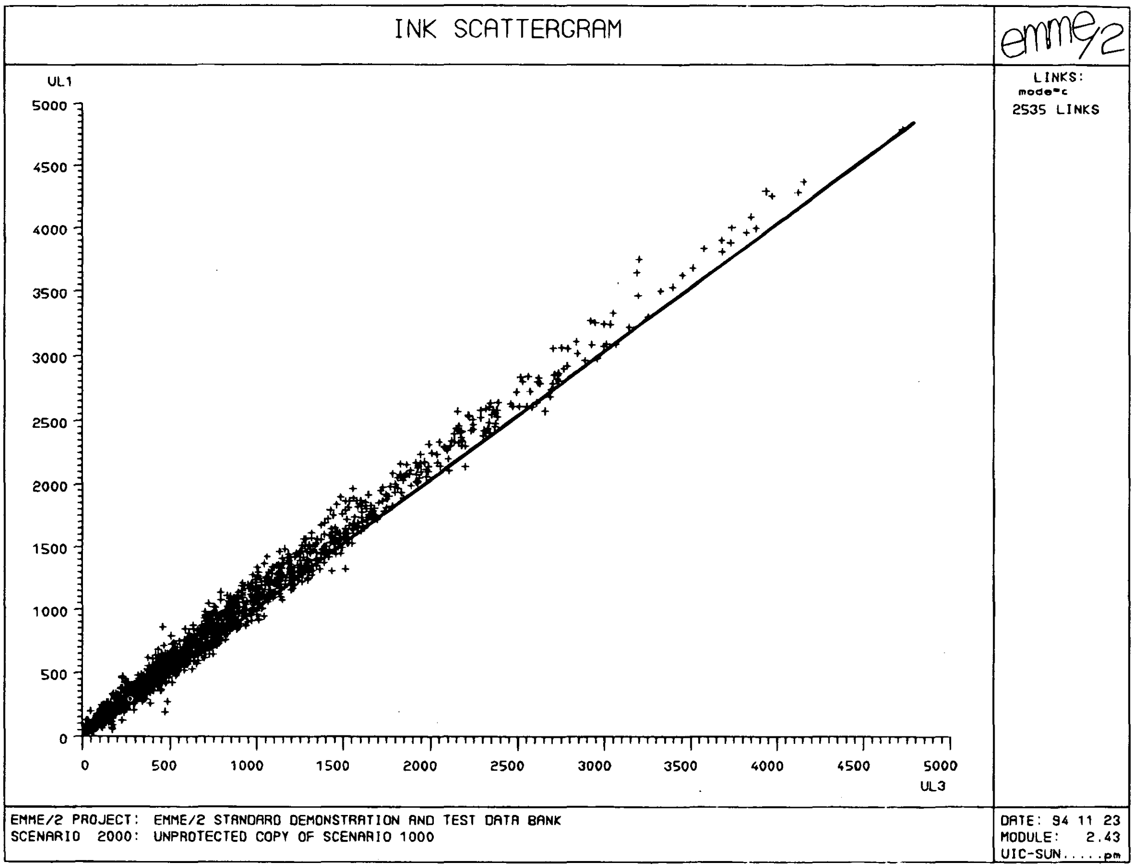


FIGURE 2 Link flows solution of sequential procedure versus link flows solution of combined model.

formance of the macro with respect to the computer requirements. If they can add to the detail in representing travel behavior without overburdening the computational effort, then such a modified macro can be seen as a powerful planning tool. Meanwhile, the implementation of a combined model in EMME/2 can meet some of the modeling requirements arising from modern urban transportation planning practice and motivate transportation professionals to use more sound planning methods. The quality of the results obtained to date seems to encourage the use of the macro in planning studies.

ACKNOWLEDGMENTS

The research presented in this paper began with a visit by the first author to the office of INRO Consultants in Montreal, with the support of the University of Illinois at Chicago. The authors also are indebted to three anonymous referees, whose comments greatly improved the quality of the paper.

REFERENCES

1. Boyce, D. E., Y.-F. Zhang, and M. R. Lupa. Introducing "Feedback" into Four-Step Travel Forecasting Procedure Versus the Equilibrium Solution of Combined Model. In *Transportation Research Record 1443,* TRB, National Research Council, Washington, D.C., 1994, pp. 65–74.
2. Florian, M. Nonlinear Cost Network Models in Transportation Analysis. *Mathematical Programming Study,* Vol. 26, 1986, pp. 167–196.
3. Patriksson, M. *The Traffic Assignment Problem: Models and Methods.* VSP BV, Utrecht, The Netherlands, 1994.
4. Evans, S. P. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research,* Vol. 10, 1976, pp. 37–57.
5. Frank, C. *A Study of Alternative Approaches to Combined Trip Distribution-Assignment Modelling.* Ph.D. thesis. Department of Regional Science. University of Pennsylvania, Philadelphia, 1978.
6. LeBlanc, L. J., and K. Farhangian. Efficient Algorithms for Solving Elastic Demand Traffic Assignment Problems and Mode Split Assignment Problems. *Transportation Science,* Vol. 15, 1981, pp. 306–312.
7. Boyce, D. E., L. J. LeBlanc, and K. S. Chon. Network Equilibrium Models of Urban Location and Travel Choices: A Retrospective Survey. *Journal of Regional Science,* Vol. 28, 1988, pp. 159–183.
8. Chu, Y.-L. Combined Trip Distribution and Assignment Model Incorporating Captive Travel Behavior. In *Transportation Research Record 1285,* TRB, National Research Council, Washington, D.C., 1990, pp. 70–77.
9. *EMME/2 User's Manual: Software Release 7.* INRO Consultants, Inc., Montreal, Quebec, Canada, 1994.
10. Boyce, D. E, M. Tatineni, and Y.-F. Zhang. *Scenario Analyses for the Chicago Region with a Sketch Planning Model of Origin-Destination Mode and Route Choice.* Final Report to Illinois Department of Transportation. Urban Transportation Center, University of Illinois, Chicago, 1992.

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

Evolutionary Transportation Planning Model: Structure and Application

DAVID M. LEVINSON

An evolutionary transportation planning model wherein the demand in a given year depends on the demand of the previous year is described. The model redistributes a fraction of the work trips each year associated with the relocation of a household or taking a new job, whereas changes in distribution associated with growth (or decline) are considered. This hybrid-evolutionary model is compared with an equilibrium model, wherein supply and demand are solved simultaneously. The reasons for preferring the evolutionary method to the equilibrium approach are several: (*a*) the ability to more easily use observed data and thereby limit modeling to changes in behavior, (*b*) additional realism in the concept of the model, (*c*) the provision of a framework for extension to integration with land use models, and (*d*) the additional information available to policy makers.

Traditionally, transportation planning models are used to forecast levels of traffic or transit ridership at a given point in time. Best practice in travel forecasting, the equilibrium approach, attempts to simultaneously (or iteratively) solve for travel demand given a congested network and to estimate network congestion given the travel demand. However, at no point in time is the demand/supply system actually in perfect equilibrium. Individuals and firms continuously enter and leave the system. Changes in system performance, such as the travel times between places, lead to further changes in user behavior, such as choice of route, mode, departure time, sequence of trips, or destination. Some of these behavioral changes are made readily with only a short lag. The disruptive nature and high transaction costs of others, such as switching jobs or moving to a new residence, mean they are undertaken rarely.

This paper presents and tests an alternative approach to travel demand modeling, which explicitly considers changes over time in work trip distribution as a result of household relocation and job switching. The behavioral theory underlying this model is not the perfect network equilibrium of Wardrop or the supply/demand equilibrium described by Boyce et al. (*1*). Rather, it is comparable to Simon's idea of bounded rationality, where the costs of changing behavior need to be considered as well as the possible suboptimality of that behavior (*2*). Thus, supply and demand are not in perfect equilibrium because the costs of moving and switching jobs are high. Traffic assignment may not be in perfect equilibrium because individuals do not have perfect information about the dynamically changing travel times between places.

The approach presented here is therefore more analogous to an evolutionary model than an equilibrium model. The dichotomy and connection between the two have long been recognized (*3*). In a strictly evolutionary model, decisions are updated continuously (or in more practical terms on some time slice such as a day-to-day basis), with some time lag between obtaining information and executing a change in behavior. Moreover, the time lag for response may vary on the basis of the type of decision and the characteristics of the individual making the decision. In this paper's hybrid-evolutionary model, day-to-day decisions are still treated as though they are in equilibrium, but long-term decisions are lagged. In this case, only a fraction of work trips are redistributed every year, with congested travel times on the basis of the previous year's results serving as the source of impedance. In addition, trips from new homes and jobs are also distributed on the basis of those times. One key question is, To what extent do different travel patterns emerge from the evolutionary modeling approach compared with an equilibrium approach?

In addition to being more realistic, one advantage to the evolutionary approach is the ability to start with observed data such as the Journey to Work census data or a trip table synthesized from traffic counts and an old trip table. The evolutionary approach (in this paper, a synthesized trip table is used as a seed) can begin with all of the information inherent in these data rather than just the impedance curves derived from them and evolve incrementally from observed conditions rather than be modeled in totality. This approach is expected to be better than simply applying zone-to-zone adjustment factors at the end of the equilibrium modeling process to correct demand for under- or overestimation because it reduces the amount of error introduced by modeling.

The evolutionary approach should also have significant advantages for future application to land use forecasting and combined transportation–land use forecasts. Although the transportation model is a largely negative feedback loop—more demand creates more congestion, which leads to less demand—the land use model is in some respects a positive feedback loop: more development increases accessibility, which leads to more development. At the extremes, positive feedback leads to the densities found in Manhattan or Hong Kong. The integration of positive and negative feedback results in a complex model that is more sensitive to historical patterns and initial conditions than a simpler equilibrium-seeking negative feedback loop. However, the model presented in this paper considers land use changes as exogenous for two reasons: (*a*) the lack of resources to calibrate a land use model to the necessary accuracy, and (*b*) the lack of support for computer modeling of land use. Planners in the Washington, D.C., area prefer a hand-crafted approach using Delphi methods for forecasting land use.

Further, many policy decisions, such as the programming of capital facilities, are made by analysis of a single equilibrium point in time. An evolutionary model can measure the transportation system over multiple time slices and give a more accurate reflection of benefits and costs.

The largest drawback to the evolutionary approach is the additional computational time required to implement the system as opposed to a one-shot equilibrium solution. If the results are not sufficiently different, or the additional information is not useful, the benefit may not be worth the additional computer resources and

Department of Transportation Studies, Department of Civil Engineering, Room 108, McLaughlin Hall, University of California, Berkeley, Calif. 94720.

complexity. A second consideration is the requirement for additional information. In an evolutionary model, the different time lags in decision making must be determined. In this case, how frequently do individuals relocate? Here, a fixed value of 22.5 percent of individuals is taken to change jobs, houses, or both every year, a figure derived from the 1991 Montgomery County Travel Survey (*4*), but future research should model the value endogenously on the basis of socioeconomic, demographic, and transportation accessibility variables for a given area or trip interchange.

Next in this paper is a discussion of model structure, which includes frameworks for modeling travel demand, a model of relocation behavior, and flowcharts of the hybrid-evolutionary and equilibrium models. This discussion is followed by a description of the model components used in this application. The model inputs of land use, demographics, and networks are presented. A comparison of the convergence properties of the two models is shown. A section comparing the results of the two models is provided. The conclusion discusses some of the questions raised by the evolutionary model.

MODEL STRUCTURE

The model structure is presented in this section. First is a look at modeling frameworks, considering equilibrium and evolution as two poles with two interim combinations of the methods, depending on the decision time horizon evaluated. Next is a presentation of how relocation is incorporated into the model system mechanically. Finally a comparison of flowcharts of the two tested models—hybrid evolutionary and equilibrium—is presented.

Travel Demand Modeling Frameworks

Several approaches can be taken in testing the concept of a dynamic demand model. Each approach is a variation on the spectrum between a lagged model, in which decisions are not simultaneously made by all commuters, and an equilibrium model. In the aggregate models tested here, it is assumed that there are two time frames for travel decisions: day-to-day and year-to-year. Day-to-day decisions include route choice, mode choice, departure time choice, and nonwork trip destination choice. Year-to-year decisions include relocation or work trip (re)distribution (for a fraction of commuters), automobile ownership, and trip (re)generation. These decisions are not entirely separable, so endogenous year-to-year decisions (location/work trip distribution) reflect changes in the day-to-day conditions. In addition, the following system variables vary annually; network, land use, demographics. Although there is a continuum of decision making in reality, this approach is taken for the sake of simplicity.

Further it is assumed that year-to-year decisions are lagged and are based on information from the previous year but that day-to-day decisions are essentially in equilibrium between demand and supply.

The models are as follows:

- Model 1. Equilibrium: equilibrium for day-to-day and year-to-year decisions;
- Model 2. Hybrid: equilibrium for day-to-day, evolution for year-to-year decisions;
- Model 3. Evolutionary: evolution for day-to-day and year-to-year decisions; and
- Model 4. Alternative hybrid: evolution for day-to-day, equilibrium for year-to-year decisions.

Because of computational intensity (3,652 days for 10 years, requiring a demand update on each day), Model 3 is not pursued here. In addition, Model 3 would need to account for variations in demand because of day- of the week and month of the year. Model 4, an alternative hybrid model, would use dynamic assignment, scheduling, and departure time, perhaps with responsive intersection control, to come up with information used in long-term decisions, which would be assumed to be in equilibrium, and is the opposite of Model 2. In all of the Model 2 runs here, the yearly decisions (trip generation, work trip distribution, and automobile ownership) are computed as lagged decisions.

Relocation

For an evolutionary analysis, a new model component is required. This concerns the decision to relocate: both moving one's home or switching jobs is a relocation decision. Here, the terms relocate and redistribute are considered synonymous, the difference in terms resulting from alternative perspectives: individuals choose to relocate while social planners redistribute individuals (match their home and workplace) in their demand models. The nature of this model is that the number of trips at time t depends on the trip pattern at time $t - 1$ plus any change forecast to happen. This is an inertial, state-dependent approach; a work trip does not change from year to year unless some outside force (a redistribution/relocation decision) causes it to change. On a much longer time scale, long-term location (and hence trip frequency/destination choice) decisions can be seen as analogous to trip chaining, where decisions are history dependent. Kitamura has shown for trip chaining that the use of lagged dependent variables is a plausible and statistically valid specification (*5*). Clearly, empirical and statistical issues will need to be further investigated for relocation choice to determine the best specification in terms of predictive value while avoiding serial correlation problems.

This model needs to rematch a fraction of all workers and jobs into work trips for each time slice (in this case, each year). Further study is necessary to understand whether these recently redistributed trips are of longer, shorter, or the same duration as average trips after controlling for the number of opportunities and competing job seekers. This question is analogous to the difference between marginal and average costs in economics. In this application, the work trip distribution impedance curves were estimated from a survey sample of the entire population (not only those who recently moved).

The following equations are used:

$$T_{ij}^{y} = (1 - R_{ij}) \times T_{ij}^{y-1} + MN_{ij}^{y} \tag{1}$$

where

T_{ij}^{y} = trips from i to j in year = y,
MN_{ij}^{y} = switched job/house and new trips (subject to redistribution),
M_i^y = trips from i in year y which switched from year = $y - 1$,
M_j^y = trips to j in year y which switched from year = $y - 1$,
N_i^y = trips from i caused by growth (not present in year = $y - 1$),
N_j^y = trips to j caused by growth (not present in year = $y - 1$), and
R_{ij} = relocation function for interchange $i - j$ (= 0.225 in this application)

subject to

$$T_i^y = \sum_{j=1}^{J} T_{ij}^y \quad (2)$$

$$T_j^y = \sum_{i=1}^{I} T_{ij}^y \quad (3)$$

$$M_i^y = R_{ij} T_{ij}^{y-1} \quad (4)$$

$$M_j^y = R_{ij} T_{ij}^{y-1} \quad (5)$$

For work trip distribution, a two-dimensional balancing procedure is used. For this, the rows (origins) and columns (destinations) are balanced. The total of origins (O_i^y) balanced here is

$$O_i^y = N_i^y + M_i^y \quad (6)$$

and the destinations (D_j^y) is

$$D_j^y = N_j^y + M_j^y \quad (7)$$

which after balancing, produces the trip table MN_{ij}^y, which is added to the fraction of trips unchanged from the previous year to obtain the final peak-period work trip table.

The following table shows the logic of whether an individual would be redistributed:

		Change Work Location	
		Yes	*No*
Change	*Yes*	Redistribute	Redistribute
Home	*No*	Redistribute	Do not redistribute

Flow Charts

Figures 1 and 2 show the flow chart of the equilibrium and hybrid-evolutionary models, respectively. The endogenous components are identified with rectangles; the exogenous updates to land use and networks are shown with curved corners.

To summarize, exogenous changes over time in the model include updates to the transportation network through the addition of links, updates of the age distribution, and household size distribution by geographical area developed from a separate demographics model and updates of the land use activity (housing units and employment by type) from regional transportation forecasts. These are discussed in detail later. Collectively, these inputs are treated exogenously because, in the short term, there is little interaction between them and travel demand. The longer the timeframe, the more that can reasonably be internalized.

Endogenous changes within this model include updates to travel times on the network, the number of work trips generated, and consequently the interchange of work trips between zones, and in a more complete extension of the proposed model, the relocation decision. Travel demand is not limited to just trips generated, but considers all of the choices in the travel demand process. Therefore, a shift in mode, route, or time-of-day is a change in demand for a facility or route just as an increase or decrease in the number of trips generated is a change in demand for the transportation system as a whole.

MODEL COMPONENTS

The model components (trip generation, trip distribution, mode choice, departure time choice, route choice, and intersection control) used here are the same as those estimated for the Travel/2 model (*6*). Briefly, these are described as follows.

Trip Generation

The person-trip generation model is in two parts (*7*): for the home end of trips, a cross-classification model based on age, household size, and dwelling unit type; for the nonhome end, generation is

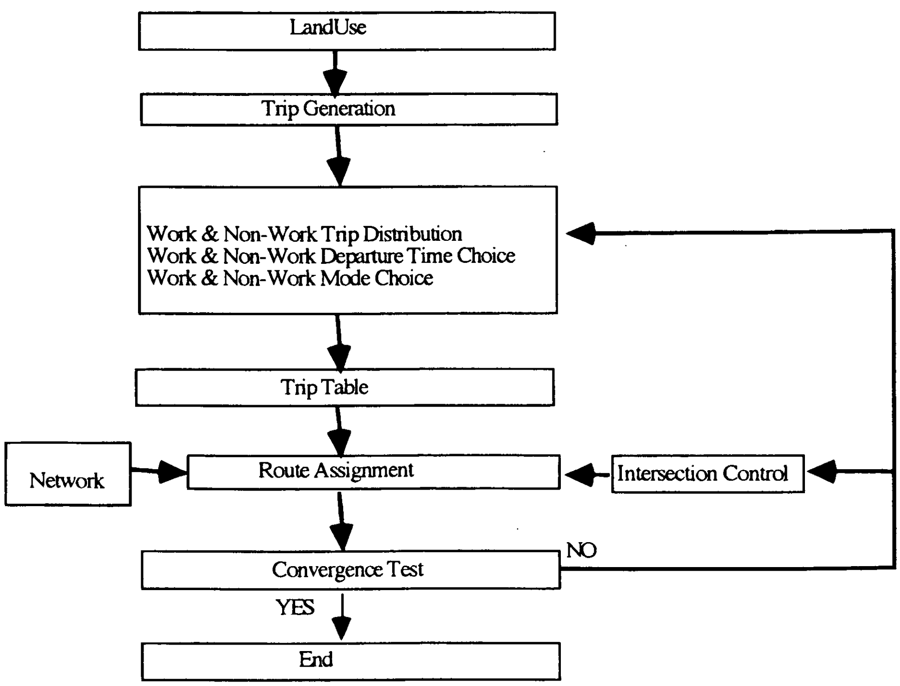


FIGURE 1 Flow chart equilibrium model.

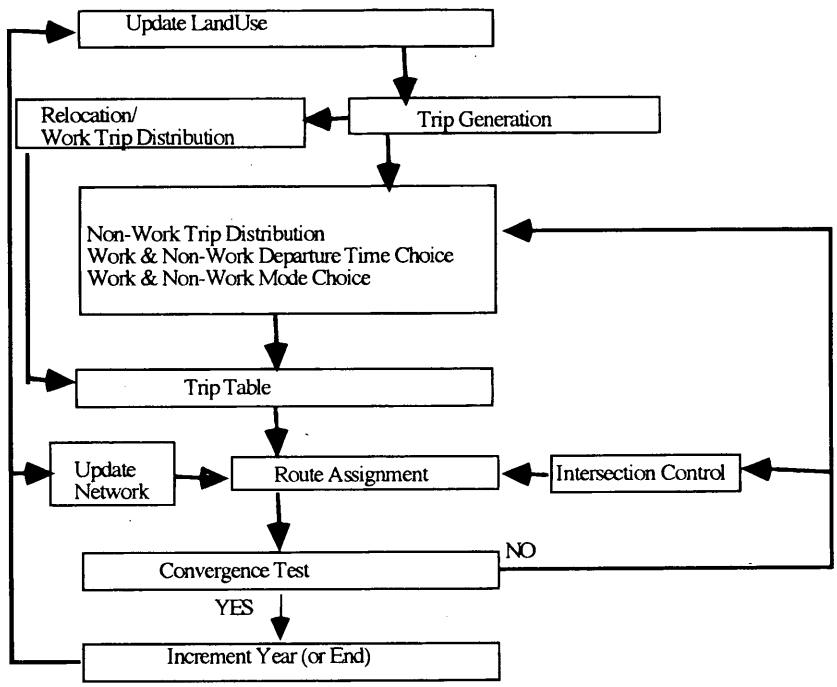


FIGURE 2 Flow chart hybrid evolutionary model.

based on the number of employees (office, retail, industrial, other). The purposes used in the model are work to home, work to other (to home), home to other, other to home, other to other (and home to work. Trip generation is computed for the afternoon peak period (3:30 to 6:30 p.m.). In the hybrid model discussed in this paper, both work to home and work to other (to home) purposes are considered "work trips"; the other purposes are considered "nonwork." Because this is a person-trip model, mode choice is estimated for both work and nonwork trips and all modes (including nonmotorized). Future research should derive trip generation from an activity approach considering activity frequency, duration, and scheduling.

Destination Choice

A multimodal trip distribution model is used in this model (*8*). A composite impedance calculated as the weighted average of the mode-specific impedances is computed, using mode shares as the weight. In the hybrid model, for nonwork trips, destination choice is computed in equilibrium with route assignment and intersection control. For all relocated and new work trips, the final travel times from the previous year are used to compute the trip distribution in the subsequent year. Other trips are carried from the previous year. Detailed information on the estimation of the initial (seed) trip table is available from the author and was not included for reasons of space.

Departure Time Choice

Departure time choice determines the proportion of peak-period vehicle trips that occur in the peak hour. It is a binomial logit model with two choices: peak hour and not peak hour. The factor that is used to determine probability of peak hour is the ratio of congested to free-flow time on a zone-interchange basis. This component is solved in equilibrium with route assignment and intersection control for both work and nonwork trips.

Mode Choice

In this application of the model, mode choice is held fixed at 1990 levels. Earlier tests of the model found little differentiation of mode choice because of the changes in network and land use between 1990 and 2000 when policies are kept fixed. In theory, this component could be solved in equilibrium with route assignment and intersection control. However, to reduce computational time and possible sources of minor variation, the zone-to-zone mode shares were therefore kept constant. Future research should consider a simultaneous approach to mode and departure time choice, and possibly destination choice, at least for nonwork activities, although various questions about the relative timing of these components would need to be resolved.

Route Choice and Intersection Control

A single-class user equilibrium assignment model provided by the EMME/2 software is used in this application (*9*). This model considers both link delay and turn delay. The inputs to turn delay (cycle length, green time per phase) are computed with an external program each iteration of the automobile assignment, and the results are fed back into the turn penalty function (*6*).

EXOGENOUS MODEL INPUTS

Two key sets of exogenous data are used in the model: land use and demographic changes by zone, and modifications in the highway and transit networks. These are described as follows.

Land Use and Demographics

The land use assumptions in this application are derived from the Round IV forecasts of the Metropolitan Washington Council of Governments and the Round IV forecast of the Baltimore Regional Council of Governments (*10,11*). In 1990, for Montgomery County, Maryland, the focus of this study, there were 280,000 housing units and 460,000 jobs, which is expected to increase by the year 2000 to 320,000 housing units and 580,000 jobs (Figure 3). These forecasts are based in large part on approved but unbuilt development (typically a 6- to 12-year inventory) and by the queue of developers who are applying for development approval. Future land use forecasts will incorporate estimates of transportation accessibility explicitly, and perhaps eventually the forecasting will be integrated. However, as noted earlier, resistance to combined transportation/land use forecasting in the Washington area is at least as political as technical. Demographic inputs (age distribution by area, average household size) are updated each year on the basis of results from an exogenous demographic forecasting process independent of any transportation variables.

Networks

A dynamic model requires that changes to the transportation network be coded to the year of change. Here the model transportation networks come from the Montgomery County Planning Department (for Montgomery County), the Metropolitan Washington Council of Governments (for the rest of metropolitan Washington) and the Baltimore Regional Council of Governments (for metropolitan Baltimore). The future network within Montgomery County has coded changes in link capacities (number of lanes) as well as additional links to the year of opening. Outside Montgomery County, the change in networks occurs for the base year and 1995. Thus, the capacity outside Montgomery County from 1990 to 1995 and from 1996 to 2000 is fixed.

MODEL CONVERGENCE

Figures 4 and 5 show convergence results for the two models, both in the year 2000 time horizon summarizing the entire model region. The results for the equilibrium model represents the value of the objective function on each iteration in the year 2000. The results for the hybrid-evolution model reflect the decisions decided in equilibrium (nonwork trip distribution, time-of-day choice, and route choice) also in the year 2000 for each iteration. In the evolutionary model, there is no convergence from year to year (as discussed in the next section on results). Figure 4 shows the total vehicles on the network, which for both runs converges to about 1 million vehicles by the 30th iteration. The equilibrium model has somewhat more vehicles than the hybrid-evolution model, although more research will be necessary to say whether this is inherent in the model structure or just an artifact of the particular data set. It should be noted that the hybrid-evolutionary model converges more quickly than the equilibrium model, probably because one major component, work trip distribution, is fixed before the model is run for a given year. By the 10th iteration the hybrid model has a demand that is substantially identical to the 30th iteration; however it takes 15 iterations for the same to be true of the equilibrium model.

Figure 5 shows the convergence of the objective function (absolute gap) for the two models. The gap is an estimate provided by the EMME/2 software of the difference between the current assignment and a perfect equilibrium assignment in which all routes used for a given origin-destination (O-D) pair would have the same

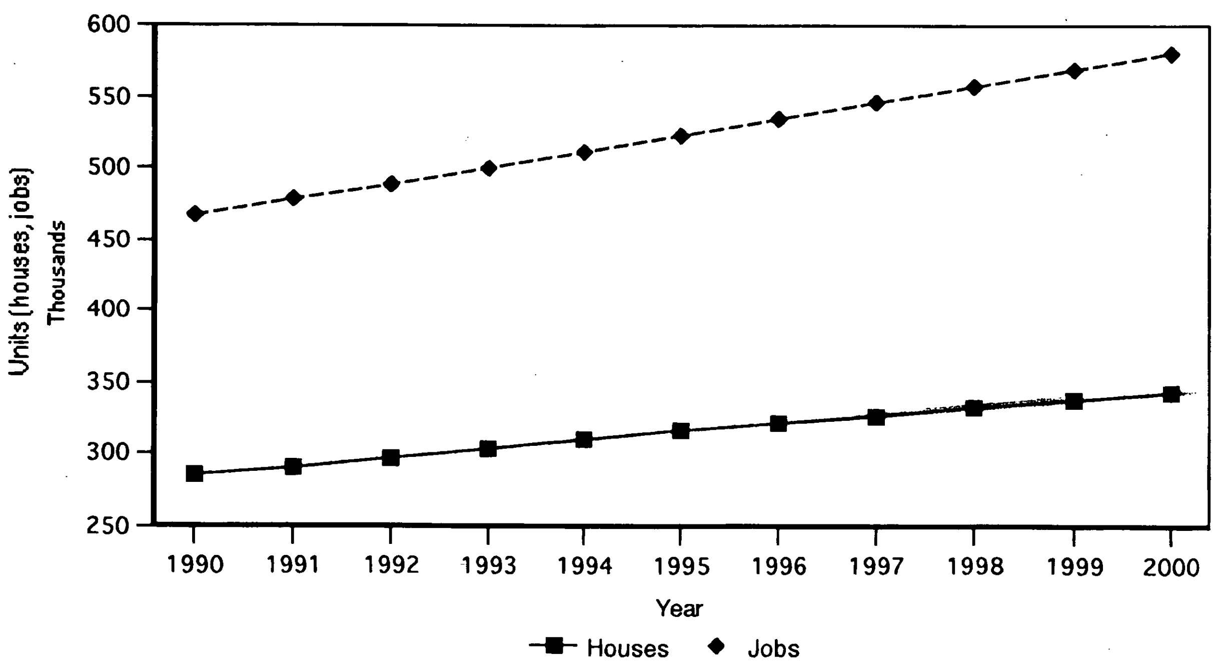


FIGURE 3 Land use activity, Montgomery County, 1990–2000.

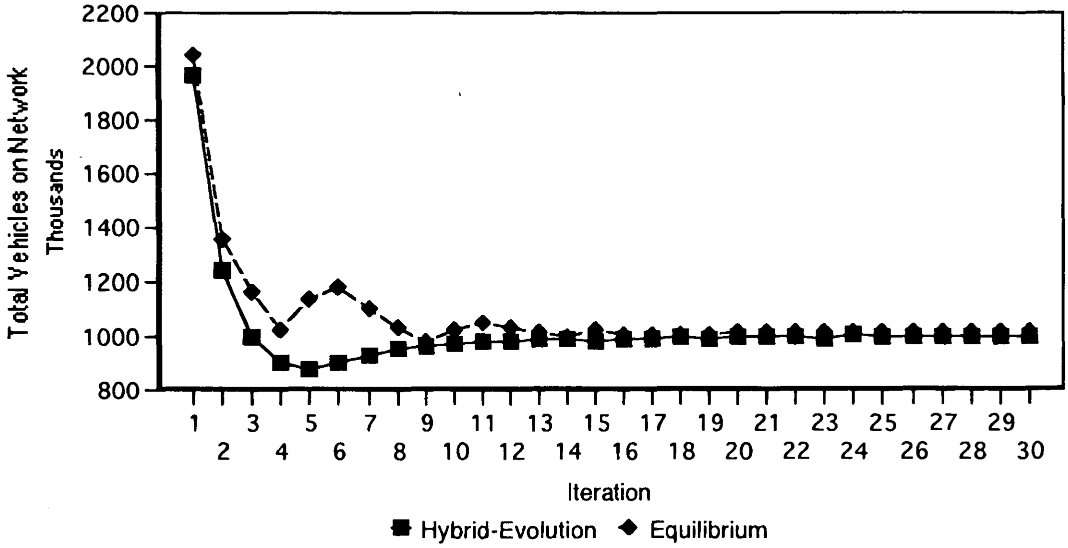


FIGURE 4 Convergence of models: total vehicles on network, 2000.

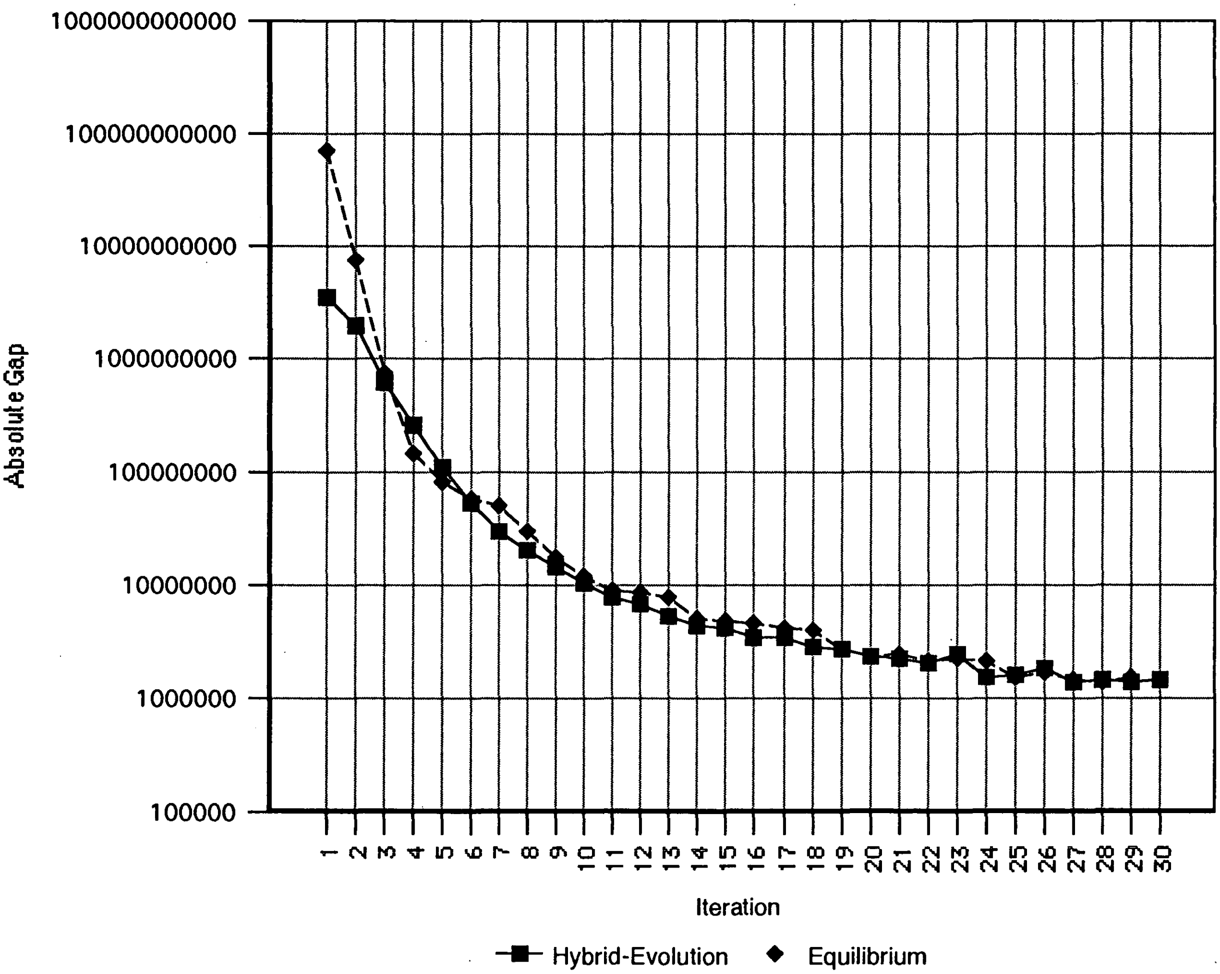


FIGURE 5 Convergence of models: absolute gap, 2000.

length (9). The value is tending to level out at about 1 million by the 25th iteration.

Thus, for any given time slice (1 year), the hybrid-evolutionary model reaches an equilibrium, although over time the equilibrium point moves. This particular structure, which is largely composed of convergent negative feedback loops is unlikely to have the potential for chaos, cascades, or catastrophes. However, depending on the rate of change of exogenous variables such as the network description or amount of development, the equilibrium point should move more or less smoothly.

RESULTS

Some summary figures are provided for the various models to compare their results. Figure 6 shows the peak hour vehicle trips (the same result as in Figure 4) for each year, again for the entire model region. The number of trips increases in the hybrid model, but is less than that in the equilibrium model. The large uptick in 1995 is caused by the increased network capacity, which was coded to come on line during the year (recall that outside Montgomery County, capacity from 1990 to 1994 is the same as it is from 1995 to 2000.) This clearly emphasizes the need for time coding of networks if this approach is to be used.

Figure 7 shows the vehicle miles traveled (VMT) within Montgomery County for the two models. Again the hybrid model is somewhat less traveled than the equilibrium model. VMT shows a sharp increase from 1990 to 1991. In that year I-270 was widened from 6 to 12 lanes through much of the county, which resulted in increased demand. Otherwise the growth is fairly smooth.

Figure 8 displays the average work trip time (in minutes), length (in kilometers), and speed (in kilometers per hour) for Montgomery County work trip origins (because this is the afternoon, origins are Montgomery County workers going home). All three values are stable across the decade, indicating that the feedback process is maintaining these attributes. In fact, speed improves over this period while travel time decreases slightly, indicating appropriate capacity increases and shifting travel patterns from suburb to suburb trips, which have higher average speeds. The difference of means tests performed over the 10 years, comparing the mean traffic zone time and speed (comparing 1990 and 2000 results f or the model) shows that the results for the year 2000 are statistically different from those in 1990 for the hybrid-evolutionary model for time and speed but the same for length. For the equilibrium model, the time, speed, and length did not show a statistical difference.

Figure 9 shows the proportion of Montgomery County links in each level of service category (LOS A through F). No trend is apparent. In fact, for the year 2000, the percentage of links better than LOS C/D is identical in both the hybrid and equilibrium models. Figure 10 shows the intersection LOS (using the critical lane volume method) for intersections in the county. Again no trend is

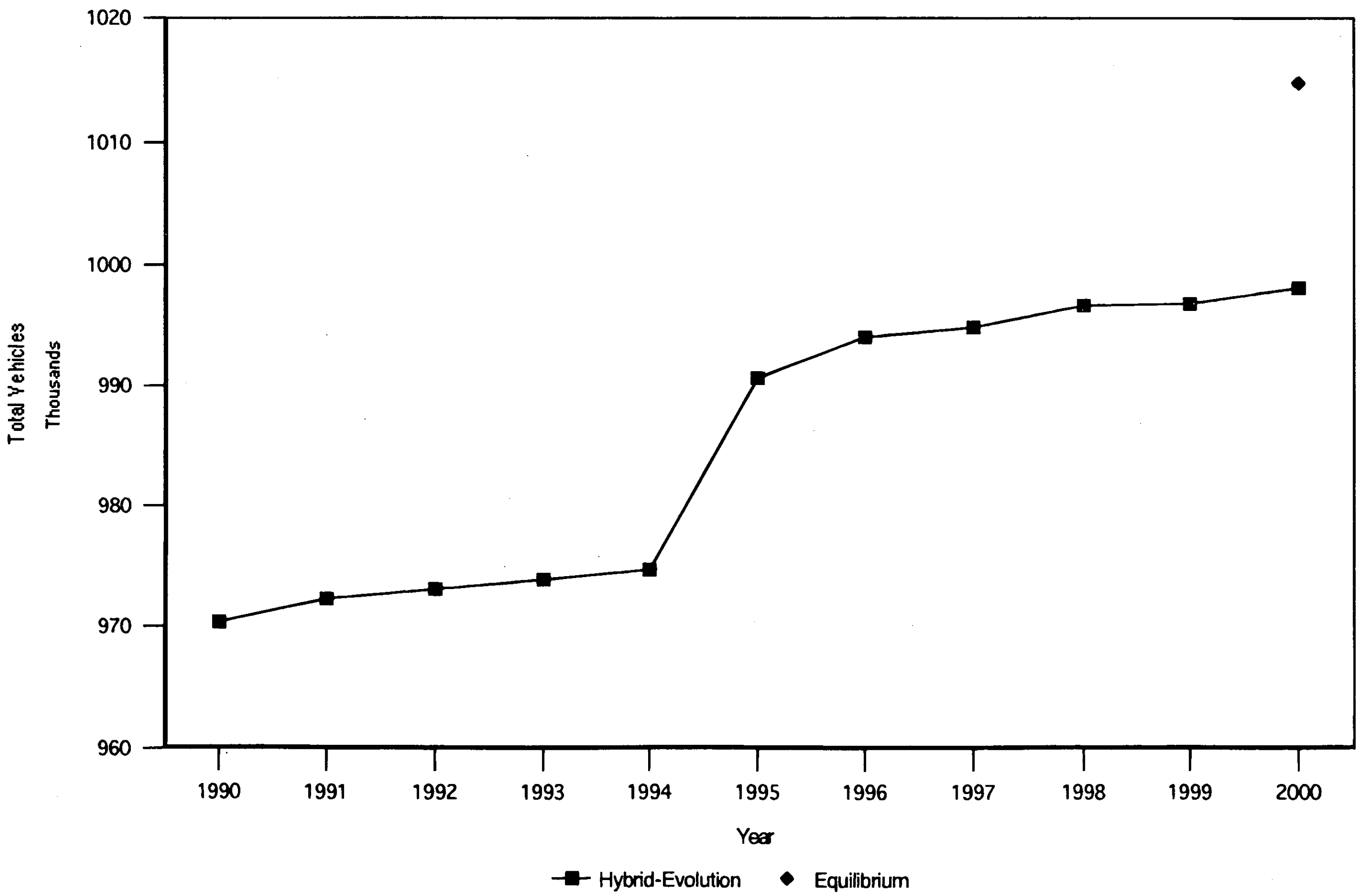


FIGURE 6 Peak-hour vehicle trips by year, entire model region.

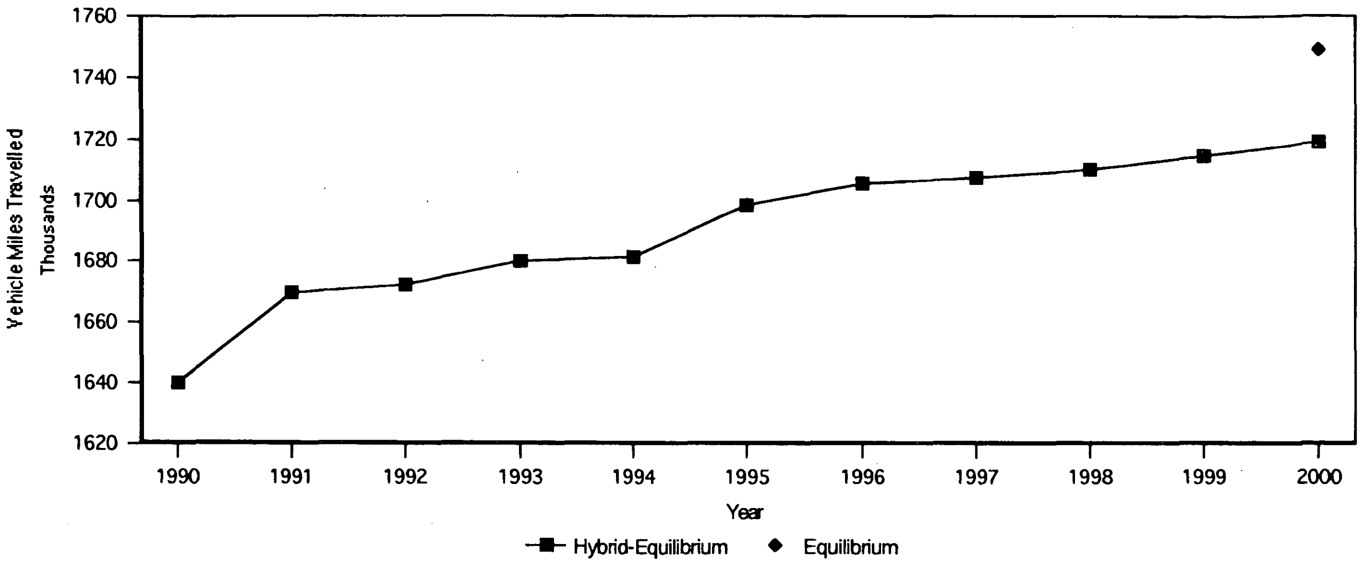


FIGURE 7 Vehicle miles traveled, Montgomery County links.

apparent, and the number of intersections above LOS C/D in both models is the same in the year 2000.

CONCLUSIONS

This paper discusses some of the implications of introducing dynamic work trip demand into the transportation planning model. As a behavioral assumption for the forecasting of a specific year in the midterm, evolution is conceptually better than equilibrium. The results were similar, but not identical between the two models. The length of the time period under study and the relative change in input data may influence model results.

The question of equilibrium or evolution is important in the context of attempts to construct dynamic models of urban structure and growth or travel demand. Most such large-scale models are now static, or dynamic in only the crudest sense, using 5-year time slices (*12*). However, the structure and function of every city, and the behavior of individuals within that city, depend crucially on their mutual co-evolutionary history. Because cities and human activity

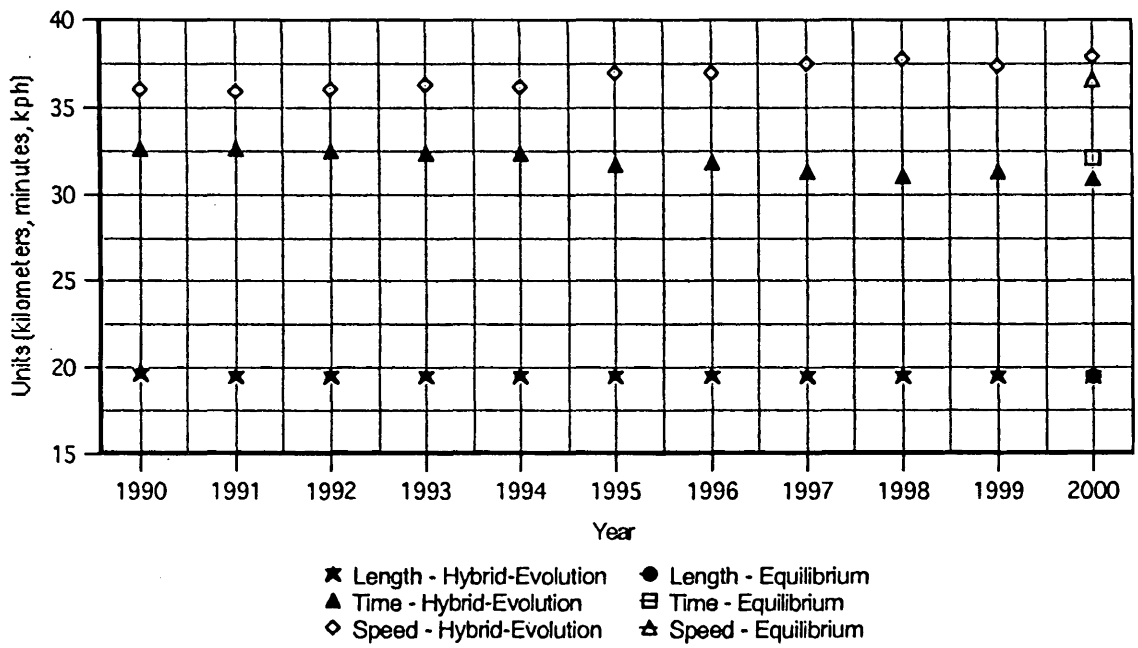


FIGURE 8 Evolutionary transportation planning model: comparison of time, length, and speed, Montgomery County work trip origins.

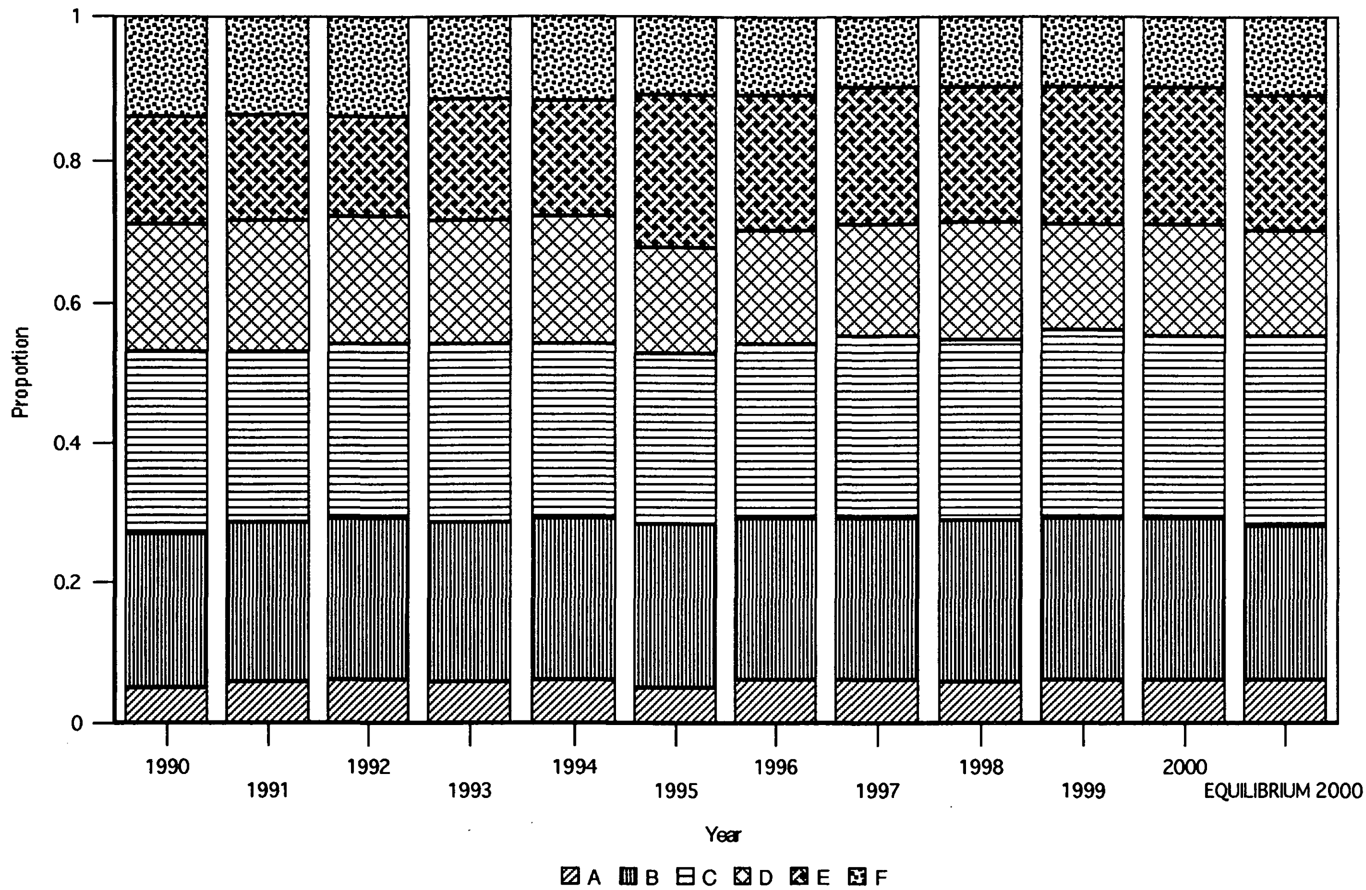


FIGURE 9 Link level of service, Montgomery County.

patterns evolve through time in complex, dynamic "environments," the interactions of urban form and human behavior do not, and should not be expected to, conform to equilibrium conditions. According to Forrester (*13,* p. 121), "The urban system is a complex interlocking network of positive and negative feedback loops. Equilibrium is a condition wherein growth in the positive loops has been arrested."

The reasons for preferring the evolutionary method to the equilibrium approach are several: (*a*) the ability to fully incorporate an observed data set such as a vehicle (or transit) trip table synthesized from traffic counts (*14*) (or transit ridership data) and the Journey to Work census data (unlike the use of equations and adjustment factors, all of the information inherent in the observed data can be used, and only the change over time needs to be modeled); (*b*) additional realism in the concept of the model; (*c*) the provision of a framework for extension to integration with land use models; and (*d*) the additional information available to policy makers for decisions such as the sequence of programming and constructing capital facilities, where the benefit depends on the timing of the facility.

This research points out the need to develop realistic behavioral models of switching in all model components. For instance, the Wardrop equilibrium principal states that no route is used between an O-D pair if the travel time is greater than on another route. But this implies perfect information. Once individuals have selected routes, their travel times change from day to day for a variety of factors. At what point does an individual decide to try another route? Under what conditions will this commuter stay with the second route or return to the first? How will advanced traveler information systems play into this? Switching is an issue in departure time choice, activity sequencing, mode choice, and nonwork trip destination selection (e.g., the choice of a grocery store). These and other questions will need to be answered as dynamic evolutionary modeling is implemented.

Some practical issues also emerge. There are not yet enough data to know the long-term temporal stability of this relocation value. What is it a function of? Are distribution curves (and other components) the same at the margins as they are on average? Further research can be aimed at implementing a full day-to-day evolutionary travel/activity demand simulation, with models of switching rather than attempting to predict the behavior of the entire population.

However, in the near term, application of supply/demand equilibrium models of travel demand is still preferable to conventional application with fixed zone-to-zone travel times independent of changes in the transportation network.

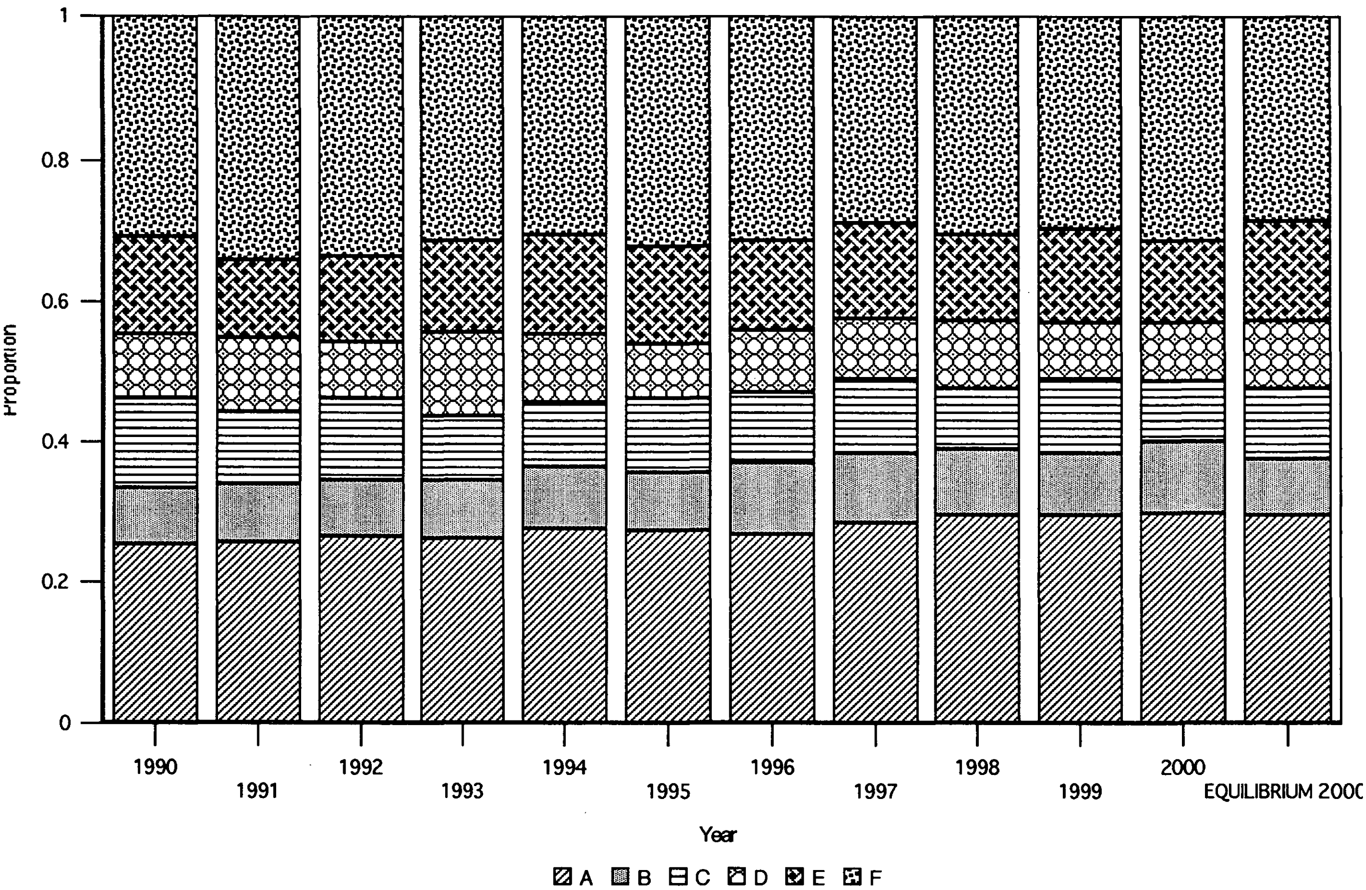


FIGURE 10 Intersection level of service, Montgomery County.

ACKNOWLEDGMENTS

The author thanks Chris Winters, Ajay Kumar, David Gillen, and three anonymous reviewers for their review of earlier drafts of this paper. The author also acknowledges the help of the University of California at Berkeley, and the Montgomery County, Maryland, Planning Department.

REFERENCES

1. Boyce, D. E., L. LeBlanc, and K. Chon. Network Equilibrium Models of Urban Location and Travel Choices: A Retrospective Survey. *Journal of Regional Science,* Vol. 28, No. 2, 1988.
2. Simon, H. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics,* Vol. 69, 1955, pp. 99–118.
3. Harris, B. Change and Equilibrium in the Urban System. In *Highway Research Record 309,* HRB, National Research Council, Washington, D.C., 1970, pp. 24–33.
4. Kumar, A., and M. Replogle. Low-Cost Travel Panel Survey. Presented at 1st U.S. Conference for Panels for Transportation Planning, Lake Arrowhead, Calif., Oct. 1992.
5. Kitamura, R. Sequential, History-Dependent Approach to Trip-Chaining Behavior. In *Transportation Research Record 944,* TRB, National Research Council, Washington, D.C., 1987, pp. 13–22.
6. Levinson, D., and A. Kumar. Integrating Feedback into the Transportation Planning Model: Structure and Application. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., 1994.
7. Levinson, D., and A. Kumar. Specifying, Estimating, and Validating a New Trip Generation Structure: A Case Study of Montgomery County Maryland. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., 1994.
8. Levinson. D., and A. Kumar. A Multimodal Trip Distribution Model: Structure and Application. In *Transportation Research Record 1466,* TRB, National Research Council, Washington, D.C., 1995, pp. 124–131.
9. *EMME/2 User's Manual: Release 6.* INRO, Montreal, Quebec, Canada, 1993.
10. *Round IV Cooperative Forecast.* Metropolitan Washington Council of Governments, Washington, D.C., 1989.
11. *Round IV Cooperative Forecast.* Metropolitan Baltimore Council of Governments, Baltimore, Md., 1991.
12. Wegener, M. Operational Urban Models: State of the Art. *Journal of the American Planning Association,* Vol. 60, No. 1, pp. 17–30.
13. Forrester, J. *Urban Dynamics.* MIT Press, Cambridge, Mass., 1969.
14. Spiess, H. *A Gradient Approach for the O-D Matrix Adjustment Problem.* Paper 692. Center for Transportation Research, University of Montreal, Quebec, Canada, 1990.

The opinions expressed in this paper are the responsibility of the author.

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

Exploring Route Choice Behavior Using Geographic Information System–Based Alternative Routes and Hypothetical Travel Time Information Input

MOHAMED A. ABDEL-ATY, RYUICHI KITAMURA, AND PAUL P. JOVANIS

A statistical analysis of commuters' route choice is presented. A binary logit model with normal mixing distribution using stated preference repeated-measurement data is estimated. General descriptive statistics are initially introduced in the paper to explore various route choice criteria and provide the basis for model estimation. The analysis is based on mail-out/mail-back surveys that were customized using routes generated by a geographic information system and information gathered from two previous route choice surveys. The results indicate the significance of travel time, travel time reliability, traffic safety, and roadway characteristics on route choice. The estimation results also underscore the influence of traffic information on route choice.

Fastest-path routing has been adopted over the years because of its simplicity and linkage with algorithms for generating equilibrium in static traffic assignment models. However, in real life, driver's routes are likely to deviate from the fastest path in significant ways. Empirical research on route choice behavior shows that drivers use numerous criteria in formulating a route: travel time, number of intersections, traffic safety, traffic lights, and other factors. Drivers' experiences, habits, cognitive limits, and other behavioral considerations may also produce variations in route selection. Viewed in this light, one can see that assuming travel time as the sole criterion of route choice is indeed an unrealistic abstraction of individual driver behavior and when aggregated at the network level may result in an inaccurate representation of traffic.

A number of studies have been performed in the past on route choice. Minimizing travel time is considered the most important criterion affecting drivers' route choice (*1–3*). Also, directness (*2*) and less congestion (*3*) were among the important reasons. Wachs (*3*) concluded that socioeconomic and demographic characteristics do not clearly relate to attitudes toward route choice criteria, whereas Jou and Mahmassani (*4*) and Mannering et al. (*5*) found that socioeconomic characteristics together with the traffic network were important determinants of route changing behavior.

An important factor that has been introduced frequently in the past few years is traffic information and its effect on commuters' behavior in general and on route choice in particular. Insights into drivers' route choice will help us understand the effect of information, which might also be a factor in new network-level traffic models.

This paper uses data collected from a route choice survey. The survey included two major components: a revealed preference (RP) section based on the attributes and perceptions of the respondents' primary (chosen) and geographic information system (GIS)-generated alternate routes and a stated preference (SP) section using repeated discrete choice scenarios. The main objective of the paper is to explore the criteria that influence commuters' route choice and to investigate the effect of advanced traveler information on route choice. The paper presents general descriptive statistics of commuters' route choice. A binary logit route choice model using stated preference data is also presented.

M. A. Abdel-Aty, Department of Civil and Environmental Engineering, University of Central Florida, P.O. Box 162450, Orlando, Fla. 32816. R. Kitamura and P.P. Jovanis, Institute of Transportation Studies, 2028 Academic Surge, University of California at Davis, Davis, Calif. 95616.

ROUTE CHOICE SURVEY

An ongoing effort for Partners for Advanced Transit and Highways (PATH) at the University of California, Davis, is to investigate the actual route choices of drivers, with the objective of developing refined route choice models that can include the effect of traveler information.

Two computer-aided telephone interviews (CATIs) were conducted in May 1992 and May 1993, respectively. These surveys investigated the actual routes used by commuters, their awareness of alternative routes, their attitudes and perceptions of several commute characteristics, and the traffic information they acquire and its effect on their route switching and choice. Several previous studies by the authors present the design of the CATI surveys and the data analyses (*6–8*).

A third route choice customized survey was developed targeting the respondents interviewed in the previous two CATI surveys. This survey contains the GIS-based RP route choice questions and the repeated discrete choice SP questions that evaluate the potential effect of advanced traveler information systems (ATIS). The survey was designed to obtain the following information:

- Route attributes considered important by the individual in the decision process that leads to the choice of a route;
- Commuter familiarity with highway and street networks and its potential effect on route choice;
- Commuter willingness to use ATIS; and
- Effect of advanced traffic information on route choice.

Response Rate

The number of targeted respondents was restricted by the availability of their addresses and the success in geocoding their home and work locations using a GIS. Home and work locations were successfully geocoded, and addresses were available (agreed to provide the address during the second CATI survey) for 263 respondents. The 263 questionnaires were customized according to each respondent's home and work locations, primary route, and travel time. The questionnaire included each respondent's primary route (from the CATI surveys), a GIS-generated minimum path route between the commuter's origin and destination, and SP choice scenarios customized using the commuter's primary route and actual travel time. The questionnaires were sent to the respondents along with a postage-paid return envelope and an incentive of $2.00. A total of 143 respondents completed and returned the questionnaires (54.4 percent response rate).

Survey Design

The revealed preference section and the stated preference section, both heavily customized for the respective respondents, are described.

Revealed Preference Section

The main objective of the revealed preference section is to understand why commuters choose a particular route (in this case their primary—or most frequently used—route); why they do not necessarily use the GIS-generated "optimal" route; how they perceive the primary and optimal routes; how familiar they are with the street and highway network; and how willing they are to use and accept the advice of an ATIS.

The primary commute route for each respondent is identified from the previous CATI surveys. Each segment of the primary route is presented to the respondent in a table; then the respondent is asked to rate a series of subjectively measured route attributes related to the primary route.

On the basis of each respondent's origin (home) and destination (work), and using GIS capabilities, the Navigation Technology's data bases are used to generate optimal routes. Navigation Technology's data bases are detailed data bases that include all the highways and streets in the study area. The optimal route is presented to the respondent in the questionnaire, followed by several questions that measure the respondent's familiarity with this route, willingness to use an ATIS, and rating of a series of route attributes. The RP data provide significant insights into the factors that influence route choice.

Stated Preference Section

The main objective of the SP section is to investigate the effect of ATIS together with roadway type, travel time, and familiarity with a particular route, on the route choice. SP methods become an attractive option in transportation research when revealed preference methods cannot be used in a direct way to evaluate the effect or demand for nonexisting services (e.g., ATIS). SP methods are easier to control, more flexible, and economical as each respondent may provide multiple observations for variations of the explanatory variables.

In this survey, each respondent is provided with three scenarios; in each, the respondent has to choose between two routes (Figure 1 shows an example of one of the scenarios). The choice is binary: Route 1 is customized for each respondent so that the SP design would be as realistic as possible, whereas Route 2 is hypothetical. For Route 1 it is stated: "Your primary route using. . . ." and then a segment of the respondent's actual route is written. The travel time of Route 1 is the respondent's actual commute time as stated in the CATI surveys, and the road type is the actual route type of the primary route (mainly freeway, mainly surface streets, or freeway and surface streets). The objective is to use the route that the respondent is familiar with and make the SP design realistic. Road type of Route 2 is one of the following: mainly freeway, mainly surface streets, or a combination of freeways and surface streets.

For the travel time on the alternative route to be as realistic as possible, and because both routes have the same origin and destination, the travel time on both routes is likely to be close to a great extent. Therefore, normal travel time on Route 2 is as follows:

0.9 * (normal travel time on Route 1)

1.0 * (normal travel time on Route 1)

1.1 * (normal travel time on Route 1)

Traffic information is available on either Route 1 or Route 2, but not both. If traffic information is available an estimation of the travel time on that day is one of the following:

0.9 * (normal travel time on the same route)

1.0 * (normal travel time on the same route)

1.1 * (normal travel time on the same route)

1.2 * (normal travel time on the same route)

1.4 * (normal travel time on the same route)

These values are chosen to be as realistic as possible to represent light and usual traffic conditions (factors of 0.9–1.1), mild traffic conditions (factor of 1.2), and heavy traffic conditions that might be caused because of, for example, an accident (factor of 1.4).

If the information system estimates an above-normal travel time, the cause of the delay is given to the respondent. The cause of the delay is either accident, maintenance, stalled vehicle, or regular congestion. ATIS were defined to the respondents as a system that can offer personalized information about a trip and give advice about other routes to take while considering current traffic conditions.

All possible combinations of the previous cases are considered, after excluding the obvious choices (e.g., if Route 1 is faster and has information that predicts no delays). In all, 68 different combinations were used, three for each respondent randomly.

FACTORS AFFECTING ROUTE CHOICE

As mentioned earlier, one of the main objectives of this study is to determine which route attributes are considered important by the individual in the decision process that leads to the choice of a route.

Respondents were asked to rank several factors that made them choose their primary route. The factors and the respondents' rankings are given in Table 1. Shorter travel time is the most important factor (first reason) for choosing the primary route (ranked as the

PART II

On the following 2 pages, we are asking you to choose from among two routes, the first is similar to your primary route, while the second is a hypothetical route.

Suppose one day you are choosing between
the following two routes from your home to work

	Route 1 Your primary route using HARVARD AVE	Route 2
1. Road type	Surface streets	Mainly Freeway
2. Normal Travel Time	15 minutes	13 minutes
3. Traffic Information		
• Estimated travel time on this day	Not available	18 minutes
• Information on the cause of the delay	—	Accident

24. Given these choices, which route would you choose on this particular day?

☐1 Route 1 ☐2 Route 2

25. When would you leave home on that day? ________ AM

FIGURE 1 Example of route choice question.

first reason by 40 percent of the respondents) followed by both travel time reliability (32 percent) and shorter distance (31 percent). About 62, 54, and 47 percent indicated that shorter travel time, travel time reliability, and shorter distance, respectively, as either the most or second important reason for choosing their primary route. Other reasons included fewer traffic signals, greater traffic safety, and lack of unsafe neighborhoods, which about 11, 6, and 4 percent of the respondents, respectively, considered the most important reason for route choice.

Table 2 indicates the factors that make respondents choose their primary route over the suggested optimal route (which was generated using a GIS system). Again, the results support the previous result that travel time minimization is the most significant factor. About 62.9 percent of the respondents indicated that they do not use the suggested optimal route because their primary route is faster. However, there exist other factors that enter into the decision to choose a particular route. Shorter distance, travel time reliability, and traffic safety were among the factors indicated by 37.8, 37.1, and 28.7 percent, respectively.

Other factors also enter into some individuals' decision to use a particular route. Number of roadway segments, freeway use, trip chaining, neighborhood security, and familiarity were among the factors less frequently stated. Overall, 10.5 percent of the respondents indicated that the suggested optimal route is the same as their primary route (they are already using the optimal route).

This result clearly shows that minimizing travel time is the primary reason for route choice, which conforms to many previous studies (*1–3*). This result also illustrates that minimizing travel time is not the only factor; there exist other important reasons, such as travel time reliability. Travel time reliability adds the measure of uncertainty to the route choice and introduces the significance of an information system that may help reduce travel time by selecting routes adaptively. In another paper by the authors in this Record, travel time variation was found to significantly affect route choice. Also, this result indicates that shortest-path criteria (either time or distance) solely are an unrealistic abstraction of individual driver behavior. It might be more realistic to include all the previous factors in determining drivers' route choice behavior and giving each factor a weight that represents its significance in the route choice.

Figure 2 indicates respondents' perception of their familiarity with the GIS-generated route (this measure might indicate the respondents' overall familiarity with their streets/highways network). The figure shows that a large majority of the respondents (73 percent) consider themselves "extremely familiar" with the suggested route, and 21.6 percent considered themselves "very familiar" with this route. The rest, about 5 percent, considered themselves "somewhat familiar" with the route. Only one respondent considered himself "not at all familiar." Also, about 54.3 percent of the respondents indicated that they had used the GIS-generated route before and 28.6 percent had used part of the route, whereas only 17 percent had not used this route.

The previous results indicate that the majority of the respondents are familiar to a large extent with their networks, which suggests that the commuters' unfamiliarity with alternative routes is not one of the main reasons that they choose a particular route; it is their perceptions of the attributes of a particular route, as discussed earlier (travel time, travel time reliability, distance, safety, etc.) that lead to a certain choice.

TABLE 1 Reasons for Choosing Primary Route

Reason for route choice	1st reason	2nd reason	3rd reason	4th reason	5th reason
Shorter travel time	58 (40.6%)	31 (21.7%)	11 (7.7%)	8 (5.6%)	4 (2.8%)
Travel time is reliable	46 (32.2%)	31 (21.7%)	21 (14.7%)	14 (9.8%)	8 (5.6%)
Shorter distance	45 (31.5%)	23 (16.1%)	17 (11.9%)	11 (7.7%)	4 (2.8%)
Fewer traffic signals	15 (10.5%)	15 (10.5%)	24 (16.8%)	24 (16.8%)	11 (7.7%)
Greater traffic safety	8 (5.6%)	15 (10.5%)	14 (9.8%)	17 (11.9%)	31 (21.7%)
No unsafe neighborhoods	5 (3.5%)	7 (4.9%)	9 (6.3%)	11 (7.7%)	19 (13.3%)
Drive more on carpool lanes	1 (0.7%)	3 (2.1%)	0	0	0

Note: Summing each column might exceed 100%, that is because some people chose two factors as 1st or 2nd reason, e.g., they consider shorter travel time and travel time reliability as the most important reason for route choice.

TABLE 2 Reasons for Not Using GIS-Generated Optimal Route

Reason	No. of respondents (percent)
Primary route is faster	90 (62.9%)
Primary route is shorter	54 (37.8%)
Travel time is unpredictable	53 (37.1%)
Primary route is safer	41 (28.7%)
Many short roadway segments	16 (11.2%)
Primary route involves more freeway segments	14 (9.8%)
Have to make stop on the way along the primary route	11 (7.7%)
Primary route does not include insecure neighborhoods	9 (6.3%)
Not completely familiar with this route	5 (3.5%)
Had a bad experience in the past with the suggested route	5 (3.5%)

Note: Multiple answers are allowed (respondents can choose more than one factor)

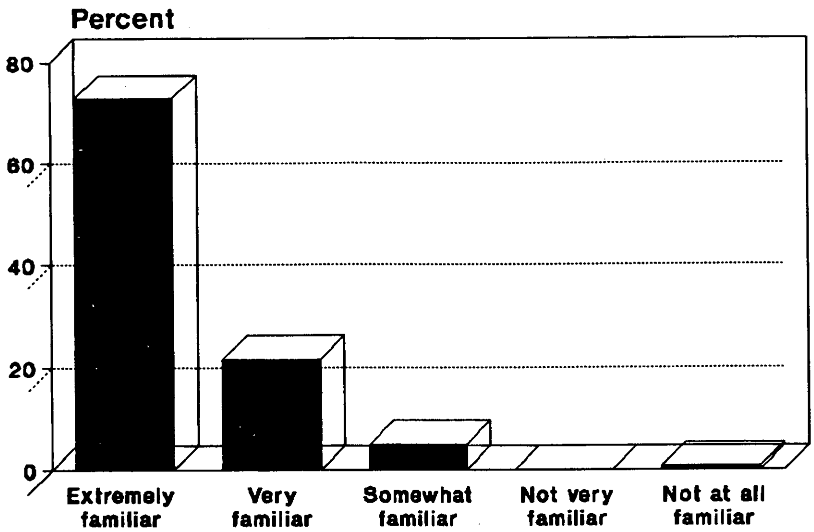


FIGURE 2 Respondents' familiarity with GIS-generated route.

ROUTE CHOICE MODEL USING STATED PREFERENCE DATA

Route choice models usually are estimated with observations of actual behavior, or RP data using discrete choice models (as presented in the previous section). However, hypothetical choice scenarios may be needed if the RP data do not provide information on preferences for nonexisting services (e.g., ATIS). In this study the SP scenarios are customized according to each respondent's route, roadway type, and travel time. This approach made the hypothetical choices realistic to a great extent, which is believed to be useful because a respondent's choices are more likely to represent the actual behavior.

As mentioned earlier, respondents were presented with three hypothetical scenarios. These scenarios are designed to investigate the respondents' route choice in the existence of ATIS.

The estimation of a logit model with repeated observations for each respondent gives rise to an obvious serial correlation of disturbances. This may be caused by heterogeneity, which refers to variations in unobserved contributing factors across behavioral units. If behavioral differences are largely caused by unobserved factors, and if unobserved factors are invariant over time but correlated with the measured explanatory variables, then estimates of model coefficients will be biased if this heterogeneity is not considered. Even without the correlation between the explanatory variables and unobserved factors, estimates of standard errors will be biased when the disturbances of a series of choices are serially correlated.

Methodological Approach

The approach taken in this paper to account for unobserved heterogeneity is to assume a parametric functional form for the pattern of the heterogeneity. The vector of observed choices or responses for individual i is defined as y_i. Each element of y_i is written as y_{it}:$t = 1, \ldots, T_i$ each of which is a repeated binary choice, expressed as the integers 0 and 1. The length of y_i is T_i, which may vary between individuals. The sample size is written as I, so $i = 1, \ldots, I$.

In the context of the short-term repeated choice sets data analyzed in this paper, it is possible to argue the existence of no state dependence (the utility of one period does not depend on choices of the previous periods) and stationarity (neither the variance of the error term nor the serial correlation between the error terms depend on time) (*9,10*). The probabilities that individual i chooses alternatives 0 and 1, p_{0it} and p_{1it}, respectively, are given as

$$p_{0it} = P(y_{it} = 0 \mid \alpha, \beta, x_{it}) = 1/[1 + \exp(x_{it}'\beta + \alpha)]$$
$$p_{1it} = P(y_{it} = 1 \mid \alpha, \beta, x_{it}) = \exp(x_{it}'\beta + \alpha)/[1 + \exp(x_{it}'\beta + \alpha)] \quad (1)$$

where

α = constant,
β = vector of parameters, and
x_{it} = vector of exogenous variables.

The influence of the unobserved variables in Equation 1 is represented by the constant term α; that is, the influence is assumed constant across individuals. The probability of observing $y_i = (y_{i1}, \ldots, y_{iT_i})$ given T_i in this specification is

$$P(y_i \mid \alpha, \beta, T_i, x_{it}) = \prod_{t=1}^{T_i} \left[\frac{\exp(x_{it}'\beta + \alpha)}{1 + \exp(x_{it}'\beta + \alpha)} \right]^{D_{1i}} \quad (2)$$

$$D_{it} = \begin{cases} 1 & \text{if } Y_{it} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Heterogeneity is introduced into the model by assuming that the probabilities p_{0it} and p_{1it} are conditional on both x_{it} and an individual specific error term, ξ_i, which represents all the other influences. Equation 1 becomes

$$p_{0it} = P(y_{it} = 0 \mid \beta, x_{it}, \xi_i) = 1/[1 + \exp(x_{it}'\beta + \alpha + \xi_i)]$$
$$p_{1it} = P(y_{it} = 1 \mid \beta, x_{it}, \xi_i)$$
$$= \exp(x_{it}'\beta + \alpha + \xi_i)/[1 + \exp(x_{it}'\beta + \alpha + \xi_i)] \quad (3)$$

The ξ_i: $i = 1, \ldots, I$ are assumed to be identically distributed with density function $f(\xi_i)$ independent of the x_i, so that Equation 2 becomes

$$p\,[y_i \mid \beta, T_i, x_{it}, f(\xi_i)] = \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} \left[\frac{\exp(x'_{it}\,\beta + \alpha + \xi_i)}{1 + \exp(x'_{it}\,\beta + \alpha + \xi_i)} \right]^{D_{it}} f(\xi_i)\, d(\xi_i) \quad (4)$$

This yields a marginal likelihood function. The unknown variables ξ_i are integrated out. Equation 4 is based on the assumption that ξ_i has a continuous distribution function. The distribution of ξ_i is called a mixing distribution. The log likelihood function is

$$L = \sum_{i=1}^{I} \ln \int_{-\infty}^{+\infty} \prod_{t=1}^{T_i} \left[\frac{\exp(x'_{it}\,\beta + \alpha + \xi_i)}{1 + \exp(x'_{it}\,\beta + \alpha + \xi_i)} \right]^{D_{it}} f(\xi_i)\, d(\xi_i) \quad (5)$$

A parametric form and $\xi_i \sim N(0, \sigma^2)$ are assumed. The integral is evaluated using Gaussian quadratures. General MLE packages such as the one provided with GAUSS statistical software (*11*) can be used to obtain maximum-likelihood estimates. The Broyden, Fletcher, Goldfarb, and Shanno (BFGS) optimization method is used in this study (*12*). The BFGS method is like the quasi-Newton method in that it uses both first and second derivative information. However, in BFGS the Hessian is approximated, reducing considerably the computational requirements, and although it takes more iterations than the quasi-Newton method it converges in less overall time.

Estimation Results

A binary logit model is developed using the methodology presented earlier. The model is developed to estimate the commuters' choice between Route 1 (customized according to the respondent's actual primary route and travel time) and Route 2 (a hypothetical alternative route). The overall observations are used to estimate the model, which gives a total of 417 observations (i.e., 139 respondents each making 3 choices).

The model is presented in Table 3. The insignificance of the constant term is a result of respondents' lack of knowledge of the attributes of Route 2 (respondents do not have complete information about Route 1 because it is customized according to their usual route). The model shows that as the percentage of "normal travel time on Route 2 to the normal travel time on Route 1" increases, the less likely the respondents are to choose Route 2. This variable shows that the respondents compare the travel time on both routes to make a route choice decision that minimizes their travel time.

The roadway type is also significant on route choice. If Route 2 involves freeway use then this route is more likely to be chosen, indicating the existence of freeway bias. The result also supports the results of the RP data, that is, commuters' preference for fewer different roadway segments on their route, which is probably the case with the use of freeways.

If the information system predicts a travel time on Route 2 that is less than the travel time on Route 1, then this increases the likelihood of Route 2 being chosen. This variable shows the importance of a travel information that provides travel time estimates.

Age was the only socioeconomic variable to enter into the model. Older respondents were found to be less likely to use Route 2, probably because this was considered an unfamiliar route. They apparently preferred to use their primary route and did not risk using an alternative.

Commuters' perception of the reliability of their actual commute route affects their choice. Respondents who perceive their actual commute route (Route 1) to have good or excellent travel time reliability were less likely to choose Route 2. This indicates their confidence in their route and consolidates the results from the descriptive statistics section.

The model shows that commute distance has a significant effect on route choice. The positive coefficient of the log of commute distance on the actual primary route (Route 1) indicates that respondents with longer distances tend to choose the alternative route (Route 2). This indicates that people with long commutes are more disposed to trying out an alternative route in an attempt to minimize their trip. The use of the log transformation indicates that this effect is nonlinear, with marginal increases in distance playing a stronger role in shorter commutes.

Finally, the significance of σ illustrates that the unobserved influences affecting a specific individual's choice are correlated from one of his selections to the next. This demonstrates the need to use a methodology, for example the normal mixing distribution in this study, to account for unobserved heterogeneity.

CONCLUSIONS

This paper is based on data collected from a route choice survey. The survey utilized innovative methods in studying route choice behavior by customizing mail-out/mail-back questionnaires. A GIS was used to generate optimal routes to understand drivers' familiarity with the highways/streets network and to study commuters' perceptions of this route in a way that helps identify the factors that influence route choice behavior. The survey included also a customized stated preference section that enables the investigation of the possible impact of ATIS on route choice.

The analysis showed clearly that minimizing travel time is the most important reason for choosing a commute route. About 40 percent of the respondents indicated that shorter travel time is their principal reason for choosing their primary route, and 63 percent indicated that they choose their primary route over the suggested optimal route because their primary route is faster.

However, minimizing travel time is not the sole reason for route choice. A large number of the respondents indicated the significance of other factors, such as travel time reliability, which illustrates the significance of the uncertainty measure in route choice and introduces the significance of an information system that reduces the level of uncertainty and helps commuters select routes adaptively. Other important factors that influence route choice are travel distance and the traffic safety on the chosen route.

Other factors appeared to enter into the route choice process, such as the number of traffic signals and stop signs and neighborhood security. These results suggest that route choice selection is a function of several factors, in which travel time would be assigned a heavy weight, and the other factors would contribute to the function according to the degree in which they influence the route choice. The survey results indicated that most respondents were familiar with the GIS-based alternative routes. Commute route choice appears to be a well-informed choice.

Modeling route choice asserted the significance of travel time on route choice, and showed clearly that ATIS has a great potential in influencing commuters' route choice even when advising a route different from the usual one. Several other commute factors were found to affect route choice, for example number of different roadway segments, freeway use, commute distance, and travel time reliability.

This paper illustrates clearly that several significant factors influence route choice, including advanced traffic information that provides

TABLE 3 Estimates of Stated Preference Route Choice Model with Normal Mixing Distribution and Gaussian Quadratures Estimation

		Coef.	t-stat.
	Constant	0.103	0.10
X_1	Normal travel time on route 2 / normal travel time on route 1	-1.585	-1.55
X_2	Freeway use dummy variable (1 if route 2 is mainly freeway or includes freeway, 0 otherwise)	0.412	3.18
X_3	ATIS dummy variable (1 if predicted travel time on route 2 < normal travel time on route 1, 0 otherwise)	1.204	3.37
X_4	Old age dummy variable (1 if > 55 years, 0 otherwise)	-0.545	-1.46
X_5	Travel time reliability on route 1 dummy variable (1 if travel time reliability on actual primary route is perceived to be good or excellent, 0 otherwise)	-0.552	-2.18
X_6	Log of commute distance in miles	0.631	1.97
σ	Standard Deviation of ξ_i	0.613	2.46

Summary Statistics
Log Likelihood at zero = -289.042
Log Likelihood at market share = -270.978
Log Likelihood at convergence = -192.961
Likelihood ratio index = 0.332
Number of observations = 417

Note: model coefficients are defined for route 2

travel time estimates for commuters. However, more work should be done to study the effect of factors that were not included in this study, along with drivers' experiences, habits, cognitive limits, and other behavioral considerations. An important factor that was raised in this paper that needs more investigation is the effect of travel time variation and uncertainty on route choice. Developing route choice models using the RP and SP data jointly also remains as a future task.

ACKNOWLEDGMENTS

The authors thank Navigation Technologies for providing the software and data bases used in developing the survey. Special thanks to the California Department of Transportation and PATH for funding this research.

REFERENCES

1. Duffell, J., and Kalombaris. Empirical Studies of Car Driver Route Choice in Hertfordshire. *Traffic Engineering and Control,* Vol. 29, No. 7/8, July/Aug. 1988.
2. Huchingson, R., R. McNees, and C. Dudek. Survey of Motorist Route-Selection Criteria. In *Transportation Research Record 643,* TRB, National Research Council, Washington, D.C., 1977, pp. 45–48.
3. Wachs, M. Relationships Between Drivers' Attitudes Toward Alternate Routes and Driver and Route Characteristics. In *Highway Research Record 197,* HRB, National Research Council, Washington D.C., 1967, pp. 70–87.
4. Jou, R., and H. Mahmassani. Comparability and Transferability of Commuter Behavior Characteristics Between Cities: Departure Time and Route Switching Decisions. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1994.
5. Mannering, F., S. Kim, W. Barfield, and L. Ng. Statistical Analysis of Commuters' Route, Mode and Departure Time Flexibility and the Influence of Traffic Information. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1994.
6. Abdel-Aty, M., K. Vaughn, R. Kitamura, P. Jovanis, and F. Mannering. Models of Commuters' Information Use and Route Choice: Initial Results Based on a Southern California Commuter Route Choice Survey. In *Transportation Research Record 1453,* TRB, National Research Council, Washington, D.C., 1994.
7. Abdel-Aty, M., K. Vaughn, P. Jovanis, R. Kitamura, and F. Mannering. Impact of Traffic Information on Commuters' Behavior: Empirical Results from Southern California and Their Implications for ATIS. *Proc. 4th Annual Meeting of IVHS America,* Atlanta, Ga., 1994.
8. Abdel-Aty, M., R. Kitamura, P. Jovanis, and K. Vaughn. *Understanding Commuters' Attitudes, Uncertainties, and Decision-Making and Their Implications for Route Choice.* Research Report UCD-ITS-RR-94-5. Institute of Transportation Studies, University of California, Davis, 1994.
9. Dunn, R., S. Reader, and N. Wrigley. A Nonparametric Approach to the Incorporation of Heterogeneity into Repeated Polytomous Choice Models of Urban Shopping Behavior. *Transportation Research A,* Vol. 21A, No. 4/5, 1987, pp. 327–343.
10. Uncles, M. *Models of Consumer Shopping Behavior in Urban Areas.* Ph.D. dissertation. Department of Geography, University of Bristol, Bristol, England, 1985.
11. *GAUSS—Version 3.0 User's Manual.* Aptech Systems, Inc., Kent, Wash., 1992.
12. Luenberger, D. *Linear and Nonlinear Programming.* Addison-Wesley, Mass., 1984.

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

Introduction of Information Feedback Loop To Enhance Urban Transportation Modeling System

KYLE B. WINSLOW, ATHANASSIOS K. BLADIKAS, KENNETH J. HAUSMAN, AND LAZAR N. SPASOVIC

The Urban Transportation Modeling System (UTMS) is a methodology used to estimate travel demand in response to changes in land use patterns, roadway characteristics, and socioeconomic factors. This demand is measured by the volume of traffic that flows through a system of streets and highways. Through the use of traffic assignment software, parts of UTMS have become automated. One of the newest automated processes is the extraction of a subarea from a larger regional model. This extraction process is important to the local planner because it maintains a link from the regional model to the local model and allows the planner to extract an already distributed trip table rather than build one from scratch. This subarea extraction process, as practiced, is a one-way information flow. The regional model is calibrated and its information is passed down to the subarea model. It is suggested that an "information feedback loop" should be inserted into the process. The subarea model information is looped back to the regional model and used in the regional calibration. The enhanced procedure is applied to a northern New Jersey network. The results show that the new methodology improved the calibration of the regional model, particularly in the vicinity of the subarea focus model. This new methodology is the key to developing subarea focus models with properly distributed trip tables. In addition, the results are used to develop general conclusions about the applicability of the feedback process.

The Urban Transportation Modeling System (UTMS) is a set of procedures used by transportation planners to estimate travel demand in response to changes in land use, roadway characteristics, and socioeconomic factors. UTMS is commonly referred to as the "four-step modeling process": trip generation, trip distribution, modal split, and route assignment (*1*). The UTMS process historically has focused on the regional impact of major transportation improvements and significant changes in land use. The regional models that have been developed to address these issues generally include only freeways, expressways, and major arterials. Roads that primarily serve local traffic are not included. Because of the desired accuracy levels, as well as technological limitations, detailed network coding for traffic signals, traffic control devices, and interchange configurations are not considered. Individual zones may be neighborhoods or even as large as municipalities.

More recently, environmental concerns as well as changes in the legislative and policy areas have resulted in closer scrutiny and analysis of smaller areas within the regional models. The need to respond to these issues, coupled with the availability of microcomputer transportation planning software packages such as QRS-II (*1*), MINUTP (*2*), and TRANPLAN (*3*), has led to the development of local area models. Compared with the regional model, the focus of the local area model is on the roadways that serve local traffic. Detailed network coding, including interchanges and traffic control devices is generally included. Individual zones may represent a residential subdivision, or a major employment center in a suburban area, or even a single block in an urban area. The questions to be answered by the local model concern the impacts of local zoning changes, major and minor residential or commercial development, and transportation system improvements such as improved traffic signal coordination and local roadway widenings.

Regional and local area models are developed to respond to different questions and to address different issues. However, they do share a large common pool of information regarding the physical characteristics of the network, as well as the demand for travel. The ability to "share" information between regional and local models has traditionally been a one-way flow. Network and travel demand information from the regional model is extracted and used as part of the development of the local area model. This paper outlines an improved flow of information that enhances the extraction process and uses the information from the local area model to create an "information feedback loop" to improve the regional model. This improvement results in a benefit at both the regional and local levels. The enhanced process is applied to a case study in northern New Jersey. The results of the case study are used to develop general conclusions about the applicability of the feedback process.

BACKGROUND

At one time the development of local area models simply did not consider the impacts of any changes that occur outside the local area boundaries. It has now become evident that transportation planning on all levels is interconnected and that the "planning-in-a-box" method of local area model development is no longer acceptable. One way in which transportation planners have attempted to respond to this need is by expanding the capabilities of its computer models to include a new analysis tool called the "subarea focus model" or "subarea windowing." The subarea focus model is a technique of extracting a subset of a larger area for use in developing a local area model.

The subarea extraction process is straightforward. Given the graphic representation of a regional transportation network, defined as a set of links (highway, roads, etc.) and nodes (origins, destination, intersections, etc.), the user first defines the limits of the study

K. B. Winslow, Parsons Brinckerhoff-FG, Inc., 506 Carnegie Center Drive, Princeton, N.J. 08540. A. K. Bladikas and L.N. Spasovic, National Center for Transportation and Industrial Productivity, New Jersey Institute of Technology, Newark, N.J. 07102. K. J. Hausman, URS Consultants, Inc., One Penn Plaza, Suite 610, New York, N.Y. 10119.

area, or subarea, by drawing a cordon line around it. Then, each link that crosses the cordon line is specified. These cordon links become the external stations for the subarea region; the traffic volumes of these links will represent all travel demand originating from or destined to the world external to the subarea. Travel demand within the subarea is unaffected by the extraction process. The result of the extraction process is a highway network containing all of the information from the regional network (number of lanes, capacities, free-flow speeds, etc.) and a travel demand matrix, or trip table, for trips with origin or destination, or both within the subarea. The extracted highway network and travel demand matrix form the basis for the local area analysis. Local streets not included in the regional model may be added as well as more detailed link coding for interchanges and divided highways. Additional information for individual links with respect to traffic control devices, local speeds, and capacities may be added as well. The travel demand matrix may be subdivided to a finer zone structure to represent specific subdivisions or employment sites.

The benefits of the subarea extraction process are threefold. First, the local area model reflects changes external to the local area. These changes include land use patterns and traffic conditions on the regional level. Second, the local area model is developed in less time. The local planner can start with the extracted network and travel demand information rather than create these components from scratch. In addition, because of the utilization of the trip generation, trip distribution, and mode choice steps from the regional model, the need to conduct traffic counts and collect origin-destination data for through traffic (i.e., traffic that has neither origin nor destination within the local area) is minimized. Third and final, the local area model should have greater accuracy, because it reflects the calibration of the regional model.

EXISTING METHODOLOGY

The subarea extraction, or subarea windowing, process is a significant step in transportation planning applications because it provides a connection between regional and local area models. In the process however, the connection is in one direction only: information from the regional model is used to develop and improve the calibration of local area models. No information from the local area models is used to improve the regional model calibration. Furthermore, the calibration of the regional model may be significantly worse for an individual area than for the region as a whole. As a result, although the local area model may benefit from the calibration of the regional model, it may also contain any biases or errors inherent in the regional model.

The traditional subarea extraction methodology, within the UTMS context, is illustrated in Figure 1. As indicated in the figure, the calibration of the regional and local area models are discrete steps within the process and the flow of information is from the regional model to the local area model only. The process starts with the calibration of the regional model, typically through iterative application of regional area trip generation, trip distribution, and route assignment steps. Once the calibration of the regional model is complete, the local planner extracts the local highway network and travel demand volumes. Other information that may be extracted include population and employment estimates, trip generation equations, and existing traffic count data. This extracted information forms the basis of the local area model. The local area network is then adjusted to better reflect local conditions. The travel demand matrix is adjusted to match existing traffic counts. The calibration of the local area model is performed, again typically through iterative application of local area trip generation, trip distribution, and route assignment steps.

PROPOSED METHODOLOGY

The proposed enhancement to the subarea extraction process adds an information feedback loop. As mentioned earlier, the calibration of the regional model may be significantly worse or biased for an individual area than for the region as a whole. As a result, it may be problematic to calibrate an extracted local area model. The proposed enhancement alleviates this problem by incorporating improvements to the regional model as part of the local area calibration process. Information from the local area model is used, or looped, to improve the regional model calibration.

The proposed methodology is shown in Figure 2. In contrast to the existing methodology, the calibration of the two models is merged into a single step. In addition, information now flows from the local area model to improve the regional calibration. The proposed methodology also begins with the calibration of the regional model, typically through iterative application of regional area trip generation, trip distribution, and route assignment steps. However, in contrast to the traditional methodology, the calibration of the regional model is not considered complete before the subarea extraction. The calibration of the local area model is performed, and the results are used to improve the regional model calibration as well. This process, or loop, is repeated until the regional model is sufficiently calibrated in the vicinity of the subarea, as well as regionally. The remainder of this paper concentrates on the application of this enhanced process to a case study in Bergen County in northern New Jersey.

CASE STUDY

The New Jersey Department of Transportation (NJDOT) currently possesses two regional highway transportation models: the North Jersey model, which includes the northern 13 counties of the state and adjacent areas in New York and Pennsylvania and the South Jersey model, which includes the southern 6 counties of the state and adjacent areas in Pennsylvania and Delaware. The development and calibration of these models is an ongoing process. Changes in technology as well as improved information sources including the U.S. Bureau of the Census and the New Jersey Department of Labor and new telephone, mail, and interview origin-destination surveys have all contributed to improve accuracy and sophistication.

The North Jersey model covers an area of 13 counties and over 200 municipalities. The network includes most of the freeways, expressways, and major arterials in the northern portion of the state: it consists of 1,377 internal traffic analysis zones and 9,970 network links, representing more than 17,773 lane-km (11,055 lane-mi) of roads. The case study uses a previously extracted portion of the North Jersey model, the Northwest Bergen County model, or Northwest model, which is shown in Figure 3. This model covers an area of one-quarter of one county and 16 municipalities. The network consists of 210 internal traffic analysis zones and 1,629 network links, representing 730 lane-km (454 lane-mi) of roads. The Northwest model includes population and employment matrixes, trip production and attraction formulas, trip distribution methodology,

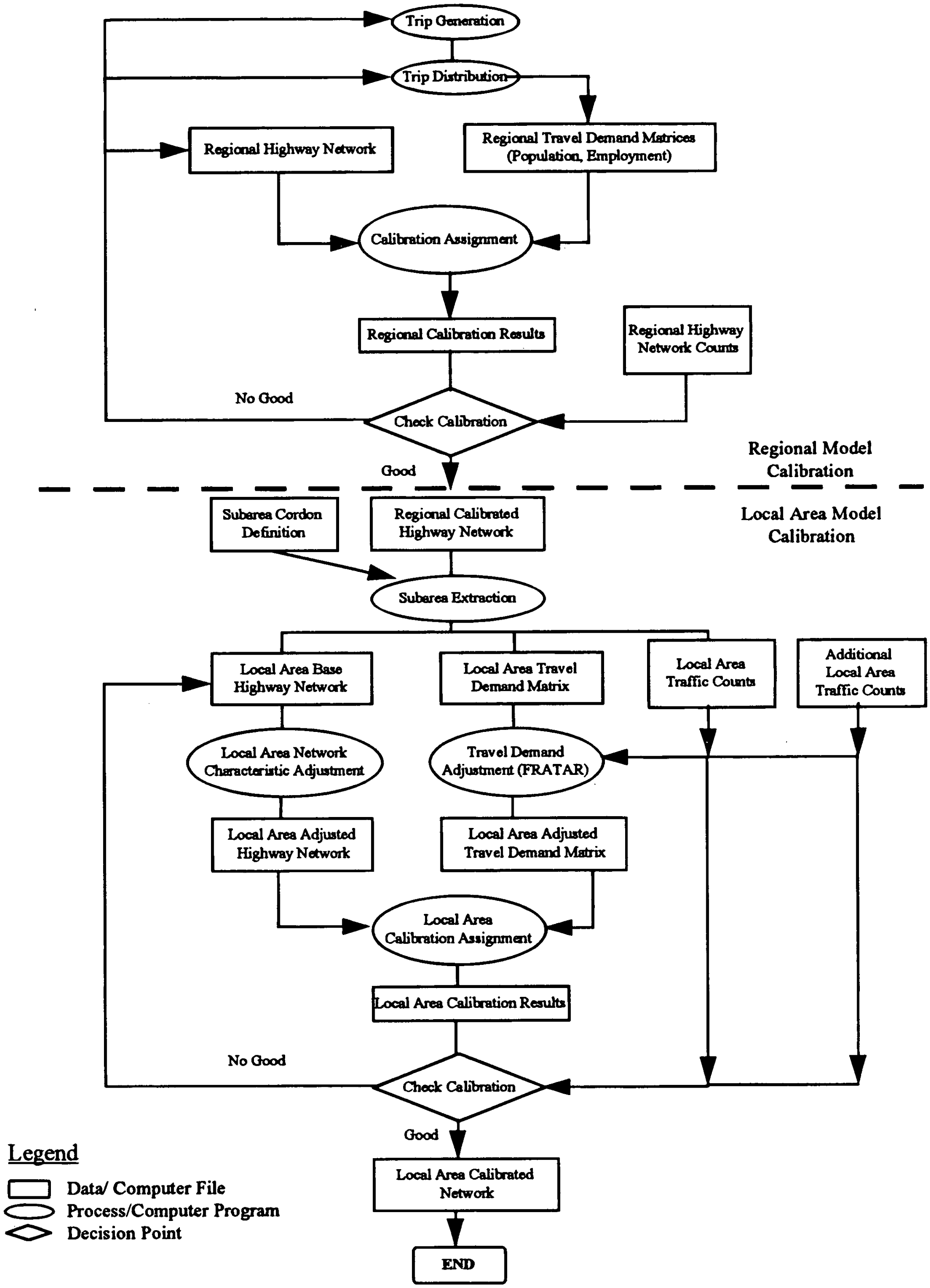


FIGURE 1 Traditional subarea extraction process.

existing travel demand matrix, roadway inventory by facility type and area type, free-flow speed and capacity information by facility type and area type, existing traffic counts, and a calibrated network.

For the case study, the northwest model had been calibrated previously as part of ongoing work being done by Bergen County. First, information is extracted from the northwest model to create the subarea model. This is accomplished through application of a route assignment with a defined subarea cordon. The output is a subarea highway network focused on the Route 4–Route 17 interchange, shown in Figure 4, and travel demand matrix.

Second, existing traffic counts to be used in the subarea model calibration are identified. To achieve good calibration of the subarea network, adequate local traffic count information must be available. As part of the subarea extraction process described above, traffic

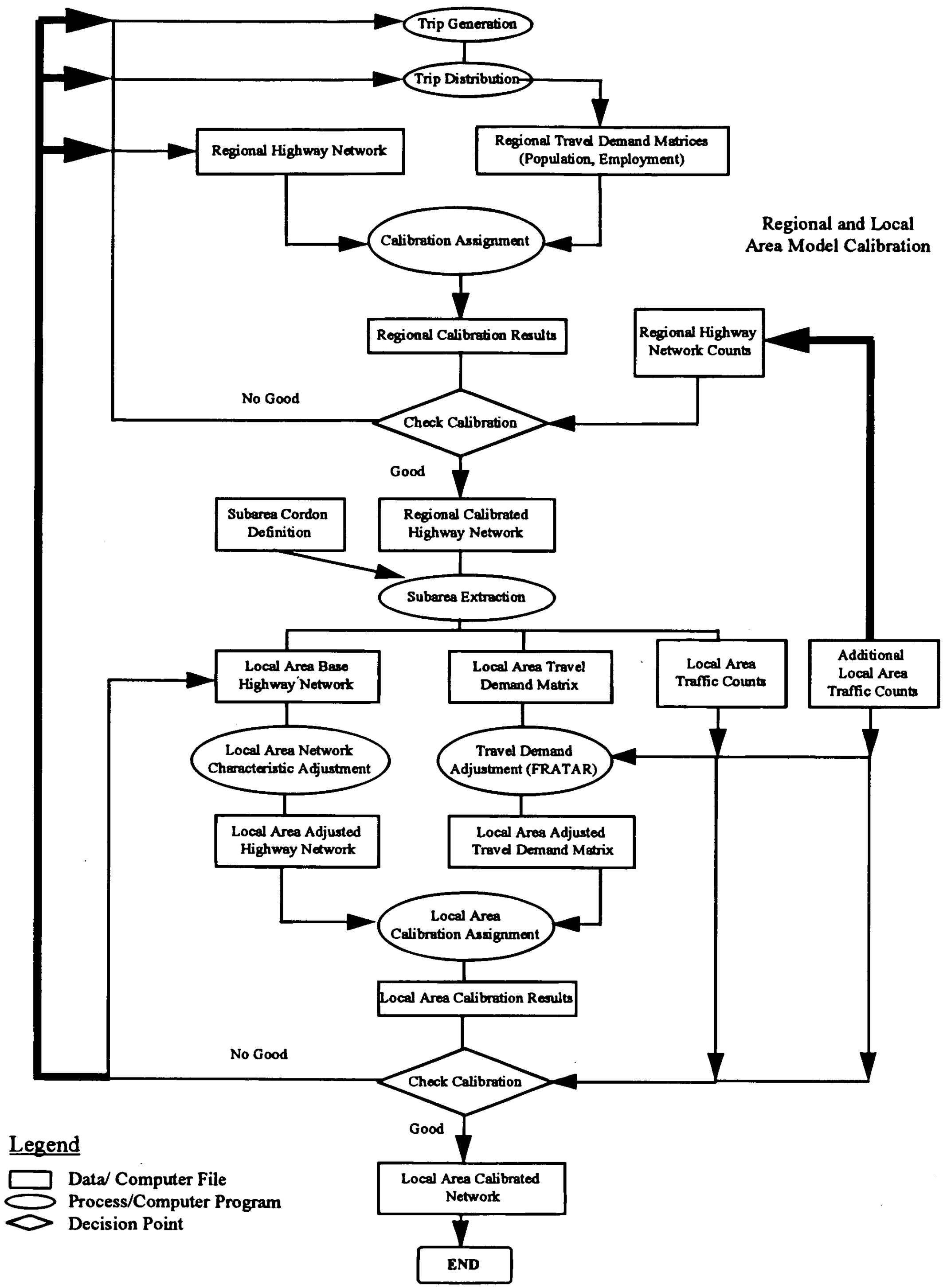


FIGURE 2 New subarea extraction process.

counts are extracted as part of the highway network. The locations of these extracted counts are noted in Figure 4. These counts alone are not sufficient to ensure good calibration because they do not include all cordon points, specifically the local road system. The subarea focus region must model the behavior of the outside world through the cordon points. Consequently, accurate traffic counts are required on all cordon points, especially in this case study because of its large "through-traffic" component, that is, traffic that neither begins nor ends within the subarea. The extracted traffic count data base is enhanced through conducting additional counts and by collecting information from local sources such as the municipal police departments and local traffic impact studies. The locations of these additional counts are also noted in Figure 4.

The traffic counts at each of the cordon points are then used as "target values" to adjust the extracted travel demand matrix. This adjustment process is typically done using the FRATAR process

FIGURE 3 Northwest Bergen County model network (*5*).

(*2*). FRATAR is a method used to adjust the trip distribution by iteratively applying factors to adjust origin and destination totals. Its shortcoming is that it is purely mathematical in nature and thus does not have a mechanism that allows it to account for network topology and performance. Hence, errors or bias in the regional trip distribution in the vicinity of the subarea would then be exacerbated. Herein lies one of the problems of the traditional methodology: it provides no ability to check the impact of the FRATAR method on trip generation or distribution. This problem is alleviated by providing a feedback loop to improve the regional calibration in the vicinity of the subarea before performing the FRATAR process.

Once the subarea travel demand matrix has been adjusted, free-flow speed and capacity adjustments are made to the extracted subarea highway network. Link speeds and capacities in the regional model are typically based on facility type (freeway, expressway, major arterial, minor arterial, etc.) and area type (central business district, urban, suburban, rural, etc.) only. This method of estimating speed and capacity is generally accurate for most links in a regional model. Consequently, it is not warranted to determine the impacts of geometric or physical attributes on each link in a regional model. However, attributes other than facility and area types do have a significant impact on both speed and capacity for individual links. For a local area model therefore, it may be warranted to identify individual links with extraordinary attributes. Consider two links representing an Interstate freeway in a suburban area. The first link is located several miles from adjacent interchanges; the second link is located in a weaving section between adjacent ramps of a major interchange. The base free-flow speed and capacity of both links would be similar; however, the effective speed and capacity of the second link is clearly significantly less. The attributes of links

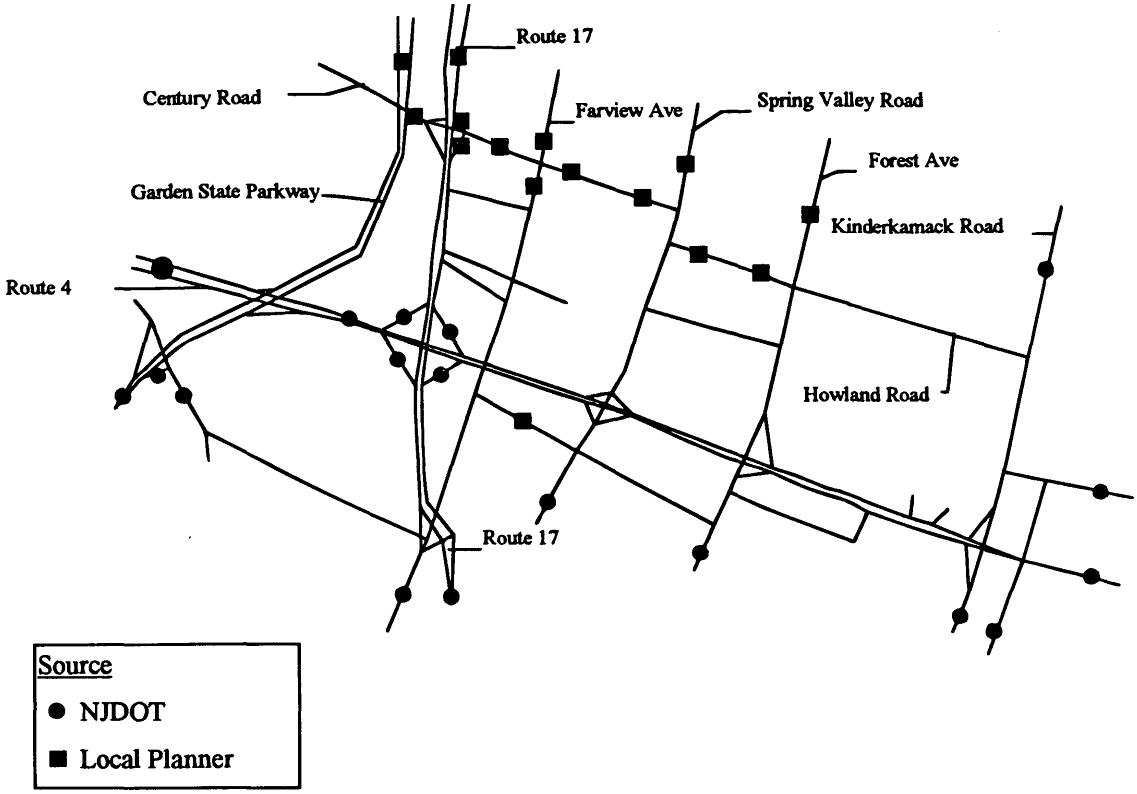


FIGURE 4 Subarea traffic count locations (*5*).

with the poorest calibration, generally the links with the lowest volumes, are adjusted first, whereas the best calibration, or highest volume links, is done last.

The subarea link attribute changes are used in the local area model calibration. Once the local area model calibration has been completed, these link changes are applied to the regional model. Also, the travel demand extraction and adjustment process may have uncovered errors—that is, too many, too few, or poorly distributed trips—in the subarea focus region. If these errors exist, they may be corrected by adjusting the trip generation or trip distribution in the regional model. The regional calibration process is then repeated using the improved information from the subarea model that is fed back to the regional model.

Once the revised regional calibration assignment has been done, statistics on two levels are checked. If the calibration is improved in the subarea focus region as well as for the regional model as a whole, the process continues with the subarea extraction and local area calibration. If the calibration improvement in the subarea focus region is at the expense of the regional model as a whole, the magnitude of the changes to the highway network or travel demand, or both, will need to be reduced. The process is an iterative one as the local planner seeks to improve the calibration at both the regional and local levels.

Finally, the subarea network and travel demand matrix are again extracted from the regional model. At this point, the revised travel demand distribution is compared with the initial extraction from the northwest model. Again, the FRATAR method is used to adjust the extracted trip table to match the traffic counts. However, the improvements to the regional model calibration in the vicinity of the subarea will result in an improved extracted trip table distribution. Consequently, the FRATAR process will have less impact on the subarea trip distribution. Enhancements to the local area network have already been incorporated in the regional model. Hence, no changes are required to the local area network. The local area assignment is then performed and the process is complete.

CASE STUDY RESULTS

To assess the success of the new methodology, calibration results of the case study are compared with the traditional method of subarea extraction. This is done for the whole Northwest model, and for the region of subarea focus—the Route 4–Route 17 subarea. The calibration is evaluated on both levels to check that better calibration in the subarea is not gained at the expense of calibration accuracy at the regional network level. The following are five ways in which the U.S. Department of Transportation (DOT) compares traffic assignment accuracy (i.e., model calibration) (*4*).

1. A comparison of total counted volume versus assigned volume across some aggregation, such as total study area or screenlines.
2. A comparison of total vehicle kilometers of travel from ground counts to vehicle kilometers of travel from the assignment results.
3. Developing a total weighted error between ground counts and assigned volumes.

TABLE 1 Performance Measures for Subarea Using Traditional Methodology

Volume Range		Number of	Total Volume		Difference		Square Error	
Lower	Upper	Records	Counted	Assigned	Assigned - Counted	Percent (%)	Root-Mean	Percent (%)
0	5,000	9	27,864	41,075	13,211	47.41	2,421	78.19
5,001	10,000	32	238,105	263,352	25,247	10.60	3,644	48.98
10,001	30,000	15	188,026	201,260	13,234	7.04	4,755	37.93
30,001	50,000	9	417,948	409,157	-8,791	-2.10	3,790	8.16
50,001	60,000	8	478,220	513,846	35,626	7.45	6,672	11.16
60,001	70,000	15	1,003,512	1,011,102	7,590	0.76	5,200	7.77
All Links		88	2,353,675	2,439,792	86,117	3.66	4,262	15.93

4. The calculation of the root-mean-square (*RMS*) errors comparing ground counts to assigned volumes by link within volume range stratification, such as

$$RMS = \sqrt{\frac{\sum_i (X_{gc} - X_{ta})^2}{N - 1}}$$

where

X_{gc} = ground count on link L_i,
X_{ta} = volume assigned on link L_i,
N = total number of links in observations group, and
i = index 1 through N.

The *RMS* error measures the deviation between two distributions—in this case counted and assigned link volumes. The percentage *RMS* error is derived by dividing the *RMS* error by the average group count for a particular group.

5. A graphic comparison of ground counts versus assigned volumes For this discussion, the Methods 1 and 4 are used as assignment calibration measures.

Using the new methodology at the Northwest model level, the planner realizes an improvement of 0.05 percent, or 3,321 vehicles (224,555 versus 221,234) in total counted versus total assigned volume. The *RMS* error improves by 13 vehicles (from 4,124 to 4,111), whereas the *RMS* percentage improves by .08 percent (28.28 percent versus 28.19 percent. Because of the minor nature of the network edits (20 out of 1,629 links) in the subarea region, one would not expect the calibration results to improve by much. But the fact that they do improve is enough to proceed with the comparison of the local area calibration results.

Table 1 presents the calibration statistics for the Northwest model in the region of the subarea focus as received from NJDOT. The absolute difference of total counted volume to total assigned volume is 86,117 vehicles, or 3.66 percent. The *RMS* error for the subarea focus region is 4,262 vehicles, whereas the *RMS* percentage is 15.93 percent. Table 2 indicates calibration statistics of the same network using the new methodology. The absolute difference of total counted volume to total assigned volume is 73,185 vehicles or 3.11 percent. The RMS error for the entire network is 4,144 vehicles, whereas the *RMS* percentage is 15.49 percent. Using the new methodology, the user has realized an improvement of 0.34 percent, or 12,932 vehicles in total counted versus total assigned volume. The *RMS* error has improved by 118 vehicles (from 4,262 to 4,144), whereas the *RMS* percentage has improved by 0.44 percent (from 15.93 to 15.49 percent). The improvements are relatively small, but by an order of magnitude greater than they were at the Northwest model level.

The third and most conclusive measure of validation of the new methodology is a comparison of the extracted subarea trip tables. Table 3 is a compressed district trip table for the traditional methodology. The 11 districts are represented in Figure 5. For this discussion, all internal zones are compressed into the first district because the subarea process does not affect them. This fact will be borne out in a comparison of the extracted trip tables.

Table 3 indicates that the total trips extracted for the subarea are 473,133. Table 4 is a compressed district trip table from the new methodology. It indicates that the total trips extracted for the subarea are 482,437, which is only 2 percent greater than the figure generated by the traditional methodology. However, the importance of the new methodology is seen in Table 5, which contains the differences between the two extracted trip tables and indicates that the distributions of each table are vastly different. As an example, the total number of trips destined to District 6 in Tables 3 and 4 is identical and equal to 9,316. However, an examination of Table 5 indicates that the origins of these trips are quite different. Using the new

TABLE 2 Performance Measures for Subarea Using New Methodology

Volume Range		Number of	Total Volume		Difference		Square Error	
Lower	Upper	Records	Counted	Assigned	Assigned - Counted	Percent (%)	Root-Mean	Percent (%)
0	5,000	9	27,864	25,824	-2,040	-7.32	1,447	46.75
5,001	10,000	32	238,105	231,941	-6,164	-2.59	3,092	41.56
10,001	30,000	15	188,026	206,182	18,156	9.66	4,897	39.06
30,001	50,000	9	417,948	411,087	-6,861	-1.64	2,843	6.12
50,001	60,000	8	478,220	536,782	58,562	12.25	7,574	12.67
60,001	70,000	15	1,003,512	1,015,044	11,532	1.15	5,201	7.77
All Links		88	2,353,675	2,426,860	73,185	3.11	4,144	15.49

TABLE 3 Extracted Subarea Trip Table Using Traditional Methodology

		Destination District											
		1	2	3	4	5	6	7	8	9	10	11	Total
O	1 Internal	10950	1818	442	10777	4204	2890	8540	7178	5729	7285	14371	74184
r	2 Century	1074	0	0	3134	575	299	319	483	956	833	0	7673
i	3 GSPNorth	1577	0	0	0	0	0	0	0	0	42592	863	45032
g	4 Rt17Nrth	7421	2827	0	209	1647	473	17521	5245	26848	72	5116	67379
i	5 ParamusW	4607	604	0	1750	172	156	1970	1378	1179	2196	1826	15838
n	6 ParamusE	3365	37	0	296	398	0	1494	2292	626	1007	1517	11032
	7 Rt4East+	7975	327	0	18320	1007	1172	2594	8703	807	3513	10873	55291
D	8 SthEast	6623	480	0	5300	863	1875	9147	1553	2368	3267	5150	36626
i	9 Rt17Sth	5975	1545	0	24520	1097	526	891	2467	0	638	7810	45469
s	10 GSPSth	7638	0	40050	0	1520	941	4112	3626	676	0	7255	65818
t	11 Rt4West+	11004	0	9816	0	2343	987	9762	4524	6282	4073	0	48791
	Total	68209	7638	50308	64306	13826	9319	56350	37449	45471	65476	54781	473133

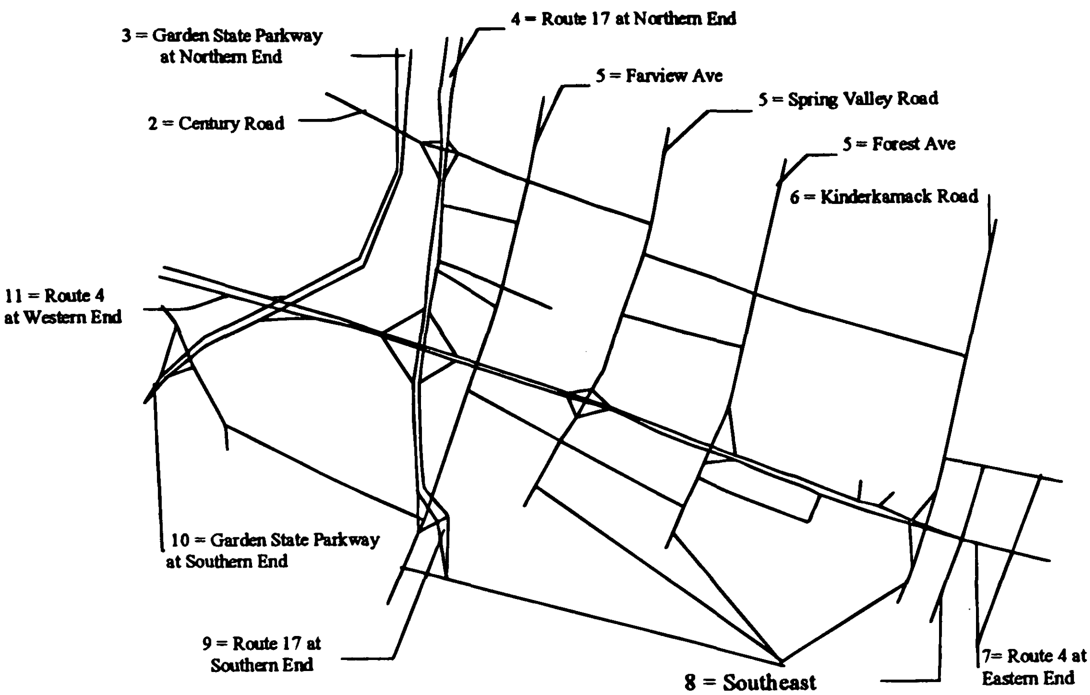


FIGURE 5 Subarea trip table reporting districts (*5*).

TABLE 4 Extracted Subarea Trip Table Using New Methodology

		Destination District											
		1	2	3	4	5	6	7	8	9	10	11	Total
O	1 Internal	9895	1707	442	10740	4473	2521	8242	6928	5529	7285	16422	74184
r	2 Century	1012	0	0	3134	572	21	319	483	956	833	0	7330
i	3 GSPNorth	1483	0	0	0	0	0	0	0	0	43242	844	45569
g	4 Rt17Nrth	7556	2827	0	209	1626	2	18206	5455	27070	72	5106	68129
i	5 ParamusW	5622	579	0	1568	202	896	2103	1535	1179	1546	1879	17109
n	6 ParamusE	1565	26	0	0	1549	0	1494	2292	626	1007	2473	11032
	7 Rt4East+	4451	327	0	18012	1128	1172	2594	8703	807	3513	14584	55291
D	8 SthEast	4679	480	0	5168	958	1875	9147	1553	2368	3267	7131	36626
i	9 Rt17Sth	5975	1545	0	24595	1097	526	891	2467	0	638	7735	45469
s	10 GSPSth	7638	0	40050	0	1520	941	4112	3626	676	0	7255	65818
t	11 Rt4West+	18333	0	9796	0	2404	1365	9242	4407	6260	4073	0	55880
	Total	68209	7491	50288	63426	15529	9319	56350	37449	45471	65476	63429	482437

TABLE 5 Trip Differences Between Extracted Subarea Trip Tables

		Destination District											
		1	2	3	4	5	6	7	8	9	10	11	Total
O	1 Internal	-1055	-111	0	-37	269	-369	-298	-250	-200	0	2051	0
r	2 Century	-62	0	0	0	-3	-278	0	0	0	0	0	-343
i	3 GSPNorth	-94	0	0	0	0	0	0	0	0	650	-19	537
g	4 Rt17Nrth	135	0	0	0	-21	-471	685	210	222	0	-10	750
i	5 ParamusW	1015	-25	0	-182	30	740	133	157	0	-650	53	1271
n	6 ParamusE	-1800	-11	0	-296	1151	0	0	0	0	0	956	0
	7 Rt4East+	-3524	0	0	-308	121	0	0	0	0	0	3711	0
D	8 SthEast	-1944	0	0	-132	95	0	0	0	0	0	1981	0
i	9 Rt17Sth	0	0	0	75	0	0	0	0	0	0	-75	0
s	10 GSPSth	0	0	0	0	0	0	0	0	0	0	0	0
t	11 Rt4West+	7329	0	-20	0	61	378	-520	-117	-22	0	0	7089
	Total	0	-147	-20	-880	1703	0	0	0	0	0	8648	9304

methodology, 740 vehicles have shifted from highways to local roads. A large number of these trips (471) shifted from Route 17.

If it is assumed that the regional trip table has a good trip distribution, and the calibration statistics indicate that the calibration in this area is improved, then it is safe to conclude that the distribution of the new methodology is superior. These comparisons support the claims that the new methodology is more sound.

CONCLUSIONS

The case study used to demonstrate the new methodology involves a regional model and a subarea of the regional model. It has been shown that by using the new methodology, improvement was realized in the calibration of the regional model, and trip distribution was improved in the vicinity of the subarea. This improved calibration process is the key to developing subarea focus models with properly distributed trip tables.

The authors believe that the new methodology with an information loop will work at all levels of the modeling process. The state DOTs in general, and NJDOT, in particular, could require that any transportation model that is funded or reviewed by the state DOT must have its basis on the DOT's statewide model. Planners would set up and collect data specific to their area and replace these new attributes back into the regional model, attempting to gain a better calibration for their specific area. Once this information has been processed by the local planner, the data can be channeled back to the state DOT. Modifications can then be made to the statewide modeling chain which would translate into new link attributes or new coefficients for production and attraction equations.

This new set of data, which is now tailored to the subarea region, would be incorporated into the modeling process. As the process continues, some of the realized benefits would be

- An enhanced statewide traffic count data base,
- Updates for the trip generation equation coefficients,
- Standardized data collection techniques,
- Incrementally better trip distributions,
- Standard statewide screenlines,
- Reduction in duplication of data collection,
- Improvements of calibration results at all levels,
- More efficient use of planning budgets, and
- Better dialogue between federal, state, and local officials.

Consistency and greater frequency between calibration updates would draw the modeling community closer to responding to changing issues in a reasonable length of time. This would eliminate the excuse of the model being "out of date." Better calibration logically yields better forecasts, and better forecasts provide planners with the needed insights to perform the land use and infrastructure planning process.

ACKNOWLEDGMENTS

This research is partially supported by a grant from DOT, University Transportation Centers Program, through the National Center for Transportation and Industrial Productivity at the New Jersey Institute of Technology. This support is gratefully acknowledged but implies no endorsement of the conclusion by these organizations.

REFERENCES

1. *NCHRP Report 187: Quick-Response Urban Travel Estimation Techniques and Transferable Parameters: User's Guide.* TRB, National Research Council, Washington, D.C., 1978.
2. *MINUTP: Technical User Manual,* Version 93-B. Comsis, Aug. 1993.
3. *TRANPLAN: User Manual,* Version 7.1. Urban Analysis Group, 1992.
4. *Traffic Assignment.* FHWA, U.S. Department of Transportation, 1973.
5. *Northwest Bergen County Land Use Model.* Garmen Associates. 1993.

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

Comparison of Alternative Methods for Updating Disaggregate Logit Mode Choice Models

DANIEL A. BADOE AND ERIC J. MILLER

An empirical assessment of alternative methods of updating disaggregate travel choice models so that their transferability from the estimation context within which they were originally developed to an application context (which differs from the original estimation context geographically or temporally, or both) is presented. The case study for the empirical tests performed is a long-term temporal transfer of work trip logit mode choice models estimated using 1964 data for the greater Toronto area (GTA) to represent 1986 work trip mode choice in the GTA. Three updating procedures that have been previously presented in the literature are examined (Bayesian updating, transfer scaling, and combined transfer estimation), plus a fourth new procedure, joint context estimation. All four procedures assume that a "small" data set of observed travel choices is available for the application context, which can be used in the updating procedure. The case study results indicate that the latter three procedures all possess merit as potential updating methods, with the choice among the three depending on such items as model specification and application context sample size. The results also indicate that if the application context sample size exceeds 400 to 500 observations, then updating may provide little or no improvement over simple estimation of an application context model, especially if "full" model specification is supported by the available data.

The spatial and temporal transferability of random utility models of travel demand is a matter of considerable practical interest. Although the empirical evidence concerning the transferability properties of random utility models is mixed (*1*), consensus exists that the potential for model transfer is greatly enhanced if local area (i.e., application context) data are used to update the model so that it better reflects application context conditions (*1–6*). At least two major reasons underlie this need to update transferred models (*4*):

1. Limitations in model specification, perhaps most notably as a result of omission of relevant variables; and
2. Differences in unmodeled "contextual factors" (geographical, historical, etc.) between the estimation and application context that affect the evolution over time of trip-makers' travel tastes and preferences.

Three major updating procedures have been presented in the literature to date:

1. Bayesian updating, in which parameter estimates from a small application context sample are combined with the estimation context parameter values using a classical Bayesian analysis to yield an updated set of parameters (*2*);
2. Transfer scaling, in which the application context utility function scales and alternative-specific constants are estimated from a small application context sample, assuming that the remainder of the utility function parameters are transferable from the estimation context (*3,4*); and
3. Combined transfer estimation, which can be viewed as a generalization of the Bayesian updating approach, which accounts for transfer scaling effects (*6*).

These approaches all assume that the estimation context model parameter values are known and that a small sample application context data set is available which permits the estimation of an application context model that is identical to the estimation context model being transferred. If, however, the estimation context data set used to estimate the original model parameters is also available (which in many instances may well be the case), a fourth approach is possible. This fourth approach, labeled joint *context estimation* involves estimating a new joint estimation/application context model, using both the estimation context and application context data sets.

This paper has two purposes. First, it provides a systematic comparison of the four updating techniques within a common empirical application. Second, this empirical application is unique in the literature because it involves assessing the relative effectiveness of the various updating procedures in achieving long-term temporal transferability of a disaggregate choice model within the same geographic area. Specifically the case study consists of updating 1964 morning peak-period work trip mode choice models developed for the greater Toronto area (GTA), Canada, over a 22-year period to reflect 1986 conditions.

The next section of this paper briefly reviews the four updating procedures. The paper's third section briefly describes the data sets used in the study, and the fourth section describes the test procedure employed. The fifth section presents and discusses the results obtained. The final section of the paper then summarizes the findings of the study and their implications for the state of practice in model updating and transfer.

MODEL UPDATING METHODS

It is assumed that a disaggregate multinomial logit choice model is to be transferred from an original (estimation) context to a new (application) context; that the estimation context parameter estimates are known; and that a small sample data set drawn from the application context that is suitable for estimating a model specified identically to the estimation context model is available.

Notation used throughout this discussion of methods includes the following:

D. A. Badoe, University of Toronto, Joint Program in Transportation, 42 St. George Street, Toronto, Ontario M5S 1A4, Canada. E. J. Miller, Department of Civil Engineering, University of Toronto, 35 St. George Street, Toronto, Ontario M5S 1A4, Canada.

$\mathbf{\Theta}$ = $[K \times 1]$ vector of utility function parameters, where $K = N + M - 1 = [^{\alpha}_{\beta}]$;
α = $[(N - 1) \times 1]$ vector of alternative-specific constants, where N is maximum number of alternatives available in choice set;
β = $[M \times 1]$ vector of parameters consisting of utility function weights for M explanatory variables in model;
$\mathbf{X}_{it}$ = $[M \times 1]$ vector of explanatory variables for alternative i for individual t;
V_{it} = systematic utility for alternative i for individual t $= \beta'\mathbf{X}_{it} + \alpha_i$ (where $\alpha_N = 0$ by definition); (1)
P_{it} = probability that individual t will choose alternative i from choice set C_t $= \exp(V_{it})/\Sigma_{j\in C_t}\exp(V_{jt})$; (2)
$\mathbf{\Theta}_1, \mathbf{\Theta}_2$ = estimates of $\mathbf{\Theta}$ derived from estimation context and application context data sets, respectively;
Σ_i = estimated parameter covariance matrix for context i ($i = 1,2$); and
$\mathbf{\Theta}_{\text{update}}$ = final estimates of $\mathbf{\Theta}$ to be used in application context, as generated by updating procedure.

Bayesian Updating

On the basis of the seminal work of Atherton and Ben-Akiva (*2*), it is well known in the literature that the Bayes theorem can be applied to the updating problem to yield asymptotically normal updated parameters with the following mean:

$$\mathbf{\Theta}_{\text{update}} = \left(\sum_1^{-1} + \sum_2^{-1}\right)^{-1}\left(\sum_1^{-1} \mathbf{\Theta}_1 + \sum_2^{-1} \mathbf{\Theta}_2\right) \tag{3}$$

and covariance matrix:

$$\sum_{\text{update}} = \left(\sum_1^{-1} + \sum_2^{-1}\right)^{-1} \tag{4}$$

Thus, to use this updating procedure, one must estimate an application context model using the available application context small sample using standard maximum likelihood methods to compute $\mathbf{\Theta}_2$ and Σ_2. These can then be combined with the known values of $\mathbf{\Theta}_1$ and Σ_1 from the estimation context using Equation 3 to yield the updated model parameters. Atherton and Ben-Akiva (*2*) used this procedure with considerable success to update a 1968 Washington, D.C., work trip mode choice model to reflect 1963 New Bedford, Massachusetts, and 1967 Los Angeles applications.

Transfer Scaling

It is well recognized that alternative-specific constants are likely not to be transferable between applications, given the extent to which systematic but unmodeled "contextual factors" are captured within these terms. It is equally true that the overall scale of the model's utility functions (which are not statistically identifiable within a standard cross-sectional model) are also likely to vary from one application to another, again because of unmodeled contextual factors.

A not unreasonable hypothesis on which to construct an updating procedure, therefore, is to assume that the utility function parameters computed in the estimation context, excluding the alternative-specific constants, are transferable to the application context, up to scale (note that, among other implications, this results in values of time—as defined by the ratios of time-to-cost parameters within the utility functions—being equal in the two contexts). Indeed, as shown elsewhere (*1,3,4*), much of the transfer bias can, in fact, be eliminated by adjusting model constants and scales. The updating problem then becomes one of determining changes in alternative-specific utility function constants and scales relative to estimation context values. For example, given a set of estimation context parameters (excluding alternative-specific constants), β_1, one can assume that the application context systematic utilities, $V_{it,2}$, take the following form:

$$V_{it,2} = \mu_{i,2}\beta_1'\mathbf{X}_{it,2} + \alpha_{i,2} \tag{5}$$

where $\mu_{i,2}$ is the ratio of the application context utility function scale to the (unidentified) estimation context utility function scale for alternative i, and all other terms are as previously defined, with the addition of the subscript 2 to indicate the application context.

Alternatively, Gunn et al. (*3*) apply scale factors to various groupings of parameters, where these groupings are defined on the basis of variable type rather than alternative. Equation 5 is thus a special case of the Gunn et al. formulation, which also includes as special cases complete reestimation of the model parameters on the basis of the application context data set (i.e., 1 μ for every parameter) and "naive" transfer of the estimation context model ($\mu = 1$).

The updated application context alternative-specific constants (α) and scale adjustments (μ) are readily estimated given an application context small sample using standard maximum likelihood methods, with the constructed variable $W_{it} = \beta_1'\mathbf{X}_{it,2}$ being the single explanatory variable in the utility function for each alternative.

Given that W_{it} is constructed using the estimated values β_1, the standard errors reported by typical logit model estimation packages will be biased downwards. If it is critical to the evaluation of the updating results to eliminate this bias, then appropriate corrections can be computed. More typically, the estimation results obtained will be sufficiently robust to allow the modeler to use the unadjusted standard errors, with the recognition that they somewhat overestimate the precision of the parameter estimates.

Successful applications of transfer scaling techniques include the following:

1. Gunn et al. (*3*), in which alternative transfer scaling schemes were applied to four different models: joint mode and destination choice models for personal business trips and shopping trips and trip frequency choice models for the same two trip purposes; with the transfer occurring between two regions in the Netherlands (Rotterdam/The Hague and Utrecht), and with the data sets for the two urbanized regions being collected 5 years apart and at different times of the year; and
2. Koppleman et al. (*4*), in which both intraregional transferability within the Washington, D.C., area and interregional transferability among the metropolitan areas of Washington, D.C.; Baltimore; and Minneapolis-St. Paul were investigated for the case of work trip mode choice models.

Combined Transfer Estimation

Implicit in the Bayesian updating approach is the assumption that $\mathbf{\Theta}_1 = \mathbf{\Theta}_2 = \mathbf{\Theta}$; that is, that the estimation and application contexts share the same underlying set of parameters. The transfer scaling method, on the other hand, explicitly assumes that a "transfer bias," $\mathbf{\Delta}$, exists, where

$$\boldsymbol{\Delta} = \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1 \tag{6}$$

Ben-Akiva and Bolduc (*6*) present a generalization of the Bayesian approach, which accounts for a nonzero $\boldsymbol{\Delta}$, and which yields the minimum mean square error estimate of $\boldsymbol{\Theta}_{\text{update}}$ achievable from a linear combination of the estimation and application context parameter estimates. As shown in Equation 6, this minimum mean square error estimate is provided by

$$\boldsymbol{\Theta}_{\text{update}} = [(\Sigma_1 + \boldsymbol{\Delta}\boldsymbol{\Delta}')^{-1} + \Sigma_2^{-1}]^{-1}[(\Sigma_1 + \boldsymbol{\Delta}\boldsymbol{\Delta}')^{-1}\boldsymbol{\Theta}_1 + \Sigma_2^{-1}\boldsymbol{\Theta}_2] \tag{7}$$

Comparison of Equation 7 with Equation 3 indicates that the combined transfer estimator reduces to the Bayesian estimator in the case of $\boldsymbol{\Delta} = 0$. In practice, the unknown transfer bias $\boldsymbol{\Delta}$ is approximated by the estimated bias $d = \boldsymbol{\Theta}_2 - \boldsymbol{\Theta}_1$. Ben-Akiva and Bolduc also demonstrate theoretically that the combined transfer estimator is superior to simply using the application context parameter estimates $\boldsymbol{\Theta}_2$, providing the transfer bias, $\boldsymbol{\Delta}$, is small. If the transfer bias $\boldsymbol{\Delta}$ is large, then the term $(\Sigma_1 + \boldsymbol{\Delta}\boldsymbol{\Delta}')^{-1}$ in Equation 7 becomes negligible and hence $\boldsymbol{\Theta}_{\text{update}} \approx \boldsymbol{\Theta}_2$.

Joint Context Estimation

The transfer scaling procedure described above for updating model constants and scales makes the following assumptions concerning the other model parameters (i.e., β):

1. $\beta_1 = \beta_2, = \beta$;
2. $\beta_1 - \beta$ is small (i.e., the sample error in the estimates of β obtained from the estimation context are small); and
3. These parameter estimates are obtained solely from the estimation context data, independent of and before consideration of application context data (which are allowed only to affect the application context constants and scales).

A much more general model that is fully consistent with the behavioral assumptions mentioned earlier is one in which β is jointly estimated using both the estimation and application context data sets, simultaneously with the estimation of the alternative-specific constants for both contexts and the scales of one context relative to the other.

The following notation is used in developing the joint context estimation procedure:

p = 1 for estimation context; = 2 for application context;

$\mathbf{s}_{it}^p$ = vector of explanatory variables for alternative i common to Periods 1 and 2 (i.e., associated with the constant parameter vector γ), but with values given for person t in context p;

α^p = vector of utility function parameters assumed to be specific to context p (at a minimum, this includes alternative-specific constants for context p);

$\mathbf{r}_{it}^p$ = vector of context-specific explanatory variables for alternative i for individual t within context p;

μ_i = utility function scale for alternative i in context 2 (context superscript has been suppressed to simplify notation; context 1 scales are assumed to be "imbedded" within α^p and γ; given this, μ_i is actually the ratio of context 2 scale to constant 1 scale for alternative i, with absolute values of either of these scales not being identifiable);

$\boldsymbol{\Theta}$ = combined vector of all parameters to be estimated within joint context model, excluding utility function scales

$$= \begin{bmatrix} \alpha^1 \\ \alpha^2 \\ \gamma \end{bmatrix}$$

$\mathbf{x}_{it}^p$ = combined vector of all explanatory variables in joint context model, for alternative i for person t in context p

$$= \begin{bmatrix} \mathbf{r}_{it}^1 \\ \mathbf{0} \\ \mathbf{S}_{it}^1 \end{bmatrix} \quad \text{for } p = 1$$

and

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{r}_{it}^2 & \text{for } p = 2 \\ \mathbf{S}_{it}^2 \end{bmatrix}$$

Given these definitions, the systematic utility components for the two contexts are

$$V_{it}^1 = \alpha^{1T}\mathbf{r}_{it}^1 + \gamma^T\mathbf{S}_{it}^1 = \boldsymbol{\Theta}^T\mathbf{X}_{it}^1 \tag{8}$$

$$V_{it}^2 = \mu_i(\alpha^{2T}\mathbf{r}_{it}^2 + \gamma^T\mathbf{S}_{it}^2) = \mu_i\,\boldsymbol{\Theta}^T\mathbf{x}_{it}^2 \tag{9}$$

Given the explicit accounting for changes in scales and constants between the two contexts, the usual IID Gumbel Type I distribution is assumed for the random utility terms in each context, leading to conventional multinomial logit choice models:

$$P^p{}_{it} = \frac{\exp(V_{it}^p)}{\sum_{j\in C_t} \exp(V_{jt}^p)} \qquad p = 1,2 \tag{10}$$

If n_p is the number of observations in the context p data set and y_{it}^p is the observed choice indicator for person t in context p (equals 1 if alternative i is chosen; equals 0 otherwise), then the joint log-likelihood function for the joint context model is simply

$$L = \ln L^* = \sum_{p=1}^{2} \sum_{t=1}^{np} \sum_{i\in C_t} y_{it}^p \ln P_{it}^p \tag{11}$$

Substituting Equations 8 through 10 into Equation 11 yields, on rearrangement;

$$L = \ln L^* = \sum_{t=1}^{n1} \sum_{i\in C_t} y_{it}^1\left[\boldsymbol{\theta}^T\mathbf{x}_{it}^1 - \ln\left(\sum_{j\in C_t} \exp(\boldsymbol{\theta}^T\,\mathbf{x}_{jt}^1)\right)\right] + \sum_{t=1}^{n2} \sum_{i\in C_t} y_{it}^2\left[\mu_i\boldsymbol{\theta}^T\mathbf{x}_{it}^2 - \ln\left(\sum_{j\in C_t} \exp(\mu_j\boldsymbol{\theta}_T\mathbf{x}_{jt}^2)\right)\right] \tag{12}$$

With straightforward changes in notation, this model is identical to that developed by Ben-Akiva and Morikawa for combining revealed and stated preference data sets within the same choice model (*7,8*). As noted by Morikawa et al. (*8*), "nested logit" full information likelihood estimation procedures can be applied to this model. Such a procedure was programmed in Fortran by the authors and used in computing the joint context model parameter estimates presented in this paper.

Joint context estimation can clearly be used as a model updating technique, providing that the estimation context data are available for combination with the small sample application context data set.

Although this will not always be the case, access to estimation context data is surely sufficiently feasible in many instances to warrant the testing of this approach compared with the previous three approaches discussed. In particular, potential advantages of joint context estimation relative to conventional transfer scaling techniques include the following:

1. It eliminates biases within the updated application context parameters caused by estimation context sampling errors [a problem discussed in detail elsewhere (6)]; and
2. It provides an operational full-information maximum likelihood procedure for parameter estimation when multiple cross-sectional data bases are available, as opposed to current methods, which are all limited information estimation procedures and hence inefficient in their use of data.

DATA

The 1964 estimation context data set is obtained from the 1964 Metropolitan Toronto and Region Transport Study (MTARTS) survey, which was a home interview survey of 3.3 percent of the households in Metropolitan Toronto and the surrounding regions, consisting of 24,000 households in total. This survey is documented elsewhere (*9*). The 1986 application context data set is obtained from the 1986 Transportation Tomorrow Survey (TTS), a telephone interview survey of 4 percent of the households in the GTA, or 67,000 households in total. This survey is documented elsewhere (*10,11*).

Both surveys were one-day travel surveys that collected generally comparable information, with the single biggest difference being that the 1986 TTS did not collect information on worker occupations and household income. Although coded to different zone systems, these zone systems are roughly similar in definition. Similarly, the study areas for the two surveys vary slightly but not significantly.

All level-of-service data required, with the exception of parking costs and transit fares (which were assembled from other sources), were generated using EMME/2 network assignment procedures applied to 1964 and 1986 road and transit networks. All costs were scaled to 1986 Canadian dollars on the basis of consumer price indexes for transportation.

RESEARCH METHOD

Test Procedure

In this study, the morning peak-period work trips contained in the 1964 MTARTS data base define the estimation context data set. Two multinomial logit work trip mode choice models are estimated using the 1964 data set: one that contains only level-of-service variables (i.e., modal travel times and costs), and one that in addition to these level-of-service variables includes as full a set of socioeconomic variables as is supported by the available data (referred to as the "fully specified" model).

The morning peak-period work trips contained in the 1986 TTS data base then define the application context data set. All four of the updating procedures discussed assume the existence of a "small" sample of trips drawn from the application context to be used in the updating calculations.

To simulate this small sample, random subsets of trip records are drawn from the full TTS data base (which consists in total of 32,328 usable records for this application). To explore the impact of sample size on updating performance, samples of 400, 800, 1,600, 3,200, and 6,400 are used (with each larger sample containing all the records included in the smaller samples). Both the level-of-service and fully specified models are then updated using each of the four updating procedures for each sample size.

In addition, level-of-service and fully specified models are estimated using each of the 1986 small samples. These small sample models are then used to compare the impact that the information contained in the transferred models contributes to predictive performance in the application context with respect to simply using the available application context data.

The performances of the four updated models and the 1986 small model are evaluated for each sample size–model specification combination in terms of how well they replicate the full 32,328 record 1986 TTS set of observed trips. The primary test statistic used is the log-likelihood value generated by the given model when it is applied to the entire 1986 TTS data set.

In addition, however, various aggregate prediction test statistics were constructed, all of which compare in various ways the aggregate number of predicted trips by mode m for a given aggregate group $g, \hat{N}_{mg}$, with the observed number of trips by this mode for this group, N_{mg}. In this paper, only one of these test statistics is discussed, the mean absolute error (MAE) defined as

$$\text{MAE} = \left\{\sum_m \sum_g |\hat{N}_{mg} - N_{mg}|\right\} \Big/ \left\{\sum_m \sum_g N_{mg}\right\} \qquad (14)$$

Two aggregations are examined: seven major destination groups and worker gender.

Model Specification and Estimation Context Parameters

Three modes are potentially included in the choice set in this study: automobile drive allway, transit allway, and walk. Although automobile passenger, automobile access to transit, and (in 1986) commuter rail modes in principle were also available, these modes were excluded from this analysis to reduce the modeling complexity with respect to specification, decision structure, and introduction of new modes (the commuter rail service did not exist in 1964). Table 1 defines the explanatory variables used in the two models, whereas Table 2 presents the estimation results obtained through standard maximum likelihood estimation of the models using the 1964 MTARTS data set.

RESULTS

Table 3 contains the 1986 full-sample log likelihood values computed for each model specification–sample size combination for the four updated models as well as the estimated 1986 small-sample models. Figures 1 and 2 display these log likelihoods for the level-of-service and fully specified models, respectively. Points to note from these figures and Table 3 include the following.

First, in view of the significant transfer bias, the combined transfer procedure as expected yields results that are virtually indistinguishable from the 1986 small-sample results. At very small samples (e.g., 400 observations), the "prior" information provided by the estimation context parameters contributes a very marginal amount of additional information (resulting in a 0.2 percent improvement in the full-sample log-likelihood value for the level-

TABLE 1 Definition of Variables

dauto	= 1 for auto-drive mode; = 0 otherwise
dwalk	= 1 for walk mode; = 0 otherwise
aivtt	= auto in-vehicle travel time (min.) for auto mode; = 0 otherwise
tivtt	= transit in-vehicle travel time (min.) for transit mode; = 0 otherwise
twait	= transit wait time (min.) for transit mode; = 0 otherwise
twalk	= transit access + egress time (min.) for transit mode; = 0 otherwise
ivtc	= auto in-vehicle travel costs ($) for auto mode; = 0 for walk mode; = transit fare ($) for transit mode
pkcst	= auto daily parking cost ($) for auto mode; = 0 otherwise
wdist	= walk distance (km.) for walk mode; = 0 otherwise
avplic	= number of vehicles per licensed person in household for auto mode; = 0 otherwise
wcbd	= 1 if worker works in PD1 (Planning District 1) for walk mode; = 0 otherwise
amal	= 1 for male worker for auto mode; = 0 otherwise
tcbd	= 1 if worker destination is PD1 for transit mode; = 0 otherwise
tgend	= 1 if worker is female for transit mode; = 0 otherwise

of-service model and a 0.1 percent improvement for the fully specified model relative to the 1986 small-sample model results). Beyond this point, however, it is clear that the transfer scaling component of the procedure completely dominates the calculations. Because in this case the transfer scaling adjusts every parameter in the model, this is equivalent to simply reestimating the model on the basis of the small-sample application context data and using the reestimated parameters directly. To the extent that this result is verified in other empirical settings, it implies that combined transfer updating contributes little relative to simply reestimating the model on the basis of the application context small sample (a theme that is discussed more generally later), except perhaps in the case of extremely small samples.

Of the remaining procedures, the joint estimation procedure always performs the best, regardless of model specification or sample size used. This is not surprising given that the joint procedure is the only full-information procedure of the three. The improvement in performance achieved with the joint procedure increases with model specification: at the application sample size of 1,600, for example, the joint procedure reduces the log-likelihood value relative to the Bayesian procedure by only about 1 percent for the level-of-service model, whereas it generates about a 4 percent improvement for the fully specified model.

Conversely, the joint estimation procedure performs marginally better than the combined transfer procedure for small sample sizes for the simpler level-of-service model, whereas the combined transfer procedure performs slightly better than joint estimation at all sample sizes for the fully specified model. In comparing these two procedures, however, it should be noted that the combined transfer procedure effectively requires the estimation of $2(N + M - 1)$ parameters, where N is the number of alternatives and M is the number of utility function parameters (excluding alternative-specific constants); that is, $(N + M - 1)$ parameters from each of the estimation and application contexts. The joint context estimation procedure, on the other hand, requires $2(N - 1)$ alternative-specific constants, N scales, and M other utility function parameters to be estimated, for a total of $M + 3N - 2$ parameters, $M - N$ fewer than the combined transfer procedure.

Given $N = 3$ in this case, for the level-of-service model ($M = 7$) 18 parameters are estimated in the combined transfer model, whereas 14 parameters are estimated in the corresponding joint context model. This is a 29 percent increase in model parameters yielding no improvement in model performance below the 800 sample level and at most a 0.5 percent improvement in full-sample log likelihood over the entire range investigated.

Similarly, for the fully specified model ($M = 12$), 28 parameters are required by the combined transfer procedure versus 19 for the joint context procedure, a 47 percent increase in parameters, which yields at most a 3.3 percent improvement in full-sample log likelihood. Thus, joint context estimation would appear to be the more parsimonious of the two updating procedures, and, hence, all else being equal, preferred.

The constant/scale updating procedure performs surprisingly well at small sample sizes. For the level-of-service model, it performs virtually as well the joint procedure up to the 1,600 observation level and it clearly outperforms the computationally more complex Bayesian procedure up to at least the 6,400 observation level. The procedure's performance relative to the others decreases with improved model specification, but it is still comparable to the joint procedure at the 400 observation level and with the Bayesian procedure up to the 1,600 observation level for the fully specified model. This sensitivity to model specification is a sensible one, given that the relative role of constants (in particular) within the model should decline as model specification improves.

Given that small sample updating generally utilizes sample sizes in the order of 1,000 or less, these results imply that updating model scales and constants—a simpler and less onerous task than Bayesian updating—may well outperform the Bayesian procedure. The

TABLE 2 1964 (Estimation) Context Model Parameter Estimates

Parameter	Level of Service (I) Model		Fully Specified Model	
	Estimate	t-value	Estimate	t-value
dauto	0.090	0.452	-1.266	-4.362
dwalk	0.924	3.853	1.592	6.029
aivtt	-0.031	-10.731	-0.009	-2.217
tivtt	-0.043	-9.584	-0.029	-6.205
twait	-0.205	-12.444	-0.202	-11.653
twalk	-0.046	-3.205	-0.026	-1.729
wdist	-1.961	-21.892	-1.884	-20.935
ivtc	-0.389	-3.697	-0.388	-3.488
pkcst	-0.333	-11.583	-0.282	-9.134
avplic			1.874	12.185
amal			0.740	4.842
tcbd			1.224	9.142
tgend			0.759	5.299
wcbd			0.943	4.951
Number of observations	8066		8066	
Log-likelihood at Zero	-5929.6		-5929.6	
Log-likelihood at Convergence	-2839.4		-2590.5	
Adjusted rho-square	0.5204		0.5625	

TABLE 3 Full-Sample 1986 TTS Log-Likelihood Values for Alternative Models and Updating Sample Sizes

Model Type	Sample Size	Log-Likelihood Values				
		Bayesian Updating	Transfer Scaling	Joint Context Estimation	Combined Transfer Estimator	Small Sample
Level of Service Model	400	-11201	-11081	-11076	-11081	-11105
	800	-11156	-11076	-11075	-11077	-11086
	1600	-11133	-11052	-11045	-11012	-11014
	3200	-11072	-11012	-10991	-10937	-10937
	6400	-11013	-11000	-10963	-10930	-10930
	32328	-10942	-10996	-10931	-10920	-10920
Fully Specified Model	400	-10277	-10128	-10100	-9762	-9774
	800	-10138	-10025	-9903	-9621	-9626
	1600	-9999	-10013	-9805	-9555	-9555
	3200	-9865	-9952	-9622	-9471	-9470
	6400	-9712	-9945	-9545	-9470	-9470
	32328	-9501	-9941	-9475	-9453	-9453

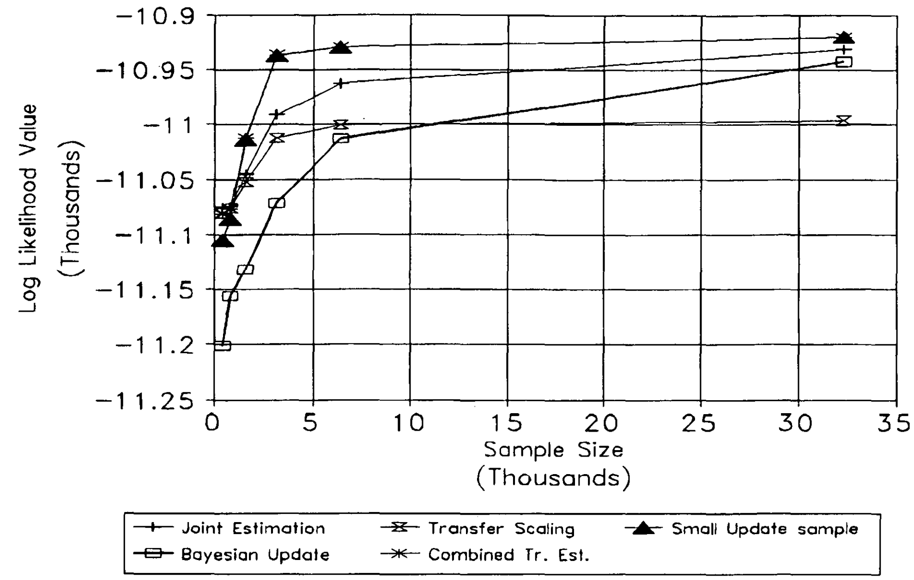


FIGURE 1 Full-sample 1986 TTS log likelihood values, level-of-service models.

results also imply that the joint estimation procedure may add little additional information to the updated model, relative to simply updating constants and scales, at least for sample sizes of 400 to 500 or less.

Comparison of the performance of the updated models with that of the 1986 small-sample models (i.e., the models simply estimated using the 1986 small samples) raises some question concerning the utility of updating a transferred model at all given the availability of an application context small sample. That is, the small-sample models outperform most of the updated models at most sample sizes. Thus, if one has a small sample of at least 400 to 500 observations, these results imply that one would be at least as well off to simply estimate an application context model, rather than to update a model developed elsewhere, especially if a relatively good specification is supported by the application data set.

Indeed, Table 3 and Figures 1 and 2 reinforce the importance of model specification in the determination of model performance by showing that the differences between the level-of-service model log-likelihoods and the corresponding fully specified model log likelihoods are far greater than the total differences between updating procedures or across sample sizes within either of the model specifications. In particular, note that the log likelihood for the 400-sample 1986 fully specified model of −9773.83 is larger than any of the full 32,328 sample level-of-service models.

Figures 3 and 4 present the aggregate MAE statistics for the four updated models and the 1986 small-sample models as a function of sample size for the level-of-service and fully specified models, respectively. The results here are less clear cut, reflecting the fact that different aggregations result in different combinations of com-

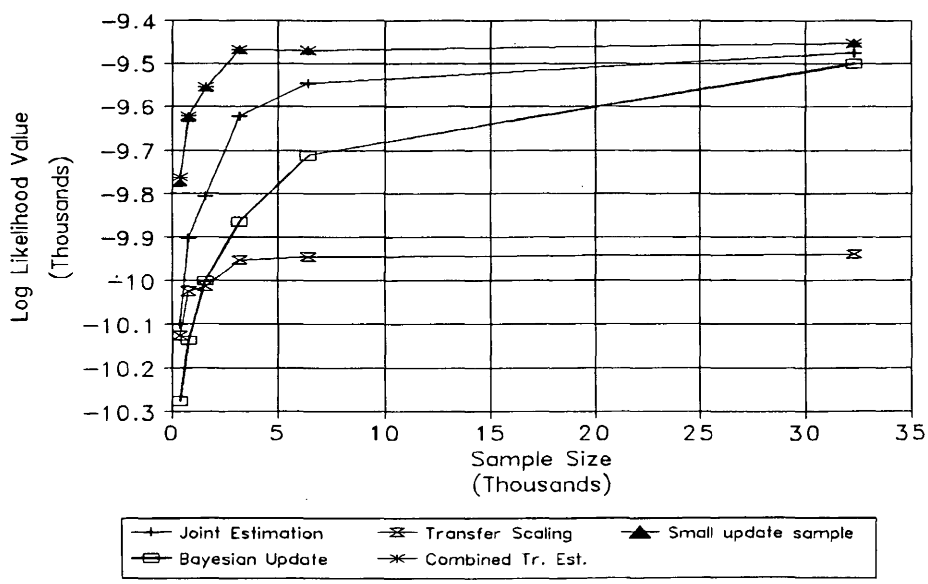


FIGURE 2 Full-sample 1986 TTS log-likelihood values, fully specified models.

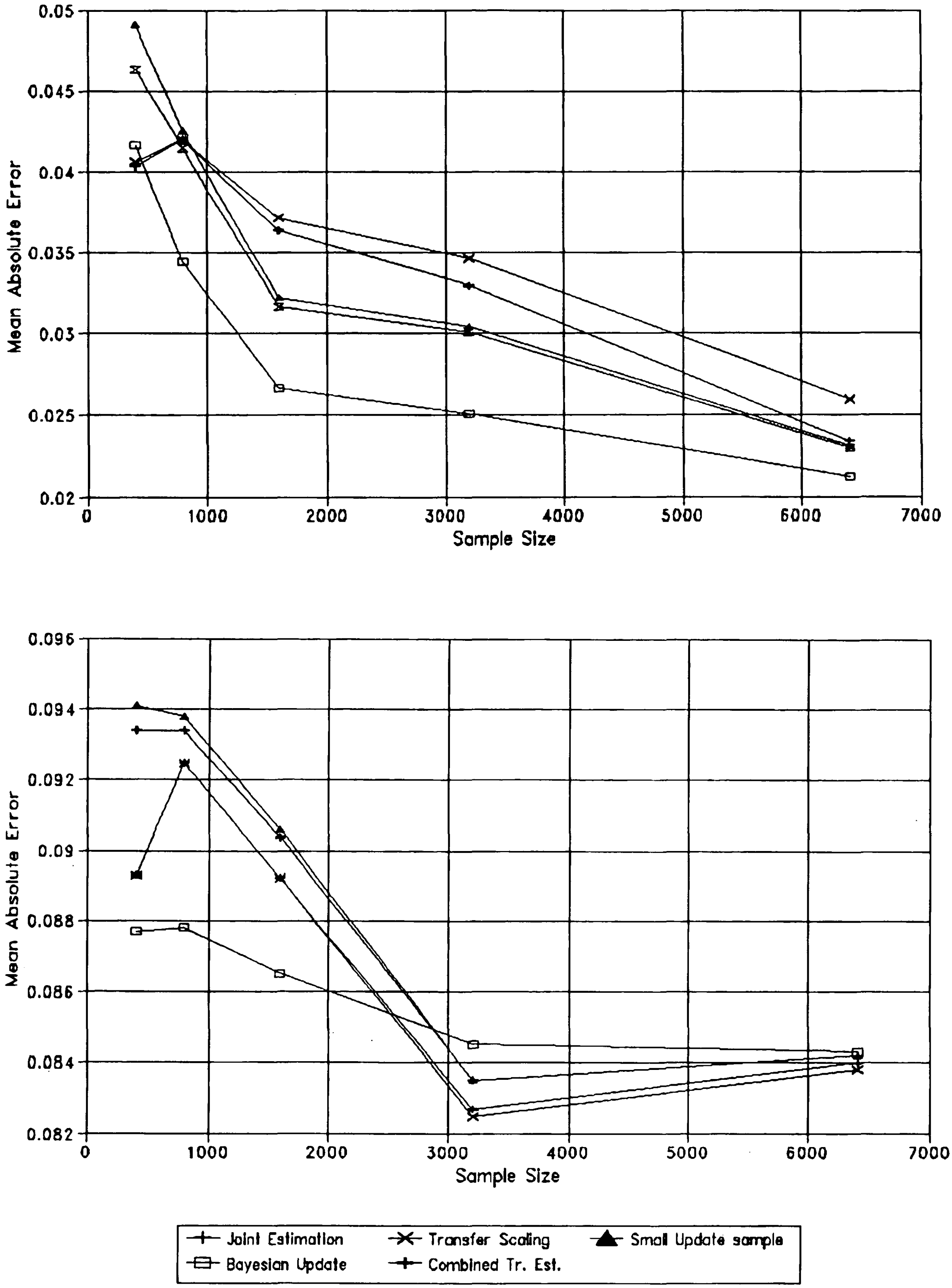


FIGURE 3 **Aggregate mean absolute prediction errors, level-of-service models:** ***top,*** **aggregation by destination region;** ***bottom,*** **aggregation by worker gender.**

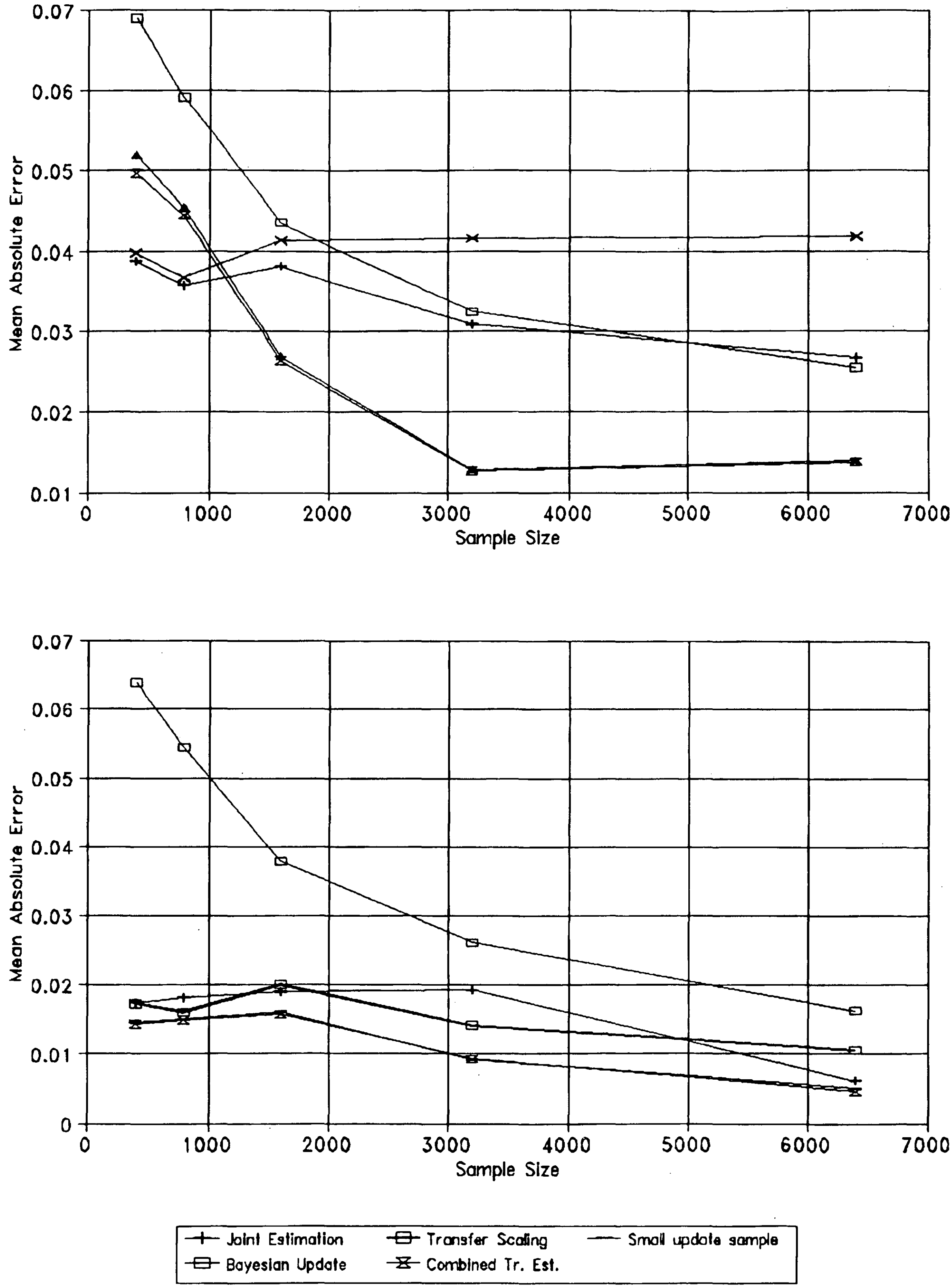


FIGURE 4 **Aggregate mean absolute prediction errors, fully specified models:** ***top,*** **aggregation by destination region;** ***bottom*****, aggregation by worker gender.**

pensating errors. Nevertheless, some general trends are evident in these figures.

In the case of the level-of-service models and small sample sizes (e.g., under 1,000), the Bayesian procedure consistently yields the best aggregate predictions, the joint estimation and constants/scales updating generally yield results similar to one another that are slightly poorer than the Bayesian results, and the combine transfer and 1986 small-sample models generally yield the poorest aggregate predictions. The results for the fully specified model are more mixed but, in general, are different from the level-of-service results in that the combined transfer and 1986 small sample model results are, overall, the best, whereas overall the Bayesian procedure performs the most poorly (especially at sample sizes under 1,000). The other two updating procedures are again fairly comparable at small sample sizes and again generally lie between the best and the worst, although in this case their performance is generally close to the best.

Figures 3 and 4 again reinforce the importance of model specification in that the aggregate prediction errors are generally smaller for the fully specified model and the sensitivity to sample size is generally larger for the fully specified model as well.

SUMMARY AND CONCLUSIONS

This paper has provided an empirical comparison of four disaggregate choice model updating procedures using two data sets from the GTA representing travel behavior at two points in time 22 years apart (1964 and 1986). All the results obtained are based on this one case study, implying the need for additional tests employing other estimation/application contexts to be able to generalize any conclusions that arise from this study. On the basis of this study's results, however, the following findings are noteworthy.

1. The combined transfer estimation procedure consistently yields the best predictive performance in the 1986 application context, on the basis of the disaggregate full-sample log-likelihood measure used. This, however, is largely the result of the dominance of the transfer scaling component of the procedure, which effectively results in the procedure corresponding to a simple reestimation of the model using the application context data set.

2. The joint context estimation yields results generally comparable to the combined transfer procedure, but with a significantly more parsimonious parameter structure. Hence, if the estimation context data set is available to support joint context estimation, generally it should be preferred relative to combined transfer estimation.

3. The computationally simpler transfer scaling procedure yields results that are similar to those of joint context estimation for small sample sizes. Hence, if the software required for joint context estimation or the estimation context data set, or both, are not available, then transfer scaling may well provide a useful and credible model update.

4. The Bayesian updating procedure is generally dominated by the other updating procedures examined, all of which explicitly deal with transfer biases in various ways. Thus, on the basis of this case study, Bayesian updating cannot be recommended as an updating procedure, especially given alternative techniques, such as transfer scaling and combined transfer estimation, which are not any more burdensome computationally and yet yield superior results.

5. Once the application context small sample reaches the 400 to 500 observation level, simply reestimating the model for the application context may yield results that are comparable or superior to any updated model transferred from an estimation context—providing that the application context data set supports development of a "fully specified" model.

6. Model specification is important in the updating/transfer process. In this case study, improving the model specification yielded far greater improvements in model performance than either "optimizing" the updating procedure or increasing the application context sample size.

In conclusion, this study indicates, in keeping with other studies cited in the paper, that updating a model estimated in another context through use of a small sample drawn from a new context significantly improves the model's transferability to this new context. In comparing the performance of a range of updating methods suggested in the literature, this study indicates that three procedures that all explicitly address the issue of transfer bias (transfer scaling, combined transfer estimation, and joint context estimation) all perform well at small sample sizes and possess merit as possible updating procedures for practical application. The choice among these methods depends on model specification, application context sample size, and availability of the estimation context data set. All else being equal, however, the joint context estimation procedure may be preferred given that it is a parsimonious, full-information approach to the problem.

ACKNOWLEDGMENTS

The work reported in this paper was funded by an operating research grant from the Natural Science and Engineering Research Council (Canada). Access to the travel behavior data bases and EMME/2 network modeling system used in the study was provided by the University of Toronto Joint Program in Transportation.

REFERENCES

1. Badoe, D. A., and E. J. Miller. Analysis of the Temporal Transferability of Disaggregate Work Trip Mode Choice Models. Presented at 74th Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1995.
2. Atherton, T. J., and M. E. Ben-Akiva. Transferability and Updating of Disaggregate Travel Demand Models. In *Transportation Research Record 610,* TRB, National Research Council, Washington, D.C., 1976, pp. 12–18.
3. Gunn, H. F., M. E. Ben-Akiva, and M. A. Bradley. Tests of the Scaling Approach to Transferring Disaggregate Travel Demand Models. In *Transportation Research Record 1037,* TRB, National Research Council, Washington, D.C., 1985, pp. 21–30.
4. Koppelman, F. S., G. K. Kuah, and C. G. Wilmot. Transfer Model Updating with Disaggregate Data. In *Transportation Research Record 1037,* TRB, National Research Council, Washington, D.C., 1985, pp. 102–107.
5. McCoomb, L. A. Analysis of the Transferability of Disaggregate Demand Models Among Ten Canadian Cities. *Transportation Forum,* Vol. 3-1, 1986, pp. 19–28.
6. Ben-Akiva, M. E., and D. Bolduc. Approaches to Model Transferability and Updating: The Combined Transfer Estimator. In *Transportation*

Research Record 1139, TRB, National Research Council, Washington, D.C., 1987, pp. 1–7.

7. Ben-Akiva, M. E., and T. Morikawa. Estimation of Travel Demand Models from Multiple Data Sources. *Proc., 11th International Symposium on Transportation and Traffic Theory,* Yokohama, Japan, July 1990.
8. Morikawa, T., M. E. Ben-Akiva, and K. Yamada. Forecasting Intercity Rail Ridership Using Revealed Preference and Stated Preference Data. In *Transportation Research Record 1328,* TRB, National Research Council, Washington, D.C., 1991, pp. 30–35.
9. Traffic Research Corporation. *An Analysis of Some Travel Trends Between 1956 and 1964 Based on Home Interview Surveys.* Metropolitan Toronto Planning Board, Ontario, Canada, 1965.
10. *The Transportation Tomorrow Survey, Design and Conduct of the Survey.* Ministry of Transportation, Toronto, Ontario, Canada, 1987.
11. Data Management Group. *TTS Version 3: Data Guide.* University of Toronto Joint Program in Transportation, Ontario, Canada, 1990.

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

Competing Risk Hazard Model of Activity Choice, Timing, Sequencing, and Duration

DICK ETTEMA, ALOYS BORGERS, AND HARRY TIMMERMANS

Recently hazard models have become increasingly popular in transportation research for modeling duration processes of various kinds. The application of hazard models is extended to the field of activity scheduling to account for the continuous nature of the decision-making process underlying activity performance. A competing risk hazard model of the accelerated time type, which describes simultaneously the duration of the present activity and the choice of the next activity, is presented. Both a generic and an activity-specific version of the model were estimated. The covariates used in the model represent factors that affect activity scheduling, such as time of day, opening hours, travel times, priorities, and time budgets. An interactive computerized data collection procedure was used to obtain specific data needed to calculate the covariates. The estimated models performed satisfactorily, suggesting that competing risk models are a useful tool for describing activity scheduling as a continuous decision-making process. This is an important finding, especially because influencing the timing of activities and trips is a subject of increasing interest to policy makers.

In past decades, activity scheduling has been a topic of increasing interest in the transportation research community (*1*). The central assumption underlying this stream of research is that people travel to participate in various activities that satisfy their personal needs. Thus, the key question in understanding how travel decisions are made and how people will adapt their travel behavior to changes in their environment is how people decide about activity performance and related travel behavior. More specifically, it requires an understanding of the activity scheduling process, which encompasses decisions about which activities to perform, at which locations, at which times, in which sequence, and which travel modes and routes to use.

Modeling efforts in transportation have addressed several aspects of activity scheduling. For instance, discrete choice models of destination choice, mode choice, and route choice are well known and widely applied, whereas multidimensional models encompassing several of these choices, often using a nested logit approach, are becoming increasingly popular (*2*). More specific applications include models of combined activity and destination choice throughout the day (*3*) and trip chaining models, describing the sequencing of activities (*4,5*). Other approaches describe the choice of complete activity patterns explained by their scheduling convenience (*6*) or the planning phase that precedes activity execution (*7*).

Another approach in choice modeling with possibly relevant implications for activity scheduling is the development of dynamic discrete choice models (*8,9*). These models typically describe how choice behavior develops over time. By including state dependence and heterogeneity, the choice made at time t is explained partly by choices made previously so that changes in behavior are modeled rather than independent choices. Models of this type have been applied in the analysis of panel data to describe vehicle transactions and various kinds of travel behavior (*8,7*). Applications in the field of activity scheduling, however, are scarce.

Dynamic discrete choice models, however, do not go without severe computational difficulties, especially if the number of alternatives and waves is large, which is typically the case in activity scheduling analysis. Furthermore, activity performance is increasingly regarded as a continuous process, in which individuals can decide during activity performance to end an activity and start another one. The decision whether to continue or stop will therefore depend strongly on time and duration of the present and previous activities. Thus, the probabilities of pursuing different activities and travel to different locations will change continuously over time. Both static and dynamic discrete choice models do not explicitly account for this duration dependence.

Recently, hazard models have gained increasing interest in transportation research as a means to describe the duration of processes such as activity performance (*10*). Hazard models therefore are promising tools for incorporating duration dependence into activity-based approaches and taking into account the continuous nature of the implied decision making. The specific contribution of this paper to the literature is the introduction of a competing risk hazard model to activity scheduling modeling to describe not only activity duration but also activity choice. Spatiotemporal constraints were incorporated by using specific individual data on available locations and hours obtained by using a computerized interactive data collection procedure. The results indicate that transitions between activity types can be described by a competing risk model with covariates accounting for spatiotemporal flexibility of activities.

The remainder of this paper is organized as follows. In the second section, hazard models are introduced and discussed. Special attention is given to competing risk models and issues of heterogeneity and risk interdependency. In the third section, the model that was used in the present research is discussed. The model structure and the covariates, representing spatiotemporal constraints, are outlined. In the fourth section, the data collection procedure, which was performed using a recently developed interactive computer procedure, is described. The fifth section describes the results of the analyses. Different specifications of the competing risk model are discussed. Finally, the sixth section summarizes the findings and addresses directions for future research.

THEORETICAL BACKGROUNDS OF HAZARD MODELS

Basic Concepts

In this paper, a series of hazard models is applied to describe and analyze activity scheduling processes. Because hazard models have

Urban Planning Group, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

not been widely used in transportation research the basic principles of hazard models will be discussed and summarized to allow a better understanding of the empirical findings of this study.

Although hazard models only recently have gained increasing popularity in transportation modeling, they have been applied for decades in other disciplines, such as industrial engineering, biology, medical science, and labor market research (11,12). Hazard models typically are applied to describe duration data such as machine failure times or patient survival times under different medications or unemployment periods. More specifically, hazard models describe the probability of occurrence of a certain event (machine failure, death, finding a job) within an interval $[t, t + dt]$, given that it has not occurred up to time t. This conditionality can be considered the key concept of hazard modeling and offers a natural framework for describing durations and intervals between the occurrence of events. For instance, in the case of activity duration, the probability of stopping an activity will be small when it has just started and will gradually increase with the time of execution. Hazard models offer the statistical tools to describe this conditional probability, which enables one to incorporate duration dependence into transportation modeling.

Mathematical Formulation

A number of functions are of particular interest with respect to the mathematical description of hazard models. First, a probability density function $f(t)$ giving an unconditional distribution of durations T within a population can be defined as

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \quad T \leq t + \Delta t)}{\Delta t} \tag{1}$$

The probability that in a specific case the event will occur before time t is then

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{2}$$

It follows that $f(t)$ is the first derivative of $F(t)$ with respect to time. A key function in hazard modeling is the survivor function $S(t)$, giving the probability that the process has survived until t:

$$S(t) = 1 - F(t) = P(T \geq t) = \int_t^{\infty} f(u)du \tag{3}$$

The hazard function $h(t)$, finally, describes the probability of occurrence at t conditional on survival until t:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|T \geq t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{4}$$

In principal, the hazard can take many different forms (see Figure 1). It can be monotonically increasing (*a*), U-shaped (*b*), monotonically decreasing (*c*), or constant (d). Lawless (11) and Kalbfleisch and Prentice (13) give examples of shapes that are typical for certain types of duration processes. Given that the shape of the hazard yields important information about the nature of the process under study, remarkably little attention has been paid to the specific shape of the hazard in transportation applications. The emphasis has been primarily on the influence of covariates influencing the scale of the hazard, indicating longer or shorter durations in general.

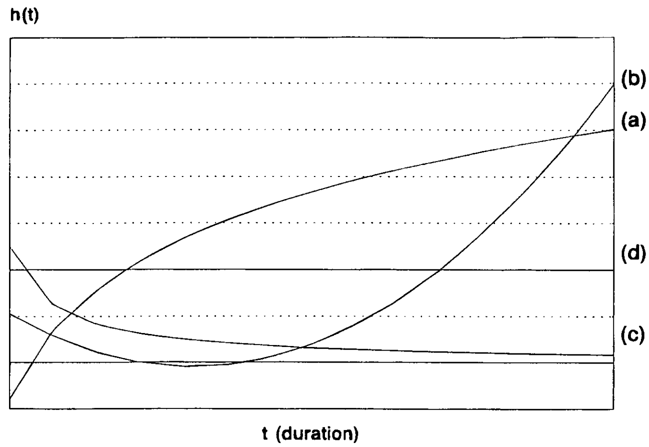


FIGURE 1 Some hazard functions.

The shape of the hazard function is determined by the distributional assumptions that are made for the probability density function $f(t)$. A number of different distributions can be chosen (11), resulting in different hazard functions. Some distributions and their related hazard functions are listed below. For a detailed review of possible distributions the reader is referred to Lawless (11) and Kalbfleisch and Prentice (13).

1. Exponential distribution:

$$h(t) = \lambda \qquad t \geq 0 \tag{5}$$

2. Weibull distribution:

$$h(t) = \lambda\beta(\lambda t)^{\beta-1} \qquad \lambda,\beta > 0 \tag{6}$$

3. Log normal distribution:

$$h(t) = \frac{\dfrac{1}{(2\pi)^{1/2}\,\sigma t} \exp\left[0.5\left(\dfrac{\log t - \mu)}{\sigma}\right)^2\right]}{1 - \displaystyle\int_{-\infty}^{t} \frac{1}{(2\pi)^{1/2}} e^{-\mu^2/2}\, du \left(\frac{\log t - \mu}{\sigma}\right)} \tag{7}$$

4. Log logistic distribution:

$$h(t) = \frac{\lambda\beta(\lambda t)^{\beta-1}}{1 + (\lambda t)^{\beta}} \tag{8}$$

The choice of a specific distribution and related hazard function usually will be made according to hypotheses based on existing theory. However, testing of different distributions with different scale and shape parameters may often lead to a better insight into the duration process under study.

Parametric Hazard Models

Apart from duration dependence other factors also influence activity duration and timing. For instance, the start of an activity may be influenced by opening hours, time of day, or priority of the activity. To incorporate such explanatory variables into the model two model types can be used. The first is known as the proportional hazard model, which takes the following form:

$$h(t \mid X) = h_0(t)g(X) \tag{9}$$

where

X = a vector of explanatory variables and
$h_0(t)$ = the baseline hazard function.

The baseline hazard is the hazard function assuming that all covariates X have a value of 0. $g(X)$ is usually defined as *exp* (βX), where β is a vector of parameters. The function g thus acts multiplicatively on the baseline hazard. This causes the property of proportionality, implying that the ratio of hazards for specific sets of covariates (h_1/h_2) remains constant over time. This assumption however can in some cases be undesired. For instance, Leszczyc and Timmermans (*14*) found that intershopping trip times differed, depending on the store chains that were visited. In this research, different duration processes may be expected for different types of activities. In such cases the proportion of hazards of different destinations is likely to vary over time. Accelerated lifetime models can be used to describe such cases. These models are log linear for T:

$$\log T = X\beta + \epsilon$$

The hazard function in this case can be shown to be

$$h(t|X) = h_0\,(te^{-X\beta})e^{-X\beta} \tag{11}$$

Thus, the effect of the covariates X is on t rather than on the baseline hazard. The models are not proportional and offer greater flexibility in modeling durations of alternative processes. This, however, comes at the cost that heterogeneity cannot be incorporated into the model. In both cases different forms of the hazard function are obtained by taking different distributions for the baseline hazard h_0 as described earlier.

Competing Risk Models

The previous description has considered durations of processes with only one exit. However, the ending of an activity can be the means to starting various new activities so that there will be different possible exits. A competing risk model was used to describe transition rates to these competing risks. In this case, hazards are defined specifically for different exits:

$$h_k(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t,\, D_k = 1 | T \ge t)}{\Delta t} \tag{12}$$

where $h_k(t)$ is the probability that exit k occurs at time t and D_k is a dummy variable indicating whether or not exit k was chosen.

The relation between hazards and survival functions for specific exits and joint hazard and survival functions is given simply by

$$h(t) = \sum_k h_k(t)dt \tag{13}$$

$$S(t) = \sum_k S_k(t) \tag{14}$$

Parametric versions of the proportional and accelerated time type are written as follows:

$$h_k(t \mid X) = h_{0k}\,(t)e^{X_k\beta_k} \tag{15}$$

$$h_k(t \mid X) = h_{0k}(te^{-X_k\beta_k})e^{X_k\beta_k} \tag{16}$$

where different distributional assumptions can be made for h_{0k}, the k-specific baseline hazard.

The probability that, if an exit is chosen, this exit will be k, π_k can be calculated as follows:

$$\pi_k = \int_0^\infty S(s)\, h_k(S)\, ds \tag{18}$$

Lancaster (*12*) shows that under the assumption of stationarity ($h_k(t)/h(t) = m_k$ for all t) and a Weibull distribution for the probability density function π_k can be written as follows:

$$\pi_k = \frac{\exp\,(x_k\,\beta_k)}{\sum_j \exp\,(x_j\,\beta_j)} \tag{19}$$

Thus, the well-known logit model can be regarded as a competing risk model under strong assumptions. This example clearly illustrates that competing risk models offer the attractive opportunity of relaxing the static assumption underlying discrete choice modeling and incorporating dynamic aspects into consumer behavior research. In a number of transportation applications, especially where the timing of travel decision is concerned, this might be a valuable contribution.

Notwithstanding these attractive features, some issues should be addressed in the application of competing risk models. Competing risk models fall in the class of models with multivariate lifetime distributions, with different distributions of lifetimes T_i according to the competing risks. If information on all lifetimes T_i is available, one can test for independence of the various lifetime distributions. However, in the case of competing risks only min ($T_1, \ldots, T_k$) is observed so that the assumption of independence cannot be tested. This is caused by the fact that it is principally impossible to discriminate between different multivariate distributions $f(t_1, \ldots, t_k)$ that give rise to the same cause-specific hazard functions based on min ($T_1, \ldots, T_k$) only (*11*). Recently, Han and Hausman (*15*) introduced a proportional hazard model that allows for testing of independence among risks. In their approach, time is divided into T discrete periods and a proportional hazard model is formulated in an ordered logit or ordered probit form. Interdependency can then be incorporated by correlations in the stochastic terms of the model.

A second issue that should be addressed is the problem of heterogeneity within the sample. In case of observed heterogeneity, characteristics of subjects that can easily be measured, such as sociodemographics, influence the observed behavior. Heterogeneity can then be accounted for by including the sociodemographics as explanatory variables in the model. Unobserved heterogeneity exists when unobserved characteristics of subjects in the sample (e.g., motivations, tastes, and preferences) correlate with the observed behavior—in this case activity choice and duration. Not accounting for heterogeneity may lead to biased results. For instance, Meurs (*9*) found that linear regression models without heterogeneity lead to underestimation of elasticities. The effects of ignoring heterogeneity in duration models are less clear cut. Studies by Hensher (*16*) and De Jong et al. (*17*) seem to suggest that including heterogeneity does not have a dramatic effect on the parameter estimates of the explanatory variables but has a larger impact on the shape and scale parameters of the distribution of the baseline hazard. An additional complication arises when multiple observations for one subject are included in the sample, for example, if panel data or multispell duration data are used. If heterogeneity exists, the observations of one subject will be interdependent. By treating the observations as independent, one can easily overestimate the effects of state and time dependence and habit persistence (*18*).

To account for heterogeneity in proportional hazard models usually a heterogeneity term is introduced, which is a random variable with a certain (often gamma) distribution (*13,16*). Lancaster (*12*) and Sueyoshi (*19*) extend the inclusion of a mixing distribution to the competing risk case. By specifying mixing distributions V_j for competing risks, the joint distribution can be used to account for interdependency between risks. However, in the case of accelerated time models, introduction of a heterogeneity term is not possible because of identification problems (*20*).

COMPETING RISK MODEL OF ACTIVITY CHOICE, SEQUENCING, TIMING, AND DURATION

In the current research, the sequencing and timing of activities during the course of a day are of interest. Two models were estimated to describe this process. Both models describe the transition from one activity to another. The competing risks by which the origin activity can end are the possible following activities. The dependent variable in the models is the duration of the first activity that is equivalent to the time until a transition to another activity takes place. However, the covariates used to explain the duration are generic in one model and specific for various activity types in the other model. For this application, models should be obtained in which the proportion of transition probabilities to different activity classes can change over time. This implies an accelerated time formulation for each model. The generic model can be specified as

$$h_k(t) = h_0(te^{-\beta X})\, e^{-\beta X} \tag{20}$$

where

$h_k(t)$ = hazard function for transition to any activity k,
h_0 = baseline hazard,
X = vector of generic covariates, and
β = vector of generic parameters.

The specific model is given by

$$h_k(t) = h_0(te^{-\beta_k X_k})e^{-\beta_k X_k} \tag{21}$$

where

$h_k(t)$ = hazard function specific for transition to activity k,
X_k = vector of covariates specific for activity k, and
β_k = vector of parameters specific for activity k.

Obviously, the choice for the accelerated time model has some implications. First, it does not allow incorporation of heterogeneity into the model. In addition, it is not possible to readily test for independence of activity choices. Hence, the choice of which model to use was based on a tradeoff between incorporating heterogeneity and interdependence between risks and the flexibility to allow the ratio of transition rates to different risks to vary over time. The latter should receive the priority that led to the choice of the accelerated time model in this project.

The following activities were distinguished in the specific model:

1. In-home leisure activities,
2. In-home task activities,
3. Work/education,
4. Shopping,
5. Personal business out of home (not item 3 or 4), and
6. An end state in which no further activities are performed.

The covariates X used in both models to explain activity duration and transition to other activities are derived from previous research (*21*), which revealed that spatiotemporal constraints and general characteristics of activity performance were relevant for activity scheduling behavior. The following covariates describing spatiotemporal flexibility were used:

1. The activity from which the transition takes place. Five dummies are used to represent the possible activities, being the first five activity types mentioned earlier. These are generic variables in both models. The dummies represent differences in average activity duration between different classes of activities.
2. The activity to which the transition takes place. Dummy coding was used in a similar way to represent the six possible destination states. The dummies represent the effect of the destination activity on the duration of the preceding activity.
3. START: the start time of the first activity in minutes. It is assumed that the time of day at which activities start may influence the probability of transition to another activity. For instance, the probability of switching to leisure activities may be higher at the end of the day, whereas switching to work is more likely at the beginning of the day.
4. TILSTART: the time until the next activity can start in minutes. This factor represents the influence of opening hours of facilities or the influence of fixed hours for certain activities, such as work or education. It is hypothesized that if less time remains until an activity can start, a transition to this activity is more likely to take place. If PS2 is the earliest possible start time of the destination activity, TILSTART is calculated as follows:

TILSTART = PS2 − START

This measure takes a value of 0 if the activity can start before START.

5. TILCLOSE: the time until the next activity can end at the latest in minutes. This factor represents the effect of closing times or the end of fixed hours for certain activities. The effect can be

twofold: if less time remains for the execution of an activity, it becomes more urgent so that transition to this activity is more likely to take place. However, if too little time remains for the execution of an activity a transition will become less likely. If PE2 is the latest possible endtime of the destination activity, TILCLOSE is calculated as follows:

TILCLOSE = PE2 − START

If START > PE2, then TILCLOSE is set to 0.

6. PRIOR1: the priority of the first activity on a 0 to 10 scale. It can be hypothesized that the priority of the first activity will influence the duration of this activity, in the sense that activities with lower priority are more likely to be ended to pursue other activities.

7. PRIOR2: the priority of the next activity on a scale of 0 to 10. Analogous to PRIOR1, a transition to an activity with higher priority is more likely to take place if the priority of this activity is higher.

8. TRAVTIME: the travel time between the origin and the destination activity in minutes. This factor represents the distance decay over time of switching to different activities.

9. TIMESPENT: the time spent on the destination activity type at earlier occasions during the same day in minutes. This factor represents history dependence. That is to say, the amount of time spent on an activity earlier in the day is likely to influence the probability of switching to the activity once more.

DATA COLLECTION

The competing risk model was estimated using activity scheduling data that were collected in January 1994. Subjects were 39 students of Eindhoven University of Technology, Eindhoven, The Netherlands. The data were collected using the interactive computer procedure MAGIC (*22*), which consists of two parts. In the first part general information on activity performance and spatiotemporal constraints is collected. For 31 activities the following information is recorded for each subject:

1. Will the activity be performed on the planning day according to an arrangement in which other people are involved (yes/no)?
2. What was the last time the activity was performed (days ago)?
3. What is the average frequency of performance of the activity (times per month)?
4. How long does is take to perform the activity (minimum time, average time, maximum time)?
5. How likely is it that the activity will be performed on the target day (on a 0 to 10 scale)?
6. What are the locations at which an activity takes place? Of each location the subject is asked to provide the following information: (*a*) the name of the location, (*b*) the hours at which the subject would consider performing the activity at this location (this may be a smaller range than is implied by strict opening hours), (*c*) the attractiveness of the location on a 0 to 10 scale, indicating how pleasant the location is to stay at, and (*d*) The address of the location.

The list of 31 activities is designed to cover the spectrum of daily and incidental activities and includes both in-home and out-of-home activities; it includes the following:

- Taking an educational course;
- Studying at home;
- Practicing hobbies at home;
- Buying provisions;
- Visiting post office or bank;
- Visiting a cafe, bar, or disco;
- Visiting a sports match;
- Sightseeing;
- Eating breakfast;
- Housekeeping;
- Visiting someone;
- Performing work;
- Visiting cashpoint;
- Participating in sports;
- Attending the theater or a concert;
- Eating lunch;
- Reading;
- Having visitors;
- Buying clothes or shoes;
- Engaging in club activities;
- Volunteering;
- Attending a museum or exhibition;
- Eating supper;
- Watching television;
- Getting food (snack bar);
- Visiting a specialty shop;
- Going to the movies;
- Visiting the library; and
- Visiting a restaurant.

In addition, travel distances between the locations mentioned by the subjects are requested. These data enable the calculation of the covariates described earlier.

In the second part of the procedure subjects are asked to perform a scheduling task; that is, they are requested to list all activities they plan to perform the day after the experiment. These activities are all on the list of activities used in the first part, so that detailed information on each selected activity is available. The schedule encompasses the planned activities and the sequence in which they are executed, the locations at which the activities take place, travel modes used, and the start and end times of activities. From these schedules the data used for estimation of both the models was derived. For each observed transition from one activity to another all competing risks, that is, all possible destination activities, were included in the data set. The set of alternative destination activities for a transition encompasses all activities from the list of 31 that were assigned a likelihood greater than 0 in the first part of the procedure. For each competing risk the values of the covariates in the generic and specific model were calculated on the basis of the information supplied in the first part of the procedure. The destination activity that was chosen by the subject was coded as an observed transition; the other competing risks were coded as right censored. The data set consisted of a total of 256 observed transitions and 7,041 right-censored cases.

The study described in this paper is exploratory in nature. A small sample that is not representative of the population of some geographic area as a whole has been used. The sample is homogeneous with respect to age (18 to 25 years), main occupation, and income (students). Therefore, sociodemographic variables are not included in the model. Furthermore, a data set that was collected using this procedure, as detailed information about spatiotemporal constraints on an individual level was obtained, was preferred in this case but is usually not the case with existing time budget and travel surveys.

ESTIMATION AND RESULTS

Estimation Procedure

The models described earlier were estimated using the SAS package. Independence between competing risks and homogeneity was assumed. This implies the following likelihood function:

$$L = \prod_{i=1}^{N} \prod_{a=1}^{A_i} \prod_{c=1}^{C_{ai}} f_c(t_{ia}|x_{iac})^{d_{iac}} S_c(t_{ia}|x_{iac})^{1-d_{iac}} \quad (23)$$

where

- N = number of individuals in sample,
- A_i = number of activities performed by individual i,
- C_{ai} = number of possible exits for activity a of individual i,
- f_c = probability density function of duration times for Exit c,
- S_c = survivor function for Exit c,
- t_{ia} = time at which activity a of individual i is ended,
- X_{iac} = vector of covariates associated with Exit c from activity a of individual i, and
- d_{iac} = dummy variable that indicates whether Exit c was chosen for ath activity of individual i.

As noted earlier, the fact that heterogeneity is not included in the model may affect the scale and shape parameters of the baseline hazard and the estimation of lagged effects. However, the heterogeneity in the sample used for this study is diminished by the fact that the subjects were all students who differed little with respect to sociodemographic characteristics. Further, the effect of heterogeneity on the estimation of state dependence in duration models is less clear, compared with dynamic models based on panel data. Nevertheless, heterogeneity will have some effect, and this should be considered when interpreting parameter estimates.

Generic Model

The generic model was estimated with various distributions assumed for the baseline hazard. The goodness-of-fit measures for the various distributions are indicated in Table 1. As indicated by the goodness-of-fit measures, the log-normal distribution describes the transition probabilities best.

The parameters that were estimated assuming the log-normal distribution are indicated in Table 2. Positive parameter values indicate a positive effect on the duration of the origin activity, whereas negative parameter values suggest a shorter duration. The origin dummies thus suggest that work/education usually has a relatively long duration. The positive effect on the durations of in-home leisure, in-home task activities, and personal activities out of home is smaller, whereas shopping is the activity with the shortest duration. The positive and significant sign of STARTTIME suggests that if an activity starts later, it will have a longer duration. Apparently, the probability of starting a new activity is smaller later in the

TABLE 1 Goodness-of-Fit Measures of Generic Models with Various Distributional Assumptions

distribution	loglikelihood
weibull	-1151.61
exponential	-1163.90
lognormal	-1125.08
loglogistic	-1151.00

TABLE 2 Parametric Estimates of Log-Normal Generic Model

variable name	parameter (t-value)
intercept	-0.08 (-0.18)
in home leisure[1]	0.93 (3.65)
in home task[1]	0.73 (3.08)
work/education[1]	2.22 (8.70)
shopping[1]	-0.08 (-0.24)
pers. act. out of home[1]	0.88 (3.75)
starttime	0.10 (4.02)
tilstart	0.07 (1.33)
tilclose	0.06 (2.70)
prior1	0.05 (2.05)
prior2	-0.08 (-5.71)
travtime	0.27 (9.50)
timespent	0.13 (2.25)
scale	1.60 (22.22)

[1] dummy for origin activity

day. TILSTART does not have a significant effect. TILCLOSE, however, has a significant and positive effect, suggesting that transitions are postponed if more time remains for the destination activity. Thus, if there is less time pressure for the destination activity, the preceding activity will have a longer duration, as one would expect. PRIOR1 has a significant and positive value. This suggests that if the priority of the origin activity is higher, it will have a longer duration. However, a higher priority of the destination activity has the reverse effect, as indicated by the negative value of PRIOR2. Thus, if the destination activity has a higher priority, the preceding activity will have a shorter duration. TRAVTIME has a positive and significant value. Thus, if travel time to the destination activity increases, this will postpone the transition to this activity, resulting in a longer duration of the preceding activity. TIMESPENT, finally, has a positive and significant value: that is, that the more time one has already spent on an activity, the less likely one is to switch to this activity. Apparently, time budgets exist for various types of activities that set limits to the amount of time spent on one activity.

Specific Model

The specific model was also estimated with different assumptions of the distribution of the baseline hazard. The goodness-of-fit measures are indicated in Table 3. Again, the best performance is achieved assuming a log-normal distribution.

TABLE 3 Goodness-of-Fit Measures of Specific Models with Different Distributional Assumptions

distribution	loglikelihood
weibull	-1073.56
exponential	-1106.82
lognormal	-1050.57
loglogistic	-1060.04

The parameters of the log-normal model are indicated in Table 4. In the specific model the origin activity was represented by a set of generic dummy variables; furthermore, an intercept was estimated for each destination activity. The parameters for the origin activity all have a significant positive value, so that the effects can be interpreted only relatively to each other. The estimated values suggest that work/education usually has the longest duration, whereas shopping has the shortest duration on average. The intercepts represent the effect of the destination activity on the duration of the preceding activity. A significant negative parameter value suggests that a transition to personal activities out of home will shorten the preceding activity. In-home task activities, however, are usually postponed, resulting in a longer duration of the preceding activity.

Positive and significant STARTTIME parameters were estimated for work/education and personal activities. Transitions to these activities are thus postponed if the preceding activity starts later. However, the negative value for the end state indicates that transitions to this category are more likely to happen later in the day. A significant parameter for TILSTART was found only for work/education. Thus, activities followed by work/education activities have longer durations if more time remains until this activity can start. The effect of TILCLOSE is significant only for in-home task activities. Contrary to the expectation and to the findings of the generic model, transition to this activity takes place earlier if time pressure is less, resulting in a shorter duration of the preceding activity. Parameters for PRIOR1 were significant only at the 10 percent confidence level for work/education and personal activities out of home. The signs indicate that a transition to work/education takes place earlier if the origin activity has a higher priority, which is contrary to the expectation in this paper. However, the opposite holds for personal activities out of home. The positive sign indicates a longer duration of the preceding activity if the priority is higher. Positive and significant parameters for PRIOR2 were found for in-home task activities, work/leisure, and shopping. If the priority of these activities increases, transitions to these activities will take place earlier, resulting in shorter durations of the preceding activities. A positive and significant parameter value for TRAVTIME was found for all activities, except the end state. Apparently, all are postponed if travel time increases. This effect is strongest for in-home leisure and relatively weak for work/education and in-home task activities. This indicates a weaker distance decay for obligatory activities, as one would expect. Significant parameters for TIMESPENT were found for work/education and personal activities out of home, indicating that transitions to work/education and personal activities out of home are postponed if more time already has been spent on these activities. This holds to an extreme extent for personal activities. Apparently, there are strict time budgets for personal activities. So, with few exceptions (PRIOR1 and TILCLOSE) the parameter signs are in line with common sense.

COMPARISON OF GENERIC AND SPECIFIC MODEL

In terms of interpretation of the parameters, the two models are largely consistent. Shifts in the sign of parameters appeared only in the case of TILCLOSE and PRIOR1. However, parameter values vary in sign and magnitude between destination states in the specific model, indicating that different types of activities are planned according to different criteria. Therefore, the extension from a generic model to a specific model is regarded as an important improvement. To test whether the specific model performed statistically better than the generic model, a likelihood ratio test was performed. The chi-square statistic of 149.02 with 35 degrees of freedom indicates that the specific model performs significantly better at $\alpha = 0.005$.

CONCLUSION

In this paper an alternative method of modeling activity choice, timing, and duration has been described. Competing risk hazard models of the accelerated time type were used to describe the duration of an activity, the choice of a next activity, and their mutual dependency. The estimated models performed satisfactorily, suggesting that competing risk models are a useful tool for incorporating duration dependence into discrete choice modeling. This conclusion is particularly relevant as timing of activities and trips

TABLE 4 Parametric Estimates of Log-Normal Specific Model

	DESTINATION ACTIVITIES					
	in home leisure	in home task	work education	shopping	pers. act. out of home	end state
in home leisure[1]	2.21 (13.23)[2]					
in home task[1]	1.97 (13.83)					
work education[1]	3.49 (21.17)					
shopping[1]	1.08 (4.82)					
pers. act. out of home[1]	2.34 (16.69)					
intercept for destination	-3.23 (-1.21)	1.93 (2.03)	-2.01 (1.85)	-1.32 (-0.75)	-2.34 (-3.66)	4.62 (3.70)
starttime	0.20 (1.09)	-0.11 (-1.83)	0.31 (4.25)	0.22 (1.85)	0.14 (2.98)	-0.55 (-6.88)
tilstart	0.21 (0.81)	0.09 (0.91)	0.35 (2.06)	--	0.12 (1.78)	--
tilclose	0.22 (1.24)	-0.17 (-3.40)	0.05 (0.64)	0.11 (1.22)	-0.02 (-0.41)	--
prior1	-0.03 (-0.54)	0.06 (1.39)	-0.15 (-1.93)	0.03 (0.23)	0.06 (1.84)	0.05 (0.60)
prior2	0.00 (0.56)	-0.16 (-4.06)	-0.11 (-2.59)	-0.25 (-3.29)	-0.05 (-1.93)	--
travtime	0.41 (3.52)	0.15 (2.46)	0.10 (1.77)	0.24 (2.53)	0.26 (6.77)	-0.04 (-0.54)
timespent	0.08 (0.28)	0.14 (1.08)	0.27 (2.32)	--	15.51 (4.06)	--

scale = 1.34 (24.81)

[1] dummy for origin activity
[2] t-values in parentheses

becomes the subject of policy making to reduce congestion and preserve mobility.

In this application two models were estimated: a generic model and a specific model, which is conditional on the destination state. The specific model was superior to the generic model because the goodness of fit of this model was significantly higher. Furthermore, the estimated parameters reflected differences between scheduling criteria for different activity types, which would remain unrevealed in the generic model. Parameter estimates suggest that spatiotemporal constraints such as time of day, opening hours, and travel time play an important role in activity scheduling and timing. Also the history of the pattern and priorities of activities influence timing and choice of activities. The inclusion of these covariates was enabled by a specific computerized data collection procedure, which provides an extensive record of individual activity scheduling processes. Application of such a procedure can be considered a prerequisite if one wants to obtain models that are capable of describing travel behavior at a detailed level. Both the generic and the specific models were estimated by assuming different specifications of the baseline hazard. Of the tested distributions, the log-normal distribution provided the best fit in predicting activity transitions.

The modeling approach described in this paper is a first step in a new direction of modeling and simulating the performance of activity patterns. However, improvements still need to be made. An issue already raised is how heterogeneity and interdependency of risks should be handled. In this study flexibility of the model structure was allowed to prevail over heterogeneity and interdependency. Future research, however, should address possible ways of unifying the above properties into one model structure. Another development would be the extension of the unconditional competing risk models described above to conditional models, where transition probabilities are dependent on both origin and destination states.

REFERENCES

1. Jones, P. *New Developments in Dynamic and Activity-Based Approaches to Travel Analysis.* Gower, Aldershot, England, 1990.
2. Algers, S. Integrated Structure of Long-Distance Travel Behavior Models in Sweden. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., 1993, pp. 141–149.
3. Kitamura, R., and M. Kermanshah. Sequential Model of Interdependent Activity and Destination Choices. In *Transportation Research Record 987,* TRB, National Research Council, Washington, D.C., 1984, pp. 81–89.
4. Kitamura, R. Incorporating Trip Chaining into Analysis of Destination Choice. *Transportation Research,* Vol. 18B, 1984, pp. 67–81.
5. Timmermans, H., and P. Van der Waerden. Modelling Sequential Choice Processes: The Case of Two-Stop Trip Chaining. *Environment and Planning A,* Vol. 24, 1992, pp. 1483–1490.
6. Recker, W. W., M. G. McNally, and G. S. Root. A Model of Complex Travel Behavior: Part 2: An Operational Model. *Transportation Research,* Vol. 20A, 1986, pp. 319–330.
7. Ettema, D., A. Borgers, and H. Timmermans. Simulation Model of Activity Scheduling Behavior. In *Transportation Research Record 1413,* TRB, National Research Council, Washington, D.C., *1413,* 1998, pp. 1–11.
8. Heckman, J. J. Statistical Models for Discrete Panel Data. (C. F. Manski and D. McFadden, eds.), In *Structural Analysis of Discrete Data with Econometric Applications* MIT Press, Cambridge, Mass., 1981.
9. Meurs, H. *A Panel Data Analysis of Travel Demand.* University of Groningen, The Netherlands, 1991.
10. Niemeier, D. A., and J. Morita. Duration of Trip-Making Activities by Men and Women: A Survival Analysis. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1994.
11. Lawless, J. F. *Statistical Models and Methods for Lifetime Data.* Wiley, New York, 1982.
12. Lancaster, T. *The Econometric Analysis of Transition Data.* Cambridge University Press, England, 1990.
13. Kalbfleisch and Prentice. *The Statistical Analysis of Failure Time Data.* Wiley, New York, 1980.
14. Popkowski Leszczyc, P. T. L., and H. J. P. Timmermans. Consumer Store Choice Dynamics: An Analysis of the Competitive Market Structure for Grocery Stores. *Environment and Planning A* (in press).
15. Han, A., and J. A. Hausman. Flexible Parametric Estimation of Duration and Competing Risk Models. *Journal of Applied Econometrics,* Vol. 5, 1990, pp. 1–28.
16. Hensher, D. The Timing of Change for Automobile Transactions: A Competing Risk Multispell Specification. Presented at 7th International Conference on Travel Behavior, Valle Nevado, Santiago, Chile, 1994.
17. De Jong, G., R. Kitamura, and J. Klooster. A Disaggregate Model of Vehicle Holding Duration, Type Choice and Use. Presented at 7th International Conference on Travel Behavior, Valle Nevado, Santiago, Chile, 1994.
18. Hensher, D. A., and F. L. Mannering. Hazard-Based Duration Models and Their Application to Transport Analysis. *Transport Reviews,* Vol. 14, 1994, pp. 63–82.
19. Sueyoshi, G. T. Semiparametric Proportional Hazards Estimation of Competing Risk Models with Time-Varying Covariates. *Journal of Econometrics,* Vol. 51, 1992, pp. 25–58.
20. Ridder, G. The Non-Parametric Identification of Generalized Accelerated Failure-Time Models. *Review of Economic Studies,* Vol. 57, 1990, pp. 167–182.
21. Ettema, D., A. Borgers, and H. Timmermans. Using Interactive Computer Experiments for Identifying Activity Scheduling Heuristics. Presented at 7th International Conference on Travel Behavior, Valle Nevado, Santiago, Chile, 1994.
22. Ettema, D., A. Borgers, and H. Timmermans. Using Interactive Computer Experiments for Investigating Activity Scheduling Behaviour. *Proc., PTRC 21st Summer Annual Meeting,* University of Manchester, England, 1993, pp. 267–282.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Effects of Different Data Collection Procedures in Time Use Research

NELLY KALFS

A field that might play an important role in the future of travel demand analysis and modeling is time use research, although some issues need to be resolved. One such issue deals with the data collection procedure. To provide guidelines to researchers, the strengths and weaknesses of three data collection systems are reviewed. One system relied on the traditional paper-and-pencil diary; another system was a self-administered electronic diary (computer-assisted self-interview, or CASI), and the third was based on an interviewer-administered electronic procedure (computer-assisted telephone interview, or CATI). These systems are compared in terms of the validity of the time use statistics, the unit response rate, and time involved in conducting the survey. The results show that none of the data collection systems is best in all aspects: the unit response rate is highest in CATI, and the time to conduct the survey is lowest in CASI. As far as the validity is concerned, one method was not found to be best for all activities. The comparison clearly shows that relatively large differences exist among the procedures. Consequently one must be careful using the results of studies that are based on different data collection systems. One specific activity that is of increasing interest to policy makers, that is travel, is illustrated.

One of the challenges facing the field of time use research today is to clearly show its usefulness in guiding policy. A policy issue that can be addressed with time use data is travel behavior. According to earlier reports (*1,2*) time use research could play an important role in the development of new travel demand models (especially from the perspective of the activity-based approach), although there are some issues yet to be resolved. One of these issues is related to the fact that many researchers who work in the time budget field prefer to measure time use by means of a diary. In this approach respondents are asked to report their activities for at least 24 hr chronologically. Traditionally this is done with paper and pencil.

The use of a diary is preferred, because this type of data collection technique is expected to produce the most accurate results [see, for instance, work by Juster (*3*), Niemi (*4*), and Gershuny and Robinson (*5*)]. However, diary surveys are expensive, and they demand a lot of time, from both respondents and researchers, who fill in the diary and process the data, respectively. These disadvantages and the fact that advances in computer technology have changed the methods of data collection have led to the development of an alternative diary form; the electronic diary. This electronic diary resembles the traditional one to a great extent, but the coding of activities is done by a computer-assisted tree-structured questionnaire (*6,7*). It was expected that this form of diary avoids many of the disadvantages of the hand-written diary.

In the first place, the use of a computer in data collection can considerably reduce the amount of work. A second advantage is that coding is automatic. Another presumed advantage of computer-assisted data collection is that it can improve data quality, if careful attention is given to automatic branching and coding, consistency checks, and help screens (*8,9*).

Until now little evidence has been given for these general findings. Although the literature on (time use) data collection mode comparisons is extensive (*10–15*), comparisons with computer-assisted data collection are rare. Therefore, a comparison was made between the electronic and the conventional paper-and-pencil procedure (paper-and-pencil interview, or PAPI). One environment was a self-registration method (computer-assisted self-interview, or CASI) and the other an interviewer-administered procedure (computer-assisted telephone interview, or CATI). In this study these data collection procedures are compared in terms of the validity of the time use statistics, the unit response rate, and time to conduct the survey.

The simplest measure to establish is the unit response rate because the maximum number of possible participants is known or is possible to estimate. The time to conduct the survey is the most difficult criterion, as the design of the surveys differed considerably, and some time components were unknown. Therefore, attention is paid to only two aspects: interviewers' time and assistance; and coding and editing time. These aspects reveal, for the surveys compared here, how much time can be gained by using the computer-assisted interviews instead of the PAPI.

Validity is measured indirectly, because it is not normally known what real time use is actually like (*15*). Juster (*11*) formulated criteria that implicitly assume that greater detail in reporting and the ability to account for time lead to more valid reporting. Besides this, the number of mistakes made by respondents and coders was evaluated.

For the purposes of this comparison, a strict experimental design was not used to evaluate the effect of the various aspects of the data collection modes. In the first place, the designers had no influence on some of the design characteristics. Second, systematic variation in characteristics does not appear to be practical for the specific surveys that were considered. Therefore, three existing procedures designed to be as efficient as possible given the mode of data collection were compared. Consequently, the data collection method is not the only factor that differs among the procedures: there are also differences in the methodology of the diaries (report and coding of activities and time and the kind of information requested) and the implementation aspects of the surveys (selection of households, individuals, and days).

In the next section an overview of the characteristics of the time use diaries and the implementation features of the three surveys are presented. Because the interest of this study was mainly in the influence of the registration method, some methodological issues are discussed briefly. These issues concern the corrections that have to be made before possible effects of the data collection procedure could be examined more clearly. Then a summary of the results of the comparison is given, and finally the impact of the differences

Ministry of Transport, Public Works and Water Management, Transport Research Centre, P.O. Box 1031, 30000 BA Rotterdam, The Netherlands.

between the data collection procedures for the time use on one specific activity, travel, is illustrated.

METHODOLOGY OF DIARIES

The information gathered with a diary "can show for an individual what activities were done during the defined period, how many times, in what order, at what time, for how long, where and other objective and subjective information connected with the activities" (*16*).

The diary itself can be designed in many different ways. The activity categories may be precoded or open, the time interval may be fixed (periods of 5, 10, 15, or 30 min are the most common) or open (asking until what time an activity lasted), the activity code itself can be varied, and the diaries may provide space for recording only one (primary) or multiple simultaneous (primary and secondary) activities (*17,18*).

It is difficult to tell which design is preferred. The most important problem seems to be the balance between the task of the respondent and the processing of the data (*10,19*). This problem has been the major reason for designing an electronic diary. In the opinion of the author an electronic diary is less demanding for the respondent and it handles the data processing very well. However, before the features of the electronic diary are outlined, the design of the PAPI diary that was used for the comparison will be presented.

PAPER-AND-PENCIL INTERVIEW

Between 1987 and 1988, the Netherlands Central Bureau of Statistics conducted time use research. On the basis of a pilot study from 1986 in which several methodological variants of the design were examined on response rate, selectivity of the response, and data quality, a design was chosen for the main survey (*8*).

The diary design included the following features: for registering activities the closed variant was chosen with a precoded list of activities. In total, the list contained 106 activities divided into nine groups. The activities were registered in fixed intervals of 15 min, and secondary activities were ignored. Thus, the respondents had to specify their activities and code them for each interval of 15 min. Next to the code, a verbal description of the activity could be given. In the events that concurrent activities were performed during a certain interval, the respondent had to choose either the activity that had lasted more than 7 min or the so-called productive activity. For the productive activities, supplementary questions had to be answered in the diary.

COMPUTER-ASSISTED SELF INTERVIEW AND TELEPHONE INTERVIEW

The electronic diary was developed by the Sociometric Research Foundation, and the main purpose of the study performed with this diary was the registration of time use. The diary is characterized by the following features: activities are recorded by answering a tree-structured questionnaire, the time interval is open, and activities that last for 10 min or more—with the exception of travel—and secondary activities have to be reported. In total, 368 activities are distinguished.

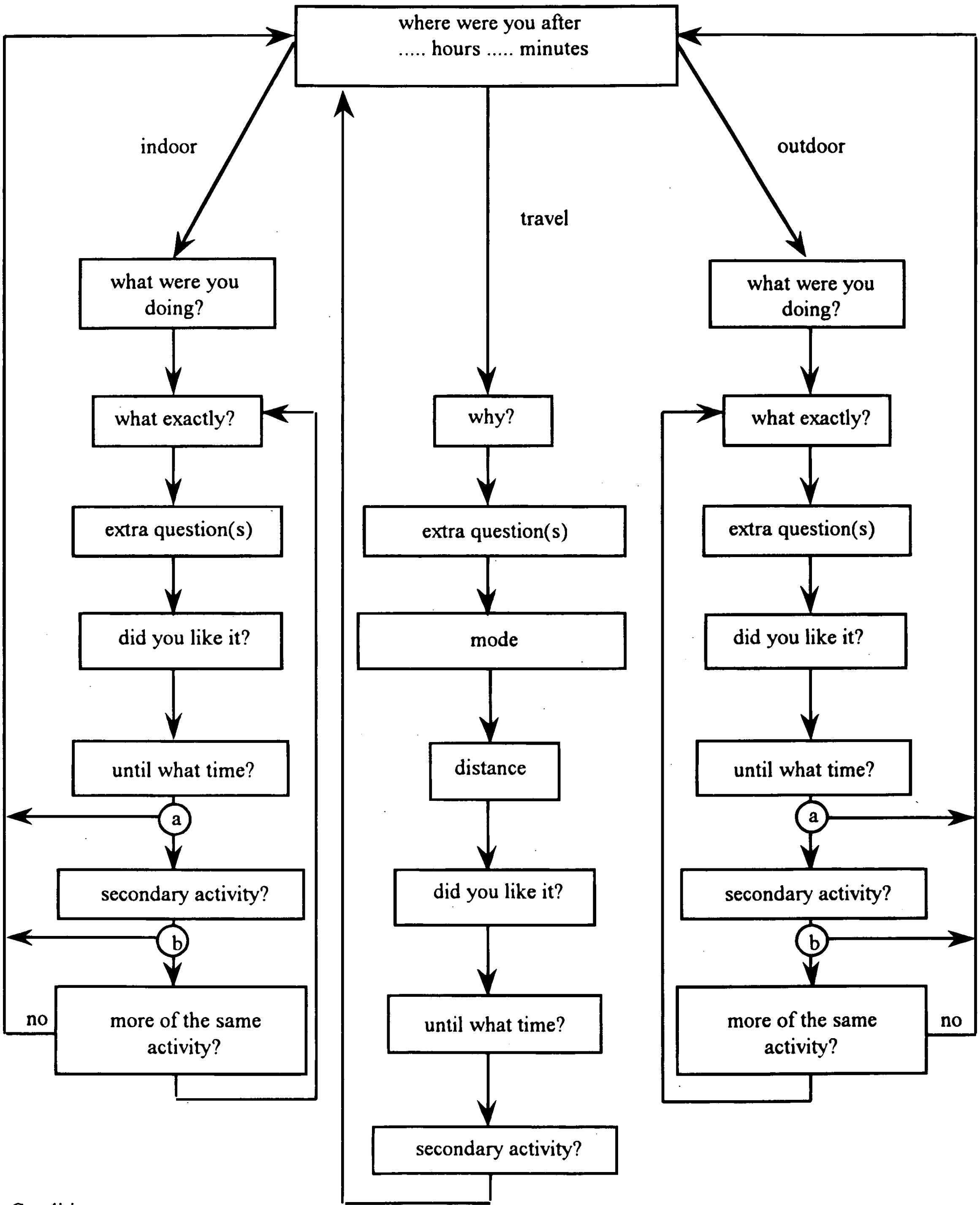
What is meant by "activity coding by answering a tree-structured questionnaire" can best be illustrated by showing part of the questionnaire (Figure 1). First, the respondent has to choose between a number of main categories. Subsequently, the activity is recorded in more detail. The code of the activity is generated automatically as it consists of the sequence of chosen answers in the tree. This procedure leads to a unique code for each activity.

If, for example, the activity is "preparing supper," the activity code is 213: first the respondent chooses the main category "running the household" (2), second "preparing food" (1), and then "supper" (3). It is obvious that this procedure saves the respondent and the research organization a lot of time with respect to coding and that more activities can be distinguished than in a coding list because the task for the respondent is easier: answering questions rather then referring to a list.

An overview of the questionnaire is presented in Figure 2. In this figure the central question in the electronic diary is: Where were you after . . . hours . . . minutes? The answer to this question determines whether indoor, outdoor, or travel activities are shown. Next, the activity is registered in detail.

What were you doing?	What exactly?	Further?
1. job	1. main job	
	2. second job	
	3. unpaid job	
2. running the household	1. preparing food	1. breakfast
		2. dinner
		3. supper
		4. other
	2. washing up	
	3. making tea or coffee	
	4. cleaning the house	1. dust
		2. etc., etc.
3. personal care	1. taking a shower	
	2. etc., etc.	
4. etc. etc.		

FIGURE 1 Part of electronic questionnaire.



Conditions:
a: all activities, excluding night sleep
b: if the reported activity is doing the housework, obtaining goods/services, work/job, media activities

FIGURE 2 Flow chart of electronic questionnaire.

Specific additional information on the activity can be easily asked and may apply to all activities or to a specific kind of activities (for the latter see, for instance, the questions asked about travel in Figure 2). For all activities, the respondents were asked whether they like performing the activity, and for nearly all activities, whether they performed another activity concurrent with the primary activity. The amount of inadequate information in the electronic diary is kept to a minimum by having immediate checks on the upper and lower boundary of answers, the time measure, and the sequence of activities, for example, report of travel between indoor and outdoor activities.

Two drawbacks of the electronic diary are that it uses an open-time interval and that respondents are not provided with a list of activities. These could lead to confusion about the level of detail expected in the diary (*19*). To compensate, an example was given as an introduction to the diary exercise and an extra question was placed after each activity to check whether the answer given represented the activity performed and its duration. Possible mistakes could be reported in answer to this question and in the response to an open question at the end of the questionnaire. To get an accurate account of time use, if the duration of the reported activity exceeds a certain length of time (4 hr or more where "work" or "job" is reported, and 3 hr or more for all other activities), the respondent is asked whether he or she is certain that all activities have been reported.

With this electronic diary, data can be gathered in different ways; the questionnaire can be filled out by either a respondent or an interviewer. In this research, data were collected both ways. The same questionnaire was used for both procedures, and therefore the coding structure of the activity and time, the recording of the primary and secondary activity, and the extra information requested about the activities were all the same. In this paper these two forms are compared with the PAPI diary. But first the implementation characteristics are discussed.

IMPLEMENTATION OF DIARIES

In principle, the selection scheme for obtaining a sample of diaries that fully reflect the time use for activities in the population is as follows: first, a probability sample of households must be selected; second, individuals within the household must be determined at random; and, finally, the dates for which time diaries are to be filled out by each respondent must be selected by a probability method (*20*). The most important decisions about these aspects follow.

Paper-and-Pencil Interview

For the PAPI survey a sample of households was randomly drawn from an address directory. All members of these households who were at least 12 years of age were requested to keep a diary for 2 consecutive days. These days were determined at random by interviewers according to a scheme.

One person per family was interviewed face-to-face on characteristics of the household. The remaining information was collected in writing and collated later by the interviewer. The respondents were asked to record their activities as soon as possible after the activities took place. To boost response, each household was offered a gift.

The survey was held over 2 years. As a result, the sample of addresses was spread over the year to make it possible to present results for periods of less than 1 year. For the purposes of this comparison, data were provided for only one member of a household (where possible, randomly selected) who responded in October and November 1988.

Computer-Assisted Self Interview

One of the environments in which the electronic diary was used was the NIPO Telepanel. This panel consists of a sample of about 1,000 households that were randomly drawn from an address directory (the same one that was used for PAPI). The NIPO has provided these households with a computer and modem, and with these facilities data are collected entirely automatically (*9,21,22*).

The time use survey in the Telepanel had to be conducted within 1 week in the 1988. One week in November was chosen because earlier research has shown that time use in the period between October and November is closest to the annual average (*23*). At the end of October, all individual Telepanel members 12 years and older were asked whether they wanted to participate in the survey. To encourage them, a lottery with one prize was announced. Of the willing respondents, only one person per household was randomly selected. This person was asked to fill in the diary for three different days. A fixed scheme of assigning days to persons was used and respondents were given notice in advance of the days for which they had to report their activities. It was not a requirement that Telepanel respondents should fill in the electronic diary several times a day because during that time the computer would have to be switched on. Instead they were asked to fill in their diaries as soon as possible after each given day.

Computer-Assisted Telephone Interview

The other environment in which the electronic diary was implemented was a centrally administered telephone survey conducted at the University of Amsterdam. Interviews were conducted 7 days a week during the afternoon and evening in October and November 1988. The sampling source consisted of a telephone directory. Again one person per household aged 12 years or older was randomly selected to report his or her activities for 3 days. During the first call the respondents were asked about their previous day's activities and about their demographic characteristics. Afterwards, attempts were made to make appointments for interviews 2 and 4 days later. If the respondent could not be contacted on these days, the interviewer tried to make appointments for 3 and 5 days later for the same diary day as before. If even this was impossible, another diary day was selected because it was felt that the recall period would otherwise become too long.

These data clearly indicate that dissimilarities in the implementation exist. Consequently, the registration process is not the only factor that differs between the procedures. To see if time use differences resulted from the registration method, some methodological issues had to be resolved. This is the topic of the next section.

METHODOLOGY OF COMPARISON

Before it is possible to concentrate on whether differences exist between the data collection procedures, two methodological issues need to be discussed briefly. The first one relates to the classification of the activities, and the second one relates to weighting.

To allow a comparison between the data collection systems, the lists of activities had to be brought into line with each other. To do this, the classification schemes were judged by four different researchers; after discussion they came up with one scheme of 75 activities for the different surveys. For the evaluation of the time use estimates, the activities were finally grouped together even more (29 activities) because many activities were performed infrequently, which hampered the comparison.

A second important issue is that it is advisable to correct the data for differences between the procedures with respect to implementation aspects. The most important aspects are assumed to be nonresponse and the sampling source. The coverage of the population may be selective and incomplete and may therefore bias the results. To examine whether the coverage is selective or incomplete, researchers often search for variables that are expected to be strongly related to the topic of the research and for which the population distribution as well as the distribution for the respondents is known. If differences between the sample and population distributions are found, the sample results can be weighted to adjust for a possible bias. In this study, weighting was done in such a way that all days of the week were equally covered in the three samples and that the distribution of certain demographic variables that are assumed to be strongly related to time use, perceived position on the labor market, age, gender, marital status, and urbanization in the samples corresponded to population estimates. By applying this weighting, it was possible to examine more clearly the influence of the data collection procedure on the time use.

RESULTS

The most salient finding of this research is that none of the procedures emerges as the outright winner in all respects:PAPI achieves the highest amount of detail (in PAPI, respondents reported on average 22 primary activities, in CASI 20, and in CATI 19), but in CATI there are fewer mistakes (in CATI 2 percent of the total time consisted of mistakes; in CASI, 4 percent; and in PAPI, 7 percent) and a higher response (CATI, 52 percent; PAPI, 45 percent; and CASI, 38 percent). The cost in terms of time is favorable in CASI (CASI, 120 h; CATI, 1,240 hr; and PAPI, 1,460 h), whereas social desirability has less of an effect in PAPI and CASI. For more details about these results we the reader is referred to other work by Kalfs (*24*).

Further, it was found that each of the data collection procedures gives more valid results than the others for certain specific kinds of activity. The registration of routine activities is best in PAPI because in this procedure the recall period is short and respondents have the time to think about their time use thoroughly and are confronted with them in a list. Activities that may lead to a lack of understanding on the part of the respondent (for instance travel, child care, and social contacts) are best reported in CATI, because a well-trained interviewer is in a position to significantly improve data quality by posing additional questions and clearing up misunderstandings. On the other hand, the interviewer can also have a negative effect on the validity because the respondent may try to create a positive impression of him or herself on the interviewer by reporting socially desirable activities. This was found to be true for the time people spent watching television: in CATI a tendency was found among more highly educated respondents to answer in a socially desirable way, leading them to underestimate the time they spent on this activity. To get an idea about the impact of the differences, one activity, travel, will be discussed in detail.

One of the activities for which the highest effect of the data collection procedure was found on time use was travel. In Table 1, the deviations between the data collection systems, in terms of the time use statistics that are normally used (mean total time, mean participation rate, and mean participation time), are relatively large.

In Table 1 time use statistics were also included from another study, the "trip diary survey" for 1988, to validate the results of the various data collection procedures. In the trip diary survey (TDS), a one-day PAPI diary is used in which respondents have to fill in all the trips they make during a designated day. Comparison of the results of TDS with the outcomes of the time use surveys reveals that the participation rate in CASI and CATI corresponds to the rate in TDS. In PAPI the participation rate is much lower. Examination of the participation time indicates that about equal estimates are found for TDS, PAPI, and CATI, but the CASI estimate is much higher.

The results suggest that the time use on travel is underreported in PAPI, given the lower participation rate and is overreported in CASI, given the higher participation time. The question then is:

TABLE 1 Time Use Statistics for Travel

Data Collection Procedure / Time Use Statistics [d]	Mean Participation Rate [a] (%)	Participation Time [b] (Minutes)	Total Time [c] (Minutes)
PAPI (n=954)	57 (2)	86 (4)	49 (3)
CASI (n=1785)	88 (1)	107 (3)	94 (3)
CATI (n=1208)	91 (1)	84 (2)	76 (2)
Trip Diary Survey (n=21300)[e]	88 (.)	80 (.)	70 (.)

( ) Unweighted Standard Error (.) Unknown Standard Error

a Participation Rate = percentage participating per day for all respondents

b Participation Time = time use in minutes per day for participating respondents

c Total Time = time use in minutes per day for all respondents

d It is possible that participation rate x participation time ≠ total time due to rounding

e CBS (1989). *De mobiliteit van de nederlandse bevolking*. Centraal Bureau voor de Statistiek, Voorburg.

How can we explain the lower participation rate in PAPI and the higher participation time in CASI?

The lower estimate for the participation rate in the PAPI procedure is probably generated by the time interval and the lack of a check on travel. Because of the fixed interval of 15 min, travel of short duration was not reported in the PAPI diary, whereas these activities had to be reported in the electronic diary, which included a check on travel. It was thought that the lack of a check on travel would strongly influence the time use on travel concerning shopping and leisure (especially entertainment) because this kind of trip is often not viewed (and thus not reported) as trip making by respondents (*6,25*).

The time use for travel seems to be overestimated in CASI because of the mistakes respondents made. A closer inspection of the data showed that some respondents in CASI had made mistakes with respect to shopping and leisure time trips (B. Bosch, unpublished data, 1991). Diaries were found in which only one trip was reported, when in fact three activities were expected: two trips—one from home to another location and vice versa—and the activity performed away from home. When this kind of mistake was made, the time use for the activity performed between two trips was added to the time use on travel. It is possible that more respondents have filled in the diary in this way than the ones who reported that such a mistake was made.

These explanations for the deviations in the time use for travel in PAPI (time interval, omission of certain trips) and CASI (omission of certain activities performed between two trips) were examined further. The influence of the time interval was estimated on the basis of the total time in CASI and CATI for travel activities of short duration. The missing trips in PAPI were reconstructed in two ways. First, the activity performed was determined when a change in the activity's location took place. Second, the total time for travel was divided by its purpose on the basis of major clusters of activities (11 categories). The purpose of a trip was inferred from the activity at the origin or the destination of that trip (*25,26*). Besides this information, data were available in CASI and CATI with respect to the distance traveled; this information made it possible to examine whether the duration per distance unit traveled (1 km) differed between CASI and CATI or, in other words, in CASI whether the time spent on the activity performed between two trips could have been added to the time use for travel.

In the first place, an estimation was made of the total time involved in trips of short duration. Calculated in minutes, the total time involved in trips of short duration was only 3 to 4 min. This result illustrated that the time interval can explain only some of the large differences in the total time among the three procedures.

More support was found for the idea that certain trips were not reported in the PAPI diary. The distributions of activities before or after a change in place (moving back to or away from the dwelling) indicated that, overall, in PAPI, travel was reported in only 36 percent of the activities. Activities that are realistic alternatives for travel (active leisure, such as walking or cycling for recreational purposes or walking the dog) consisted of 6 and 10 percent, respectively.

If these estimates are taken together, it is found that in 52 percent of all changes in the activity's location trips were reported and in 48 percent trips were not reported. This 48 percent mainly consisted of everyday needs (25 percent), visit/party, and similar social activities (21 percent), other shopping (15 percent), and working time (11 percent). Thus, trips with respect to these latter activities were underreported in PAPI. Further evidence for this statement is provided by the data in Table 2, in which the total time for travel is distinguished by the purpose of the trip. Moreover, if the CATI estimates are taken as criteria, one can see that the time use for shopping and entertainment trips in particular are underestimated in PAPI. The trips related to these activities are also underestimated in CASI, but the differences between CASI and CATI are smaller.

Looking now at CASI, a more striking result was found with respect to the category "traveling around." This activity concerns trips that are undertaken without a specific purpose of the trip. This estimate is much higher for CASI than for PAPI and CATI. It supports the idea that respondents in CASI tend to forget to report the activity they performed between two travel activities.

TABLE 2 Mean Total Travel Time for Specific Activities

Mean Total Time	PAPI (n=954)	CASI [a] (n=1785)	CATI [a] (n=1208)
Activity	Minutes		
Work	13	15	15
Domestic activities	1	3	3
Child care	1	1	1
Shopping	5	7	13
Personal needs	4	5	6
Education	7	7	6
Organisations	1	3	3
Entertainment	8	12	16
Active leisure	2	5	6
Passive leisure	2	1	1
Travelling around	6	29	5
Total[b]	49	87	75

[a] The total times for CASI and CATI are corrected for mistakes

[b] Details may not add to totals due to rounding

TABLE 3 Mean Duration for Travel Activities (min/km)

Duration in Minutes / Direction of Travel	CASI Mean	CATI	Difference
At home → travel → not at home CASI (n=1859), CATI (n=1685)	5.8 (0.3)	6.0 (0.2)	-0.2
Not at home → travel → not at home CASI (n=548), CATI (n=548)	6.1 (0.4)	5.5 (0.2)	0.6
Not at home → travel → at home CASI (n=1873), CATI (n=1645)	5.5 (0.3)	6.1 (0.2)	-0.6
At home → travel → at home CASI (n=792), CATI (n=157)	38.0 (2.0)	16.5 (1.4)	21.5

( ) Unweighted Standard Error

This possible effect was examined further by calculating the mean duration per kilometer for various situations in which travel was reported. The results are indicated in Table 3 and clearly support the existence of this effect. In the table, one can see that large deviations between the two procedures are found for only one category: at home → travel → at home. The mean duration per kilometer for this category in CASI is more than twice as high as the estimate in CATI. Thus, it appears that in this situation the time use for the activity performed between two travel activities is added to the time spent on travel.

In general, Table 3 results show that in PAPI as well as in CASI, trips associated with shopping and entertainment are particularly underreported. Consequently, the time spent on activities for which the trips are missing is probably overestimated in PAPI. The situation in CASI is different; despite underreporting of shopping and leisure time trips, the time use for travel is overestimated primarily because of the time spent on traveling around. Most of the time use for this latter activity consists of the time spent on shopping and entertainment trips, and the time actually spent on these activities. Therefore, if the total time for shopping and entertainment is added to the total time for travel, one would expect roughly equal results for PAPI and CASI. This is indeed the case, as one can see in Table 4; the overall estimate for PAPI is 168 min, and for CASI it is 169 min.

In summary, it is concluded that the time spent on travel is underreported in PAPI because of a lower participation rate. The overreporting in CASI is because of a higher participation time. This under- and overreporting is mainly caused by the same factor: a problem in correctly classifying trips related to shopping and entertainment. In PAPI, respondents forgot to enter these kinds of trips in the diary. In CASI, however, trips associated with these activities were reported because there was a check on travel in the electronic diary, but at the same time the shopping and entertainment activities were themselves omitted relatively frequently. The CATI interviewer was well trained to clear up misunderstandings with respect to these activities and therefore less-biased estimates were obtained.

CONCLUSION

In this paper, two forms of an electronic diary have been compared with the traditional PAPI. These diary forms are compared in terms of the validity of the time use statistics that are normally used, the unit response rate, and the time involved in conducting the survey. The results show that there is no single method that is best in all aspects. Each of the data collection procedures examined here reveals aspects for which better results are obtained if the results of the other procedures are compared. The same result was found with respect to the time spent on specific activities, which would suggest that one should use different methods if one is interested in the time use for all kinds of activities. However this use would be both expensive and impractical. Thus, the question remains: Which method should be chosen if one has to rely on only one method?

Of course improvements in the questionnaire are still possible, particularly in the electronic diary. Given the large differences between the procedures with respect to the time involved in conducting the survey, which especially favors CASI, it is clearly worthwhile to put extra effort into improvements. To start with, sound training of the respondent, especially for problematic activi-

TABLE 4 Mean Total Time for Shopping, Entertainment, and Travel

Mean Total Time / Activity	PAPI (n=954) Minutes	CASI (n=1785)
Shopping	36 (2)	15 (1)
Entertainment	83 (4)	60 (3)
Travel	49 (3)	94 (3)
Total	168 (5)	169 (5)

( ) Unweighted Standard Error

ties, can help every data collection method to provide better data. Besides this kind of assistance, it is also possible to improve the precision of time use reporting in the electronic diary by implementing additional checks on activities and time reports. However, it is difficult to see any way of avoiding bias because of social desirability. In general, therefore, if more effort were made in the design of the diary and if equivalent guidelines were applied to each diary, it would be possible to get data of comparable quality from PAPI and CASI, although some problems still remain in the case of CATI.

Although the analyses have touched on only three particular data collection procedures within one specific topic, and although they have been mostly concerned with aggregated measures, there is much more to be learned from data collection procedure comparisons in which computers are used, especially given the technological advances being made nowadays. Therefore computer-assisted data collection is a topic to research more closely in the years to come. Travel behavior is another topic for future research.

REFERENCES

1. Pas, E. I. Emerging Directions in Travel Demand Modeling: a New Role for time use Studies? In *Fifteenth Reunion of the International Association for Time Use Research:* (N. Kalfs and A. S. Harvey, eds.), NIMMO, Amsterdam, The Netherlands, 1994, pp. 19–30.
2. Pas, E. I., and A. S. Harvey. Time Use Research: Implications for Travel Demand Analysis and Modelling. In *Understanding Travel Behaviour in an Era of Change* (P.R. Stopher and M. Lee-Gosselin, eds.), Pergamon Press (in press).
3. Juster, F. T. *The Allocation of Time: Empirical Findings, Behavioral Models, and Problems of Measurement.* Survey Research Center, University of Michigan, Ann Arbor, 1990.
4. Niemi, I. Systematic Error in Behavioural Measurement: Comparing Results from Interview and Time Budget Studies. Presented at Meeting of the International Association for Time Use Research, Madrid, Spain, July, 1990.
5. Gershuny, J., and J. Robinson. Estimating Paid Work Time. Presented at Meeting of the International Association for Time Use Research, Rome, Italy, June, 1992.
6. Kalfs, N. *Hèt Construeren van meetinstrumenten voor quasi collectieve voorzieningen en huishoudelijke productie.* Research Memorandum 861117. Sociometric Research Foundation, Amsterdam, The Netherlands, 1986.
7. Verwey, M., N. Kalfs, W. E. Saris, and M. de Pijper. Tijdsbestedingsonderzoek middels Tele-Interviewing en de mogelijkheid voor segmentatie. *Massacommunicatie,* Vol. 4, 1987, pp. 353–369.
8. Kiesler, S., and L. S. Sproull. Response Effects in the Electronic Survey. *Public Opinion Quarterly,* Vol. 50, 1986, pp. 402–413.
9. Saris, W. E. *Computer Assisted Interviewing.* Sage University Papers, Beverly Hills, Calif., 1991.
10. Lingsom, S. *Advantages and Disadvantages of Alternative Time Diary Techniques. Working Paper.* Statistisk Sentralbyra, Oslo, Norway, 1980.
11. Juster, F. T. Response Errors in the Measurement of Time Use. *Journal of the American Statistical Association,* Vol. 81, No. 394, 1986, pp. 390–402.
12. Sykes, W., and M. Collings. Effects of Mode of Interview: Experiments in the U.K. In *Telephone Survey Methodology* (R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicolls II, and J. Waksberg, eds.), Wiley, New York, 1988, pp. 301–320.
13. Groves, R. M. *Survey Errors and Survey Costs.* John Wiley & Sons, New York, 1989.
14. De Leeuw, E. D. *Data Quality in Mail, Telephone and Face-to-Face Surveys: A Quantitative Review.* Technical Report 1A, Response Effects in Surveys. Vrije Universiteit, Amsterdam, The Netherlands, 1990.
15. Lyberg, I. *Sampling, Nonresponse and Measurement Issues in the 1984/85 Swedish Time Budget Survey.* Statistics Sweden, Research Methods Development, 1990.
16. Harvey, A. *Proposal for Multinational Cooperation in Time Budget Research.* Saint Mary's University, 1984.
17. Gershuny, J., I. Miles, S. Jones, C. Mullings, and S. Wyatt. Time Budgets: Preliminary Analyses of a National Survey. *The Quarterly Journal of Social Affairs,* Vol. 2, No. 1, 1986, pp. 13–39.
18. Stoop, I. A. L., and J. Oudhof. Designing a Time Use Survey on Informally Productive Activities. *CBS Select 5.* Staatsuitgeverij, The Hague, The Netherlands, 1989.
19. Lingsom, S. *Open and Fixed Interval Time Diaries. A Comparison Based on a Pilot Study on Time Use 1979.* Statistisk Sentralbyra, Oslo, Norway, 1979.
20. Kalton, G. Sample Design Issues in Time Diary Studies. In *Time, Goods and Well-Being* (F.T. Juster and F.P. Stafford, eds.), Survey Research Center, University of Michigan, Ann Arbor, 1985, pp. 93–113.
21. Saris, W. E., and M. de Pijper. Computer Assisted Interviewing Using Home Computers. *European Research,* Vol. 14, 1986, pp. 144–152.
22. Van Doorn, L. Het gebruik van microcomputers in panelonderzoek. *Jaarboek '87–'88 van de Nederlandse Vereniging van Marktonderzoekers,* 1987.
23. Niemi, I. *The 1979 Time Use Study Method.* Central Statistical Office of Finland, Helsinki, 1983.
24. Kalfs, N. *Hour by Hour: Effects of the Data Collection Mode in Time Use Research.* NIMMO, Amsterdam, The Netherlands, 1993.
25. Tigerstedt, R., and M. Kaartama. Estimating Trip Frequency from Time Budget Data. Presented at 9th Reunion of the Research Group on Time Budgets and Social Activities, Espoo, Finland, Aug. 1984.
26. Ås, D. (1978). Studies of Time Use: Problems and Prospects. *Acta Sociologica,* Vol. 21, No. 2, 1978, pp. 125–140.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Temporal Variations on Allocation of Time

AJAY KUMAR AND DAVID LEVINSON

A study of the allocation of time and trip making across time of day, day of week, and month of year, as well as over the past 40 years, revealed some interesting findings. People are working much more, shopping somewhat more on weekends, and staying at home less today than they did 40 years ago. Time spent in travel on each weekend day (Saturday or Sunday) exceeds that on any weekday, as it did 40 years ago. Time spent shopping on a typical day in the busiest month (December) is more than twice that in the least busy month (September). Monthly variations in daily time in travel exceed 10 percent. The time-of-day patterns of shopping and other trips for workers and nonworkers are both rational: nonworkers peak in midday away from rush hour, whereas workers peak just after work, indicating trip chaining.

Growing congestion and changing travel patterns in urban areas have forced transportation researchers to venture beyond the confines of the daily work trip. Although work trips have been the traditional focus in transportation planning and policy formulation, recent studies have shown that nonwork trips are a dominant component of daily trip making and are growing faster than work trips (*1,2*). The historic emphasis on work trips was justified by the fact that the temporal clustering of work trips resulted in peak-hour congestion, dictating most investment decisions. However, as the role of transportation planners moves from investment to management, it is worthwhile to reexamine this issue in the broader perspective of activity patterns. One purpose of this paper is to look at both work and nonwork activity patterns, across all 7 days of the week.

Nonwork trips are tied to some basic and necessary human activities, such as shopping, performing errands, and socializing. Previous studies have related trip making and activity patterns to demographics and socioeconomic conditions (*3*) and trip generation to variations in land use patterns and metropolitan size (*4*). However, these activity patterns vary even more significantly across some fundamental criteria: natural and cultural cycles reflected in the calendar and the clock.

The study of human activity patterns has engaged the attention of researchers across disciplines. Recent developments by transportation engineers and modelers include attempts to introduce the concepts of trip chaining, activity sequencing,and combined time of day and route choice into demand forecasting procedures (*5–7*). Although these models have focused on methods for simulation of activity patterns, less empirical work has analyzed their long-term stability and their placement in a broader economic context. Pioneering work quantifying the use of time has been conducted by Szalar, who compared these results internationally, and Robinson, who conducted and reported on the American portion of that study (*8,9*.) Meanwhile, sociologists have examined the impact of the increasing number of women in the labor force on the quality of life and changing roles of time at work and leisure (*10–12*); planners have studied the allocation of time by activity and location for demographic and socioeconomic classes (*13–16*); and economists have developed a theory of the allocation of time wherein individuals or households combine time and market goods to produce "commodities" (*17*).

This study, part of a larger investigation into activity patterns, evaluates empirically the influence of temporal variations on the allocation of time. Much attention has been paid to trends in activity patterns, that is, the aspects of behavior that increase or decrease as a linear function of time. Less has been placed on the cyclical aspects of time—recurring patterns over the course of days, weeks, and years. Although most previous studies of travel behavior and time usage are atemporal, assuming an average day, this study, using the 1990 Nationwide Personal Transportation Survey, investigates variations in activity patterns by day of the week and month of the year, as well as the more traditional time of day. Information on weekend travel is sparse, and this analysis partially fills that gap. Answers to a number of questions are sought: What is the difference in activity patterns on Saturday versus Sunday? How different are the weekend activity patterns from an average weekday? Is there an average weekday? Does weekend travel exhibit the same diurnal relationship as weekday travel? How different are shopping trips from other nonwork trips?

Next in this paper is a discussion of the data base used in the analysis. This is followed by a review of long-term trends in the use of time, comparing studies performed in 1954 and 1966 and the 1990 Nationwide Personal Transportation Survey (NPTS) used here. Cyclical patterns are reviewed, and several hypotheses are tested in a comparison of month-of-year and day-of-week variations, respectively. Last is a discussion of time-of-day variations across the weekdays, Saturdays, and Sundays. The paper concludes with a discussion of the relevance of considering nonwork as well as work travel and considering the temporal variations in human activity patterns.

DATA

The original data base used in this analysis comes from the 1990–1991 NPTS. The NPTS was conducted as a telephone interview survey by the Research Triangle Institute, sponsored by the U.S. Department of Transportation (*18*). The survey collected data on household demographics, income, vehicle availability, all trips made on the survey day, long trips made over a 2-week period, and traffic accidents within the past 5 years. Characteristics of trips include departure time, distance, and duration of the trip, trip purpose and mode, day of the week, and month of the year. The survey was conducted between March 1990 and March 1991 and consisted of 21,817 household interviews and 47,499 persons making almost 150,000 trips. Because each interview consists of a single day, it is important to remember that the comparisons in this study across day

A. Kumar, Montgomery County Planning Department, 8787 Georgia Avenue, Silver Spring, Md. 20910. D. M. Levinson, Department of Transportation Studies, Department of Civil Engineering, Room 108, McLaughlin Hall, University of California at Berkeley, Berkeley, Calif. 94720.

of the week and month of the year do not come from the same individual. Conclusions must therefore be treated with caution. Further research with panel data will be able to compare the same individual across these time slices, offering another perspective on this issue.

First, it may be useful to define travel, activities, and their interrelationship. Activities are of two classes: location-specific activities and travel. Location-specific activities are defined on the basis of reported destination activity (purpose) from the travel survey. Travel is the activity that links other spatially separated location-specific activities. The core of this study comes from the 1990 NPTS, which like most travel surveys, provided respondents with a choice of answering where they went next (trip purpose), how they got there (mode), and how long it took (trip duration). These location-specific activities are consolidated into the following categories: home, work, shop, and other. The time spent traveling is accumulated into the travel activity category.

Only two pieces of time information were provided: the time of departure for a trip and the travel time for that trip. To create activity data, this study takes the NPTS "travel day" data base and, by looking ahead to the departure time of the next trip, determines the duration of the stop at the destination. A number of individuals did not report the time of arrival or departure for one trip during the day. These individuals were excluded as their daily time did not add up to 1,440 min. Only individuals who ended the day at home were considered in this study, and time at home was computed on the basis of final arrival time at home and initial departure at the beginning of the day. This is added to any stops at home in the middle of the day. For the graphs and tables presented in this paper, only adults aged 18 to 65 were considered. The elderly and children clearly have different diurnal, weekly, and seasonal time allocation patterns, and these may be evaluated in further research.

TABLE 1 Long-Term Trends in Use of Time

Activity	Year	Time Spent in Primary Activities by Day of Week (in hours): Average Weekday	Saturday	Sunday
Home	1954	17.0	17.8	19.2
	1965/66	15.8	18.0	19.1
	1990/91	15.7	18.1	18.7
Work	1954	4.6	2.6	0.8
	1965/66	5.0	1.8	0.8
	1990/91	5.4	1.8	1.0
Shop	1954	0.3	0.6	0.1
	1965/66	0.3	0.7	0.3
	1990/91	0.3	0.8	0.4
Other	1954	1.2	2.0	2.7
	1965/66	1.5	2.1	2.7
	1990/91	1.5	2.2	2.8
Travel	1954	1.0	1.1	1.2
	1965/66	1.4	1.4	1.1
	1990/91	1.0	1.1	1.1
Total	1954	24.0	24.0	24.0
	1965/66	24.0	24.0	24.0
	1990/91	24.0	24.0	24.0

Note: sources 1954 data - Sebastian de Grazia, J.A. Ward
1966 data - Robinson, Szalai
1990 data - Kumar and Levinson, 1990 NPTS
see text for discussion

ANALYSIS OF LONG-TERM TRENDS

Table 1 summarizes some long-term trends in activity patterns in the United States. These data are illustrative but cannot be compared in rigorous detail because of different methodologies used in the studies as well as limitations on the reported data. The 1954 results are reported by de Grazia (*11*) from an unpublished study, A Nationwide Study of Living Habits, conducted for the Mutual Broadcasting System by J.A. Ward. The J.A. Ward study used quarter-hour diaries during March and April 1954. The 1954 sample was large; 7,000 households and 20,000 individuals. The diaries were collected from 6 a.m. to 11 p.m., the remaining time was assumed to be spent at home.

The 1966 results are drawn from tables reported by Szalai and Robinson (*8,9*) in the monumental 1966 international use of time study. The sample was much smaller, over 2,000 adults, primarily as a day after diary. The data from this study were cross-classified in numerous ways and tables. Some of the tables, such as for travel, shop, and work were directly comparable with those from the other two studies. However, the results for home and other had to be inferred from several tables and adjusted to get a best estimate. This is because a number of activities that could occur at either location (home, other) were reported by type of activity (for instance, television watching or socializing with friends) rather than location.

Despite the differences in methods, some clear trends emerge. In 1990, adult Americans are working more on weekdays and less on Saturday than in 1954. The weekday rise is principally associated with the larger number of women working outside the home. Although Schor has argued that time at work has risen for men as well, this may not show up in a travel or activity survey but rather in wage data (*12*). The Saturday drop reflects the widespread adoption of the five-day work week since 1954. The amount of time spent shopping has held remarkably steady, although even small time differences in this category represent larger-percentage differences. Americans would appear to be shopping more on weekends now than before. This is partially a result of Sunday shopping, which was rare in 1954 because of blue laws, but this also seems to be true on Saturdays.

The amount of time in travel is almost identical between 1954 and 1990, although the 1966 study shows 10 to 30 percent higher weekend and 40 percent higher weekday travel time. To what extent this is real and to what extent it is a result of survey methods is unclear. However another study by the authors (*3*) shows that time in travel in metropolitan Washington has increased between 1968 and 1988 (from 1.3 to 1.7 hr for men and from 1.2 to 1.5 hr for women on weekdays) caused by the rise in nonwork trips and the increase in workers. This increased time is not, as has often been supposed, caused by a longer duration of work trips. The most important information for transportation analysis, the amount of time spent traveling, is ironically the least clear.

The two most curious categories are home and other. Given the increase in participation of women in the labor force, time spent at home from 1954 to 1990 should have been expected to decrease on weekdays. This is supported by the data. However several interacting factors made the issue more complicated. Saturday work has decreased, which makes more time available on Saturdays (for home and shopping), and the opening of stores (and other activity locations) on Sunday enables people to get out on Sunday.

ANALYSIS OF CYCLICAL PATTERNS

The analysis of cyclical variations on the allocation of time in 1990 America takes several forms: time of day, day of week, and month of year. Five activity patterns are identified in this paper: home, working, shopping, other, and travel. "Other" activities are defined to include trips for the following: family or personal business, school or church, doctor/dentist, visiting friends or relatives, social/recreational, and any otherwise nonspecified activity (not home, working, shopping, or vacation). The other trips were grouped to maintain sample size significance and simplify the analysis. Time spent at each of the activities and diurnal variations, average frequency, and duration of activities are computed for the different time slices: month of the year, day of the week, and time of day. These are addressed in turn.

The information is presented in graphs that show the mean daily duration of each activity. Behind each graph lies a table, not presented for space reasons but available from the authors on request, which contains matrixes of the t-statistic resulting from a difference of means tests for month verses month and for day verses day. In this way, the statistical significance of differences of points on the graph could be ascertained. Monday can be compared with Tuesday, and March can be compared with April, and comparisons between any given day and the average can also be made. The statistical significance of the difference of means that are reported were developed from those tables with a report of significance indicating that the difference is significant at the 90 percent confidence level or better on a two-tailed t-test.

Hypotheses

The NPTS data base offers innumerable possibilities for analysis. Keeping the focus on temporal variations, several hypotheses are explicitly evaluated in this study. First, it is hypothesized that there is a tie between human activity patterns and seasonal cycles, which will be indicated by differences in average activity durations in winter and summer, spring and fall. These differences are expected to occur in each of the activities, with different activity-specific patterns across the months of the year.

The second hypothesis is that Saturday and Sunday behavior are expected to differ from each other and from weekdays, but weekdays are expected to be similar to each other. The difference in activity patterns between Saturday and Sunday results from a variety of obvious religious and cultural reasons. This is tested across activities.

A third hypothesis concerns the temporal distribution of regional and neighborhood shopping: longer shopping trips to stores farther away will occur on weekends. A similar pattern is also expected to emerge for other trips, which should be longer on the less-constrained weekends.

The last set of hypotheses concerns time of day: that on weekdays, workers will tend to perform shopping and other activities on the way home from work, whereas nonworkers will tend to perform shopping and other trips outside of the peak commuting hours. This results from a desire to avoid congestion during peak periods on the part of nonworkers and to minimize travel time on the part of workers by combining nonwork trips with the work trip. In short, individuals are assumed to make boundedly rational decisions on the allocation of time that produces this scheduling behavior (*3*).

Activity Duration by Month of Year

Figures 1 and 2 display average daily time distribution by month and activity. These graphs indicate seasonal variations in the time spent at various activities. It is hypothesized that there is a link between human activity patterns and natural (and cultural) cycles, which will be reflected by differences in activity durations. Future research may compare activity patterns and geography to get an indication of the relative importance of climate compared with other seasonal/cultural patterns.

Several statistically significant results are found. Time at home peaks around the December holidays (1,015 min) and reaches a nadir in April (960 min). Many of the differences between months are significant, and although some pairwise comparisons of months do not appear significant, the trends seem to be. For instance, for time at home, January does not significantly differ from February, and February does not significantly differ from March. But January differs more from March for time spent at home (than February) and is significantly different from April, all suggesting a real trend.

Time at work (per person, not per worker) is the opposite from time at home, peaking in April (275 min) and with a low in December (220 min). Moreover, time at work has a secondary valley during July because of summer vacation (250 min). The differences here are not as significant; only December is significantly different from the average month.

Time spent shopping per day peaks in December (34 min), from a September low (15 min). December, January, May, and September are significantly different from the average month, and the months with a great deal of shopping are different from those with below-average shopping.

Time at other is flat, ranging from 100 min in winter to 120 min in spring and summer. May and October are significantly different from the average months, and again, a number of pairwise comparisons are also significantly different.

Travel consumes 62 min/day in most months but in summer consumes 70 min. May, July, and August differ from the average month, and the winter months are different from the summer months.

Activity Duration by Day of Week

Figures 3 and 4 display time spent at each of the five activities (home, work, shop, other, and travel) by day of week. For each day, the total time of the five activities adds up to 1,440 min. The hypothesis is that weekday activity patterns are similar to each other but differ from weekends and that Saturday differs from Sunday.

As expected, time spent at each of the activities tends to be somewhat the same across the work week, although it differs over the weekend. However, even during the work week, some variations can be observed:

Time at home on Mondays is greater than on the other four weekdays, perhaps because of recovery from the weekend or the "3-day weekend" (associated with official holidays and personal vacation), whereas time at work is slightly less on Mondays. This difference

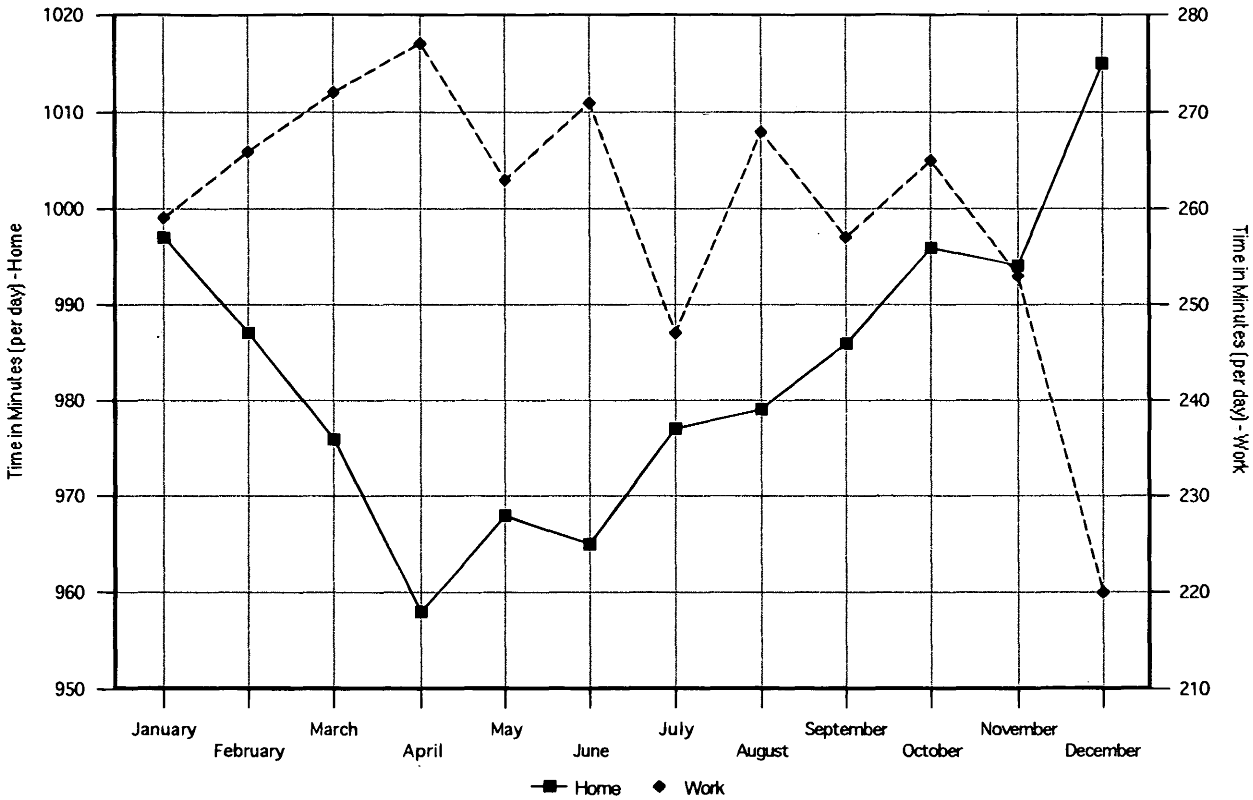


FIGURE 1 **Time at home and work by month (*18*).**

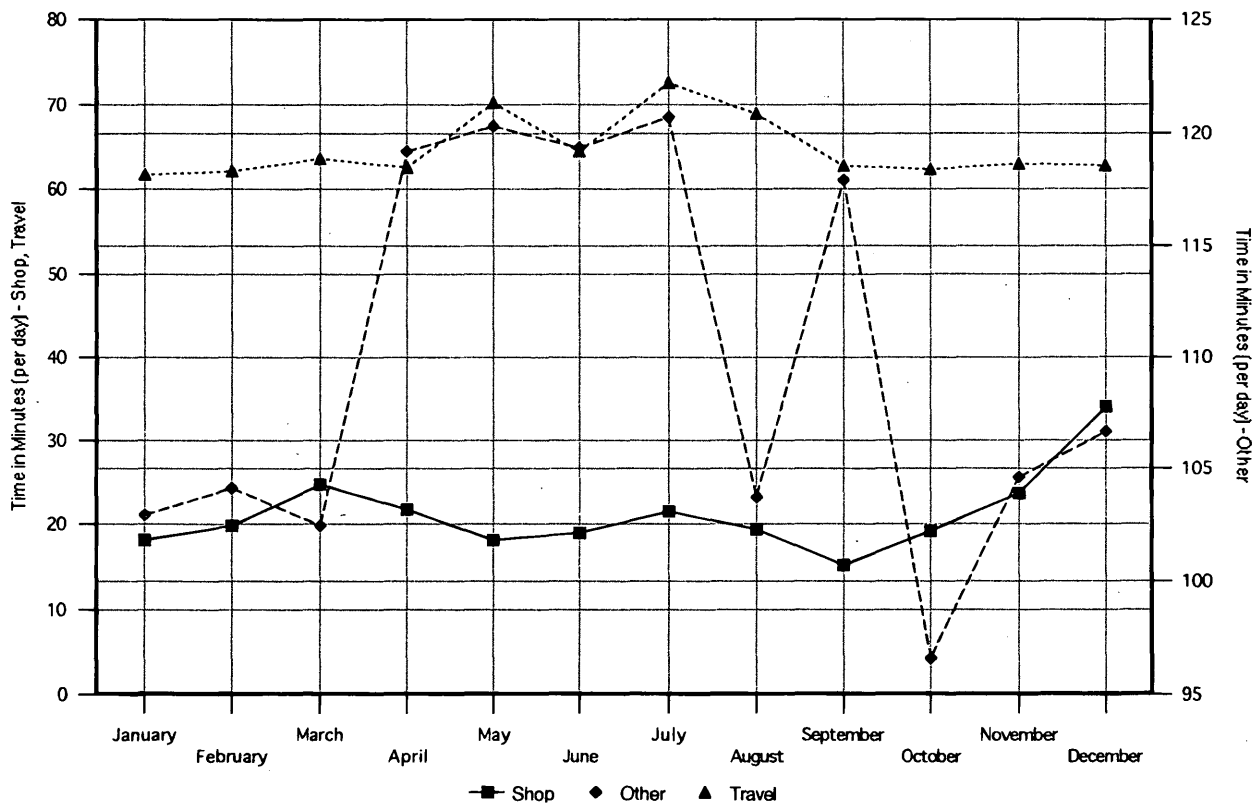


FIGURE 2 **Time shopping, traveling, and other by month (*18*).**

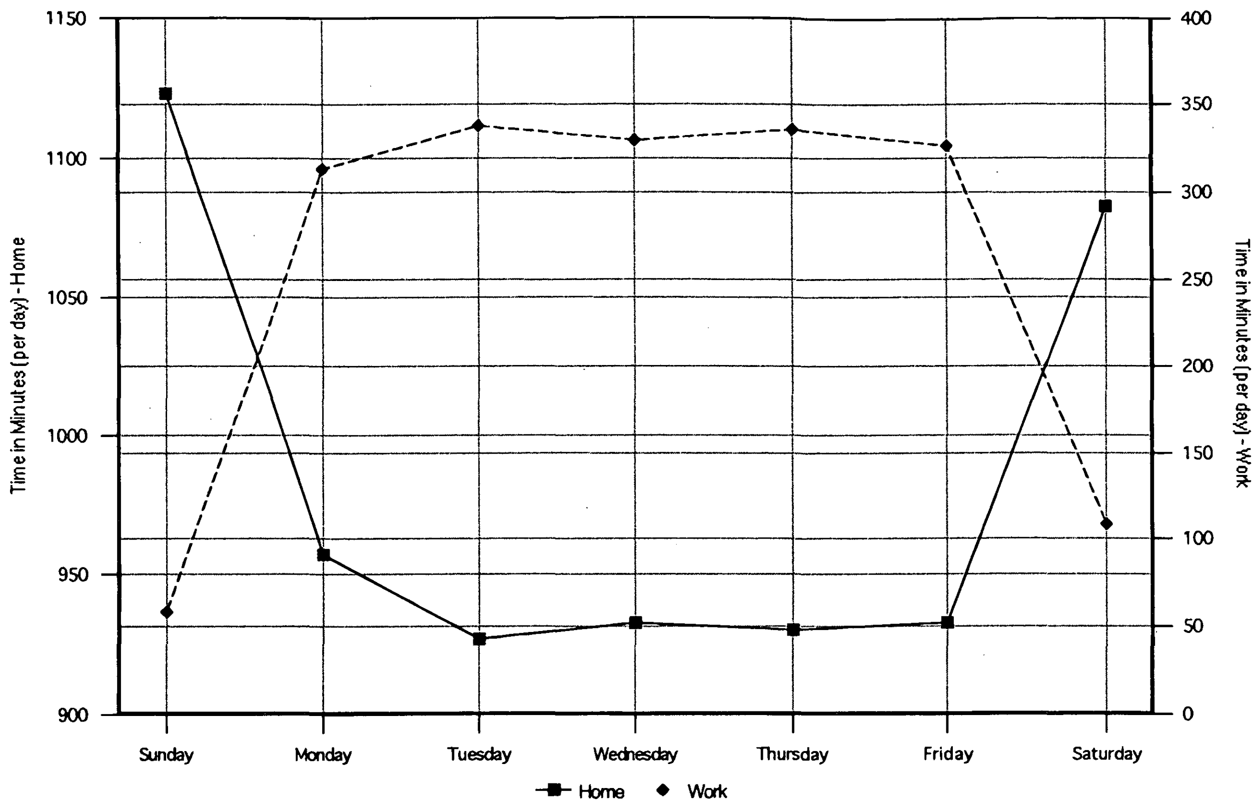


FIGURE 3 Time at home and work by day of week (*18*).

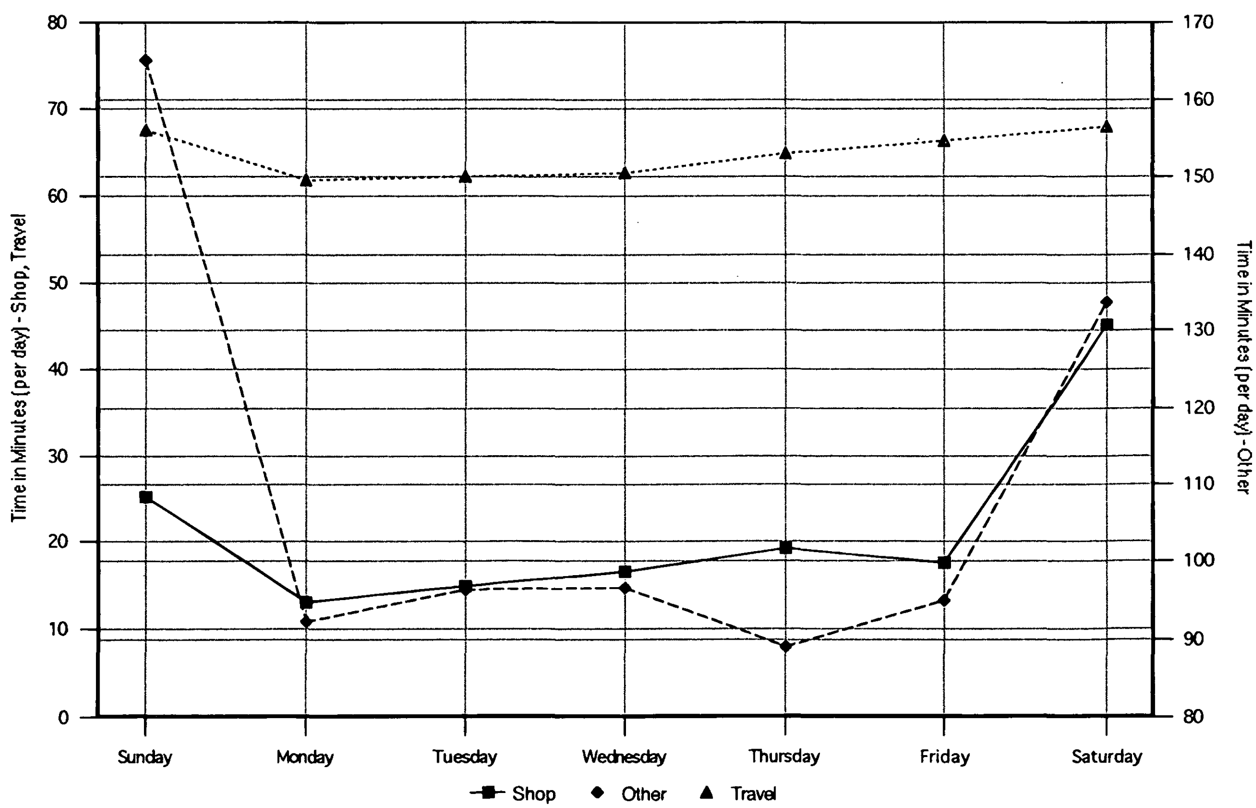


FIGURE 4 Time shopping, traveling, and other by day of week (*18*).

is statistically significant. However, the time at home on the other weekdays is not statistically different. Time at home is greatest on Sundays (1,125 min) followed by Saturdays (1,080 min). The weekends are statistically different from the weekdays and from each other, validating the hypothesis.

Also, time at work on Mondays is significantly different from that on other weekdays and, as expected, the weekends do differ from the weekdays and each other. However, Tuesday through Friday are similar.

Time spent shopping rises from Monday to Friday, with a small peak on Thursday (19 min). Shopping peaks on Saturday (45 min), followed by Sunday (25 min). Although adjacent weekdays are not different from each other (the difference between Monday and Tuesday or between Tuesday and Wednesday is not significant), the difference between nonadjacent weekdays does tend to be significant, again suggesting a trend over the week. The weekend days are significantly different from each other and weekdays.

Time at other activities is fairly flat over the weekdays, with a dip on Thursdays (90 min). Time at other activities peaks on Sundays (165 min) followed by Saturdays (135 min). The weekdays are not significantly different from each other, although the weekend days are different from each other and weekdays.

Time in travel rises slightly from Monday to Wednesday but more sharply from Wednesday to Friday. Time in travel on the weekends is greater than on weekdays, with Saturday being the highest at 68 min. However, weekdays are not significantly different from each other, and Saturday is not significantly different from Sunday, but the weekends are significantly higher than weekdays.

Trip Making by Day of Week

Figures 5 through 7 show trip frequency, duration, and distance by day of week. These figures are classified by worker and nonworker and come out as might be expected from the earlier discussion.

Figure 5 shows trip frequency. Work trips for workers basically are flat across weekdays, as are trips for shopping. Work trips are more frequent on weekdays than on weekends, and higher on Saturdays than on Sundays. Other trips are fairly consistent across weekdays until Friday, when there is a rise for both workers and nonworkers. Weekends have more nonwork trips than weekdays. However, a higher share of other and shopping trips for workers occurs on weekends than on weekdays compared with nonworkers, indicating a displacement. Again, nonworkers can make these trips on weekdays in midday, which is relatively uncongested, whereas workers must perform these activities on weekends.

Trip duration and distance by day of week, shown in Figures 6 and 7, come out as might be expected, in part because work trips are longest. Weekend work trips are shorter than weekday trips, likely because of different types of jobs (weekend employment is more often part time, retail jobs). Somewhat surprisingly, the work trip duration variances within the week show statistical significance. Among those who work, Thursday and Friday trips take longer than Monday or Wednesday trips. The Monday verses Friday difference may be explainable by congestion (there are fewer trips on Monday than other days, and many 3-day weekends begin during Friday evening rush hour). Alternatively, some of the difference may be because of trip chaining, which might add to reported times, but for

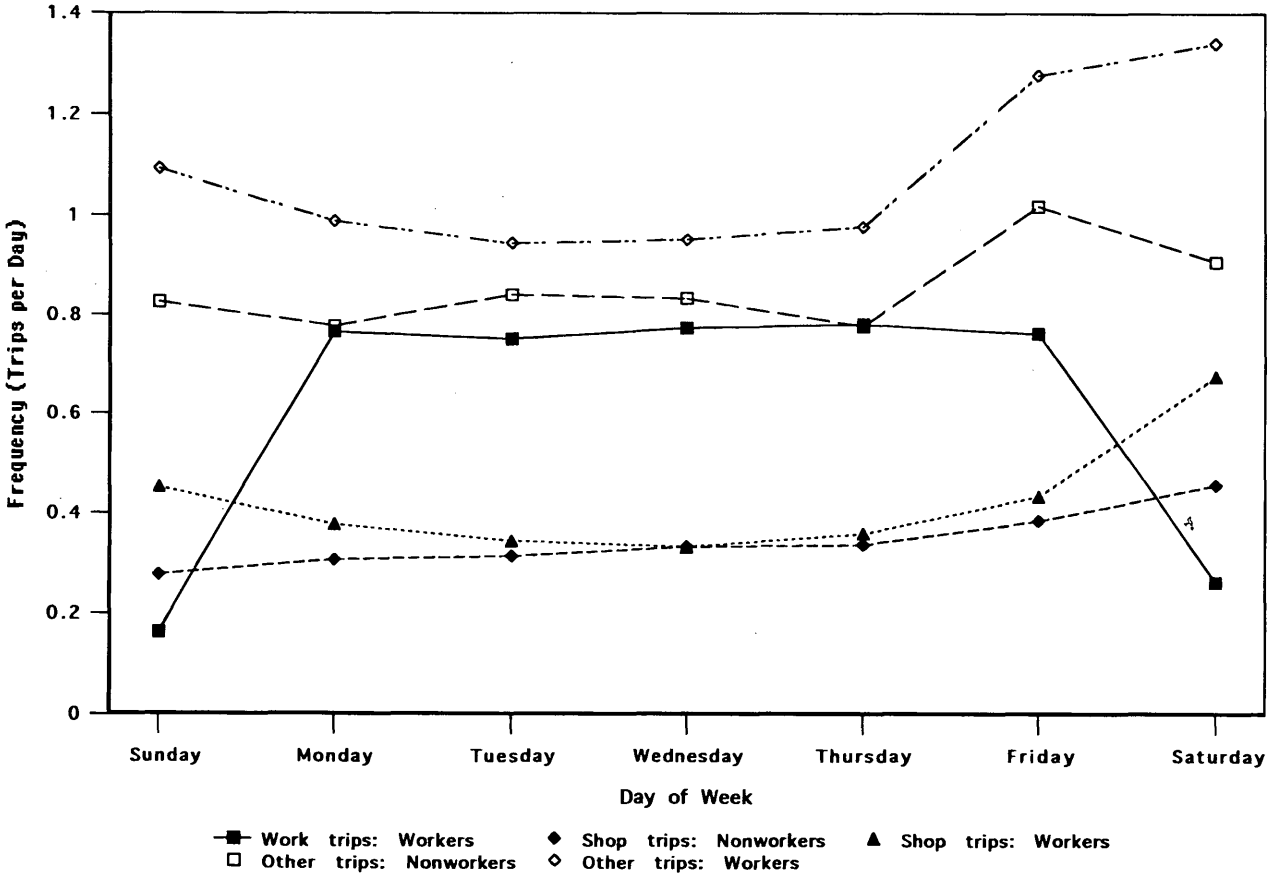


FIGURE 5 Trip frequency by day of week.

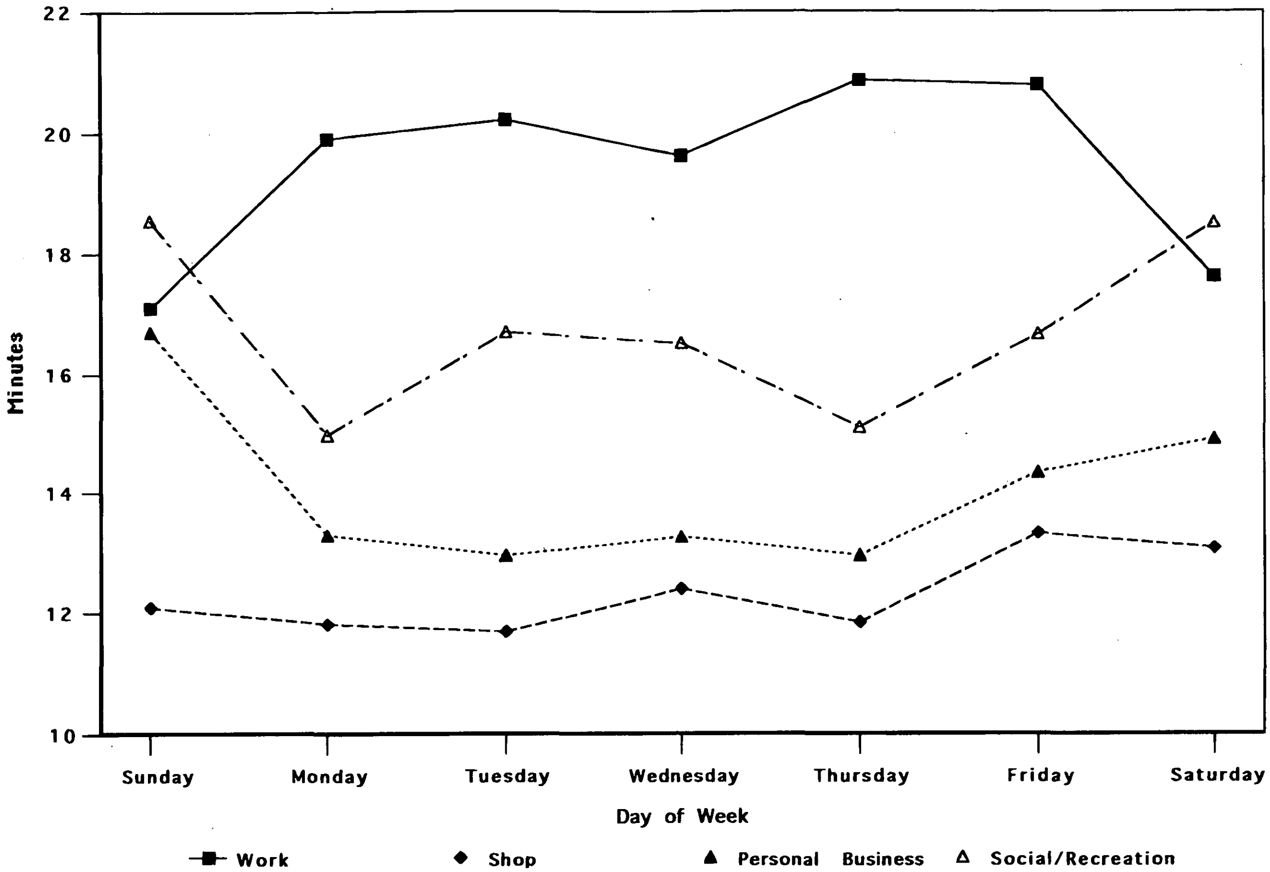


FIGURE 6 Trip duration by day of week and purpose.

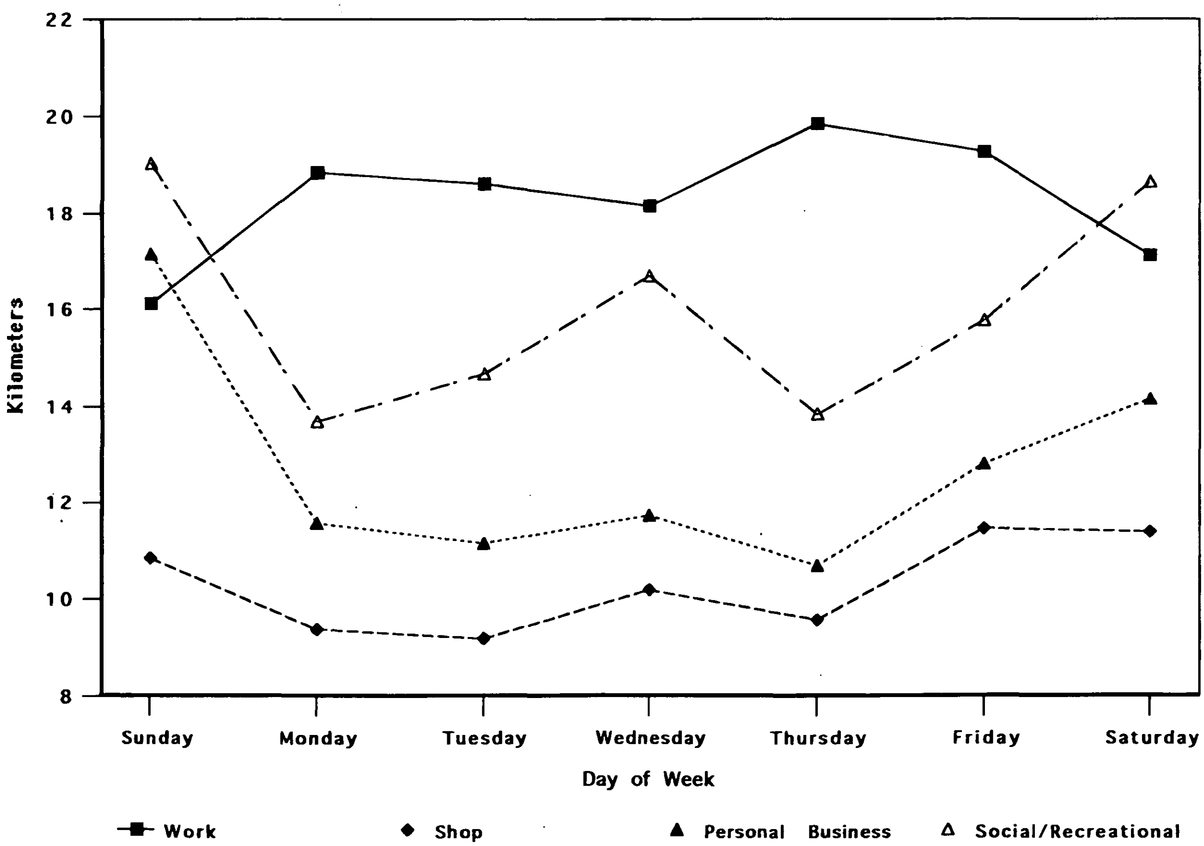


FIGURE 7 Trip distance by day of week and purpose.

some activities (getting gas, stopping at a convenience retail) may not be reported 100 percent of the time.

Interestingly, social/recreational trips are longer than personal business, which are longer than shopping trips, indicating that not all nonwork trips share the same characteristics. Other trips on weekends are longer than on weekdays, but this is hardly true for shopping trips. Personal business is significantly longer on Friday, Saturday, and Sunday than the rest of the week, and social-recreational trips are longest on the weekend and shortest on Monday.

It was anticipated that regional shopping (mall-going, shopping for durable goods, etc.) would necessitate longer trips than neighborhood shopping (groceries); they are somewhat longer in distance (6.5 versus 5.5 mi) and somewhat shorter in duration (Friday and Saturday have durations of 13 min, whereas other days average 12 min), indicating higher speeds because of both less congestion on weekends and the use of different, higher-speed roads for regional shopping as opposed to local shopping. The differences between Friday and Saturday and the rest of the week are statistically significant.

Another noteworthy point is that although the trip frequency for other trips exceeds that of the non-other categories, even for workers, the average other trip (either personal business or social recreational) is shorter than the average work trip. So their impact on total travel (e.g., vehicle-miles traveled) is similar. Fortunately, they do have different peaking patterns, as shown in the next section, and use different roadways.

Time-of-Day Distribution

The time-of-day distribution of trips for workers and nonworkers for the average weekday, Saturday, and Sunday, classified for shop and other trip purposes was analyzed. The time-of-day distribution for work trips on weekdays is well documented and has remained largely stable over the past few decades, with some peak spreading (*3*). Figures 8 and 9 indicate the time-of-day distributions for shopping and other trips, respectively.

Given the obligatory and regular nature of work trips, it is expected that workers and nonworkers will have somewhat different behavior. The hypothesis is that, on weekdays, workers will tend to perform shopping and other activities after work, often on the way home, to minimize travel through trip chaining; nonworkers, also to achieve travel economies, will tend to perform weekday shopping and other trips outside of the peak commuting hours. In addition, for a variety of religious and cultural reasons, Saturday and Sunday behavior is expected to differ from each other and from that on weekdays. Probably because of the need to rise early for work on Monday, as well the closing of shops, Sunday "ends" for most people earlier than Saturday.

Several results are found from inspection of the graphs. On weekdays, for workers, shopping trips peak after the close of work, whereas other trips have two peaks: at lunch and after the close of work. On weekdays, for nonworkers, shopping trips peak before midday and decline thereafter, and other trips peak after midday

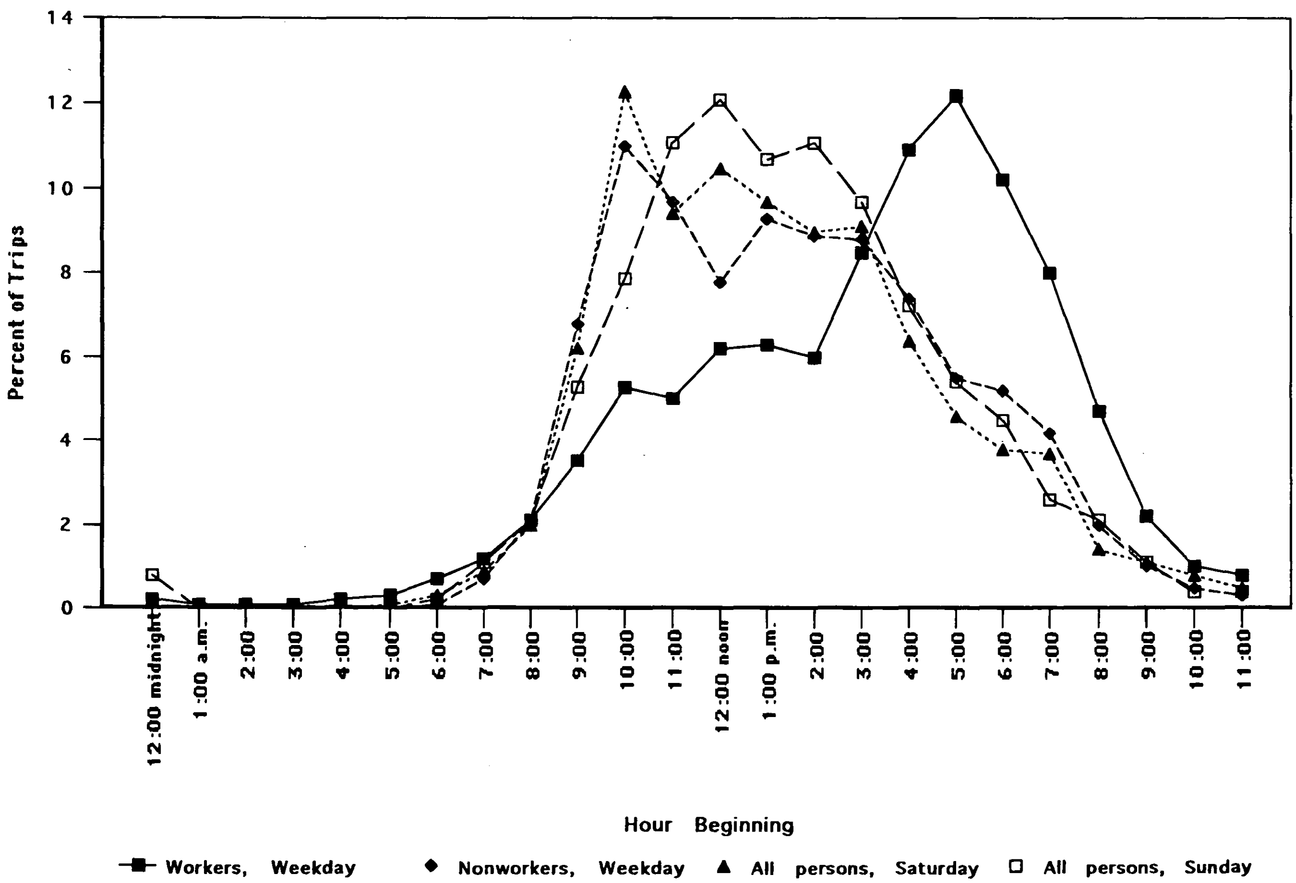


FIGURE 8 Time-of-day distribution for shopping trips.

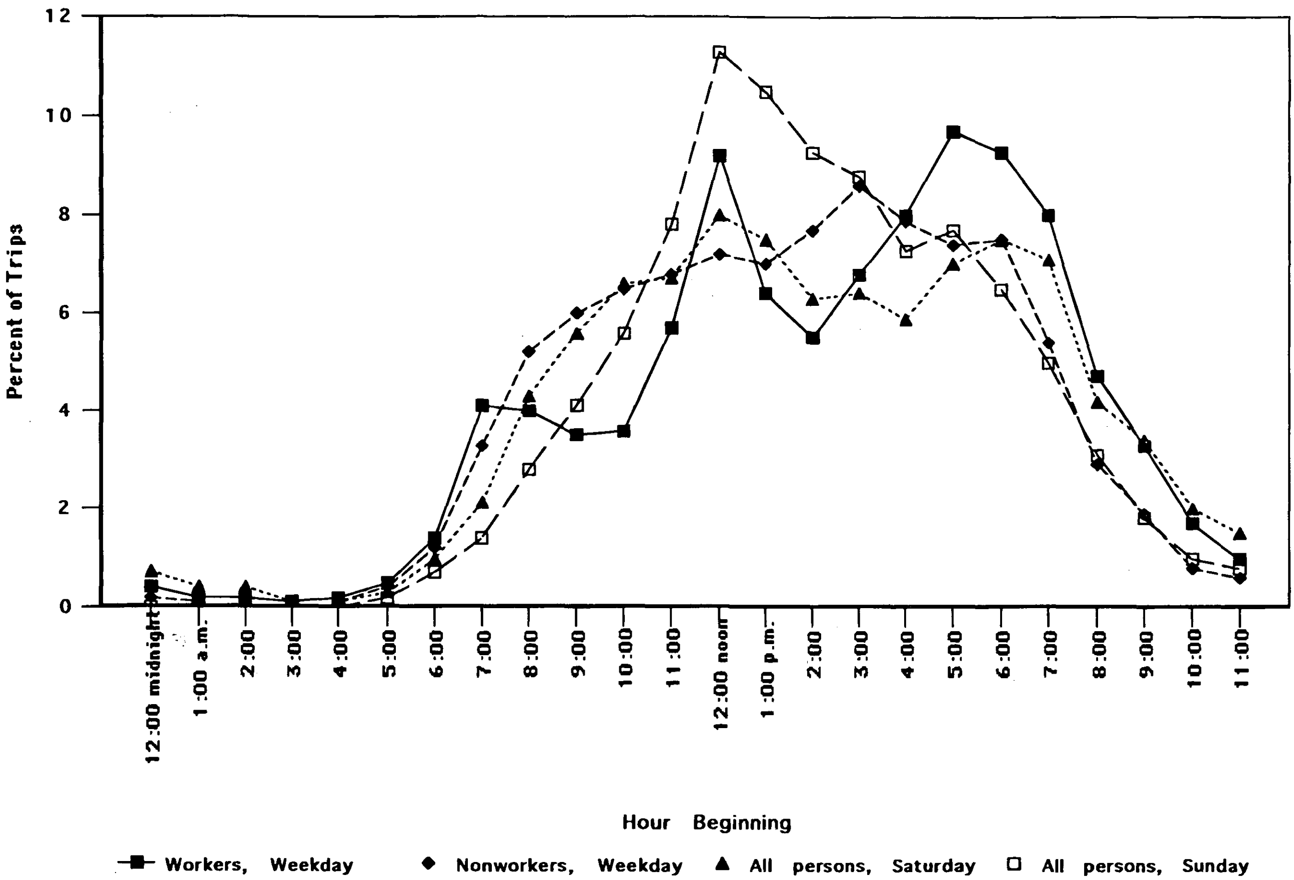


FIGURE 9 Time-of-day distribution for other trips.

(3:00 p.m.). Saturday shopping patterns are similar to nonworkers' weekday patterns, although Saturday, like a typical workday, has two peaks for other activities, at noon and 6:00 p.m. People shop earlier on Saturday than on Sunday, probably because of Sunday church-going, as evidenced by other activities (which include school and church) being conducted earlier on Sunday than Saturday.

CONCLUSIONS

The prime mover in the rise in both work and nonwork trip making over the past few decades has been the growth in women's participation in the labor force. This rise has directly increased the number of workers and thus work trips. It also resulted in the increase in per capita (if not household) incomes while reducing available time and thereby permitted the substitution of household commodities from outside the home (day care for at-home child rearing, eating out for home-cooked meals), which leads to more nonwork trips per person.

This analysis brings out some interesting results. People are working much more, shopping somewhat more on weekends, and staying at home less today than they did 40 years ago. Time spent in travel on each weekend day (Saturday or Sunday) exceeds that on any weekday, as it did 40 years ago. This finding underscores the need to focus greater attention on weekend travel. Time spent shopping on a typical day in the busiest month (December) is more than twice that in the least busy month (September). Monthly variations in daily time in travel exceed 10 percent. The time-of-day patterns of shopping and other trips for workers and nonworkers both are a result of rational decision-making processes: nonworkers peak in midday away from rush hour, whereas workers peak just after work, indicating trip chaining.

Several factors suggest that, in the future, nonwork activities will become relatively more important. First, advances in telecommunication should enable more work at home and thus free some time formerly spent commuting for nonwork trips. Second, the large increase in the number of workers in the labor market caused by women joining the workforce is ending. The share of the labor force held by men and women is equalizing. One factor that is certainly related to travel demand is income, but over the past two decades income growth has slowed (*3*). If this is in part because of the rapid rise in women's participation in the labor force (and a relatively higher labor supply), this trend of sluggish income growth may end as labor becomes scarcer and more costly. These higher incomes may result in nonwork travel and changes in activity patterns.

Thus an understanding in the patterns of nonwork activity should become even more important in coming years. This is pertinent with the growing concern about developing strategies for traffic mitigation and environmental control, which focuses almost entirely on work trips. Some of the findings of this study may be particularly relevant for effective travel demand management programs as well as monitoring environmental consequences. Most air pollution emissions analyses derived from traffic forecasting models assume the "average" day. But as can be seen from these figures, not all weekdays are created equal, weekdays differ from weekends, and travel patterns vary seasonally. As weather patterns also vary seasonally, climate-specific, as well as congestion-inspired, demand programs may be targeted to account for these variations. In addition, dynamic travel simulation models, which estimate changes over time, should incorporate variations associated with these cycles.

In brief, this study shows empirical relationships between activity patterns and trip making and natural and cultural cycles (time of day, day of week, and month of year). Although many causes can only be speculated about, the results are predictable. Further analysis is required to tie down the causes of many of these variations and determine how the same factors influence different individuals. This research should focus on the interaction of temporal, spatial, socioeconomic, and demographic characteristics of individuals in consuming various amounts of activities.

ACKNOWLEDGMENTS

The authors acknowledge Chris Winters and the staff of the Montgomery County Planning Department. The authors also thank four anonymous reviewers for their comments.

REFERENCES

1. Gordon, P., A. Kumar, and H. Richardson. Beyond the Journey to Work. *Transportation Research A,* Vol. 22A, No. 6, 1988, pp. 419–426.
2. Pisarski, A. *New Perspectives in Commuting.* FHWA, U.S. Department of Transportation, 1992.
3. Levinson, D., and A. Kumar. Activity, Travel, and the Allocation of Time. *Journal of the American Planning Association,* (in press).
4. Levinson D. and A. Kumar. *Is Residential Density a Transportation Issue?* Working Paper. Montgomery County Planning Department, Silver Spring, Md., 1993.
5. Clarke, M. I., M. C. Dix, P. M. Jones, and I. G. Heggie. Some Recent Developments in Activity Travel Analysis and Modeling. In *Transportation Research Record 794* TRB, National Research Council Washington D.C. 1981, pp. 1–8.
6. Recker, W. W., M. G. McNally, and G. S. Root. In A Modern of Complex Travel Behavior: Part I—Theoretical Development, and Part II—An Operational Model. *Transportation Research* Vol. 20A, 1986, pp. 307–330.
7. Kitamura, R. Formulation of Trip Generation Models Using Panel Data. In *Transportation Research Record 1203,* TRB, National Research Council, Washington, D.C., 1988, pp 60–68.
8. Szalai, A. *The Use of Time: Daily Activities of Urban and Suburban Populations in Twelve Countries.* Mouton Publishers, The Hague, The Netherlands, 1972.
9. Robinson, J. *How Americans Use Time: A Social-Psychological Analysis of Everyday Behavior.* Praeger Publishers, New York, 1977.
10. Cross, G. *Time and Money: The Making of Consumer Culture.* Routledge Books, New York, 1993.
11. De Grazia, S. *Of Time, Work, and Leisure.* Vintage Books, New York, 1990.
12. Schor, J.B. *The Overworked American: the Unexpected Decline of Leisure.* BasicBooks, New York, 1991.
13. Chapin, F. S., and H. Hightower, Household Activity Patterns and Land Use. *AIP Journal,* Vol. 31, No. 3, Aug. 1965.
14. Chapin, F. S. Activity Systems and Urban Structure: A Working Schema. *AIP Journal,* Vol. 34,; No. 1, Jan. 1968.
15. Chapin, F. S. Free Time Activities and Quality of Urban Life. *AIP Journal,* Vol. 37, No. 6, Nov. 1971.
16. Chapin. F. S. *Human Activity Patterns in the City: Things People Do in Time and in Space.* John Wiley & Sons, New York, 1974.
17. Becker, G. A Theory of the Allocation of Time. *Economic Journal,* Vol. 75, No. 299, Sept. 1965, pp. 493–517.
18. *The 1990/91 Nationwide Personal Transportation Survey: User's Guide for the Public Use Tapes.* FHWA, U.S. Department of Transportation, 1991.

The ideas and opinions expressed in this paper are those of the authors and not the Department of Transportation Studies.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Sample Selection Bias with Multiple Selection Rules: Application with Residential Relocation, Attrition, and Activity Participation in Puget Sound Transportation Panel

JIN-HYUK CHUNG AND KONSTADINOS G. GOULIAS

Two sources of sample selection bias emerging simultaneously from panel attrition and residential relocation and their effect on activity participation are examined. The data used were from two time points (Wave 1 in 1989 and Wave 2 in 1990) of the Puget Sound Transportation Panel. Data regarding relocation decisions, taking place between Wave 1 and Wave 2, are available for the households that participated in both waves (participants) and are not available for the households that participated in the first wave only (dropouts). Double selection was associated with the possible simultaneous or sequential decision process underlying participation in the survey and household residential relocation. The method used is based on a bivariate probit model that accounts for selectivity. The method emerges from the unknown relocation status of the dropouts in Wave 2. Subsequent creation of correction terms, needed to account for the lack of data on dropout households' activity participation in Wave 2, uses the probit model. The method, called the Tunali method, is a two-step procedure that follows the usual Heckman method. The models estimated, that is, the bivariate probit model of double-selection and activity participation linear regressions corrected and uncorrected for selection, are provided.

Dynamic analysis of travel behavior is greatly facilitated when panel survey data—information from repeated observations of the same individuals over time—are available. A common problem to all panels, however, is the potential selectivity bias emerging from attrition or refusal to participate in a subsequent time point of the survey. Ordinary least-squares (OLS) regression coefficient estimates are inconsistent if attrition occurs in a systematic way, and it is not accounted for in estimation. Analogously, selectivity bias may also emerge from other sources. For example, nonrandom residential relocation (or, more generally, migration) during the panel survey may also produce similar biases. In addition, attrition and residential relocation decision making may also be related. For example, relocating residents may be more likely to refuse participation in the panel in subsequent waves. A method is needed to remove selectivity bias in which attrition and residential relocation are considered simultaneously. This would allow researchers to test hypotheses about the relationship between attrition and relocation, derive sample weights that can be used for subsequent waves of a panel, and provide for a complete correction method for regression models that suffer from selectivity biases.

The most common selectivity bias correction method, used in transportation modeling, takes the form of an equation that represents the selection process with a discrete dependent variable (e.g., participation or nonparticipation in a survey). Another equation represents the outcome of some decision-making process (e.g., number of household trips or number of cars owned by a household). This is the equation for which consistent estimates are needed. The usual technique to account for selectivity bias has been to create "correction" terms used to augment the target regression equation and "eliminate" the selectivity bias as if it were a specification error. This method treats selectivity as a specification error and is named the Heckman correction method (*1,2*). The method has been used by Mannering (*3*), Kitamura and Bovy (*4*), Hensher et al. (*5*), and Monzon et al. (*6*). In this paper this method is called the single-selection model because it includes only one source of selectivity. When the sources of selectivity are several, similar methods can be devised and multiple correction terms can be used to eliminate the bias. These methods, however, are more complex than the single-selection method. Their complexity increases exponentially when relationships exist between the selectivity sources and when portions of the "selected" sample are unobserved (*7*).

In this paper two sources of selectivity are considered: panel attrition and residential relocation. Their effect on activity participation is also examined. The data used are from the first two time points (Wave 1 in 1989 and Wave 2 in 1990) of the Puget Sound Transportation Panel [PSTP, described by Murakami and Watterson, (*82*)]. Data about relocation decisions, taking place between Wave 1 and Wave 2, are available for the households that participated in both waves (participants) and are not available for the households that participated in the first wave only (dropouts). This precludes the use of the methods devised by Kitamura et al. (*9*) and may be the source of "double selection," as a result of the possible simultaneous or sequential decision process underlying participation in the survey and residential relocation. The method, based on a bivariate probit model, accounts for selectivity caused by the unknown relocation status of the dropouts. The lack of data on activity participation for the panel dropouts is another source of selectivity. The method creates two correction terms to be used in the Wave 2 activity participation equations.

First the paper presents a more general model of double selection. Then, the selectivity model is described with a few estimation issues. It then provides a short description of the data analyzed. Then, estimation results for the bivariate probit model of attrition and residential relocation and the augmented regressions (with the two correction terms) of activity participation are provided. A summary and conclusion are offered last.

Department of Civil and Environmental Engineering, Pennsylvania State University, University Park, Pa. 16802.

MODEL

The general model of double selectivity, of which the model used in this paper is a particular case, is formulated as follows. Each household in the sample is characterized by two discrete-outcome decisions, to participate in the Wave 2 of the panel and to change the residential location between Wave 1 and Wave 2. A third decision is characterized by a "continuous" outcome, that is, frequency of activity participation in Wave 2. Using the dichotomous variables, Y_1 and Y_2, to represent the two discrete outcome decisions and the continuous variable Y_3 to represent the continuous outcome, it is possible to write the two selection "rules" in terms of explanatory variables such as

$$Y^*_{1i\ 1i} = \beta_1' X_{1i} + \epsilon_{1i}$$

$$Y_{1i} = 1 \qquad \text{if } Y^*_{1i} > 0$$

$$Y_{1i} = 0 \qquad \text{if } Y^*_{1i} \leq 0 \tag{1}$$

and

$$Y^*_{2i} = \beta_2' X_{2i} + \epsilon_{2i}$$

$$Y_{2i} = 1 \qquad \text{if } Y^*_{1i} > 0$$

$$Y^*_{1i} = 0 \qquad \text{if } Y^*_{1i} \leq 0 \tag{2}$$

The third equation describing the continuous dependent variable is as follows:

$$Y_{3i} = \beta_3' X_{3i} + \sigma_3 \epsilon_{3i} \tag{3a}$$

where

X_{ki} = vectors of explanatory variables (k = 1,2,3),
σ_3 = unknown scale parameter, and
β_k = unknown regression coefficient vectors to be estimated with the elements of the variance-covariance matrix of (ϵ_{1i}, ϵ_{2i}, ϵ_{3i}) reported in Equation 4:

$$\sum = \begin{bmatrix} 1 & \rho & \rho_{13} \\ \rho & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} \tag{4}$$

Equations 1 through 4 describe the structure of the model under consideration. The household observations contain information on Y_{1i}, Y_{2i}, Y_{3i}, and X_{1i}, X_{2i}, X_{3i}. Y^*_{1i} and Y^*_{2i} can be interpreted as the propensity of the household to relocate and to participate in the second wave of the panel survey, respectively. Considering the two discrete outcome variables, described by Equations 1 and 2, there are four possible joint outcomes. In Figure 1 this can be indicated by a four-cell table containing the frequency of the number of households in each combination of outcomes. Assuming that the assumptions ϵ_{1i}, ϵ_{2i}, ϵ_{3i} are trivariate normally distributed with 0 mean and covariance given by Equation 4, and error terms independent across households and the explanatory variables, then it is possible to write the joint cell probabilities reported in the second part of Figure 1.

The probability density, associated with each cell, of Y_{3i}, can be written as a function of the cell probability and the trivariate normal density of the ϵ's. These components in turn can be used to derive a likelihood function for the entire system of equations and then use it for estimation via maximum likelihood. A problem arises, however, when some cells in Figure 1 are not observed.

In Figure 1 the data present four possible distinct regimes defined by the combination in outcomes depicted by the variables Y_1 and Y_2. [There will be four pairs of possible joint outcomes for Y_1 and Y_2. (0,0), (0,1),(1,0), and (1,1).] Letting [$Y_1 \times Y_2$] be the joint outcome of the two variables in Figure 1, the expectation of Equation 3a can be written as

$$E(Y_{3i}, Y_1 \times Y_2) = \beta_3' X_{3i} + \sigma_3 E(\epsilon_{3i} \mid X_{3i}, Y_1 \times Y_2) \tag{5}$$

In Figure 1 there are four distinct subsamples. One equation of the type described in Equation 5 applies to each. However, panel attrition and residential relocation are characterized by the lack of information on residential relocation of households that dropped out of the panel. In terms of Figure 1, there are only three distinct cells:

Frequencies

	Y_2	
Y_1	0	1
0	N_1	N_2
1	N_3	N_4

Probabilities

	Y_2	
Y_1	0	1
0	$BN(\beta_1 X_1, \beta_2 X_2, \rho) = \int_{-\infty}^{-\beta_1 X_1} \int_{-\infty}^{-\beta_2 X_2} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$	$BN(-\beta_1 X_1, \beta_2 X_2, -\rho) = \int_{-\infty}^{-\beta_1 X_1} \int_{-\beta_2 X_2}^{\infty} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$
1	$BN(\beta_1 X_1, -\beta_2 X_2, -\rho) = \int_{-\beta_1 X_1}^{\infty} \int_{-\infty}^{-\beta_2 X_2} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$	$BN(-\beta_1 X_1, -\beta_2 X_2, \rho) = \int_{-\beta_1 X_1}^{\infty} \int_{-\beta_2 X_2}^{\infty} f(\epsilon_1 \epsilon_2)\, d\epsilon_1 d\epsilon_2$

FIGURE 1 Four discrete outcomes and associated probabilities.

1. Participants who change residential location (movers) and took part in both panel waves (participants),
2. Participants who did not change their residential location (stayers) and took part in both panel waves, and
3. Participants in Wave 1 only (dropouts) of unknown residential relocation choice.

It is clear then, that observation of residential status is conditional on panel attrition (herein called incomplete information). In terms of Figure 1 this is equivalent to "collapsing" two cells into one. For these cells instead of a bivariate normal cell probability one obtains a univariate normal probability (e.g., corresponding to the probability of panel attrition). Estimation of Equations 1 and 2 also can be performed using a log likelihood function that is analogous to the usual bivariate probit likelihood function.

Consider Y_1 representing residential relocation status (taking the value of 1 if the household did not move and 0 otherwise) and Y_2 representing panel participation (taking the value of 0 if the household is a dropout and 1 otherwise). The cells with incomplete information are ($Y_1 = 0$, $Y_2 = 0$) and ($Y_1 = 1$, $Y_2 = 0$). The sample size of each distinct cell is $(N_1 + N_3)$ for the dropouts, N_2 for participant-movers, and N_4 for participant-stayers. The log likelihood function associated with Equations 1 and 2 is as follows:

$$L^* = \sum_{i=1}^{N4} \ln BN\,[\beta_1'X_{1i}, \beta_2'X_{2i}, \rho] + \sum_{i=1}^{N2} \ln BN\,[\beta_1'X_{1i}, -\beta_2'X_{2i}, -\rho] + \sum_{i=1}^{N_1+N_3} \ln \Phi[-\beta_2'X_{2i}]$$

where BN is the bivariate normal standard distribution and Φ is the univariate normal standard distribution (this is the effect of "collapsing" two cells because of a lack of residential relocation data on the dropouts). This function can be used to estimate the regression coefficients in Equations 1 and 2 and the correlation coefficient between their two error terms (ρ). One can use either maximum likelihood or any other method as in work by Amemiya (*10*). A pseudo t-test associated with ρ can be used to verify that a bivariate probit model is a more appropriate formulation than two univariate probit models for Equations 1 and 2. Alternatively, a nested likelihood ratio chi-square test can also be applied.

The second objective of estimation in this paper is to obtain consistent estimates of β_3 and to examine the sign and magnitude of the parameters in Equation 5. The selectivity "problem" arises when $E(\epsilon_{3i} \mid X_{3i}, Y_1 \times Y_2) \neq 0$ and OLS is used to estimate Equation 3a. For the cells in which Y_{3i} is observed a trivariate normal density applies and the related likelihood function is analogous to the complete cell membership discussed before. In this paper, instead of employing a method that involves trivariate normal densities, an alternative procedure that produces equally consistent estimates is used.

The method was devised by Tunali (*11*) and is the double-selection analog of the Heckman single-selection correction method (called the Tunali method here). It is a two-step procedure, which at the first step employs maximum likelihood estimation for Equations 1 and 2 to obtain consistent estimates of the two correction terms (λ_1 and λ_2). At the second step, the estimates of the λ's are used to correct for specification error (emerging from selection bias) in the regression of Y_{3i}. The system of the equations to consider is given by Equation 1, Equation 2, and the following, augmented continuous dependent variable regression:

$$Y_{3i} = \beta_3'X_{3i} + \gamma_1\lambda_1 + \gamma_2\lambda_2 + \sigma_3\,\epsilon_{3i}^* \tag{3b}$$

where γ_1 and γ_2 are functions of σ_3 and the correlations in Equation 4 and can be estimated by least-squares regression. λ_1 and λ_2 are the double-selection analogs of the Mill's ratios in single selection. The λ's are functions that involve data from the selection rules in Equations 1 and 2. ϵ_{3i}^* is a heteroskedastic error term. When OLS is applied to Equation 3b the usual standard errors of the coefficient estimates are biased. This is allowed for by "correcting" the OLS standard error estimates used for hypothesis testing. Estimation of the correction terms (λ_1 and λ_2), their associated coefficients (γ_1 and γ_2), and the associated standard error follows LIMDEP (*12*), which follows the Heckman two-step method.

DATA

PSTP is the first general-purpose urban transportation survey in the United States. The major goals of the panel are to (*a*) track changes in employment, work characteristics, household composition, and vehicle availability; (*b*) monitor changes in travel behavior and response to changes in the transportation environment; and (*c*) examine changes in attitudes and values of transit and nontransit users. PSTP includes household, person, trip, and attitude information of four waves, with each pair of waves a year apart. The first-wave data collection took place from September to early December 1989. The second-wave survey was conducted in the fall of 1990. An extensive description of the panel is provided by Murakami and Watterson (*8*).

In this paper, the analysis uses selected travel diary information from the first two waves. The travel diary includes continuous 48-hr activities (excluding the in-home activities) for each wave. It includes every trip a person made in 2 days. Each trip was characterized by trip purpose, type, mode, start/end time, travel duration, origin/destination, and distance. From this data set out-of-home activity engagement information can be derived using the trip purposes. The raw data were "cleaned" from any inconsistencies and the records with complete information are used here.

In the original data set, trip purposes are classified into eight different types (work, school, college, shopping, personal business, appointments, visiting, and free time). Models for all the activities considered together (sum of activities) and by grouping activities in a few categories were estimated. Assuming that a household, within a given 24-hr period, prioritizes its activity participation according to the relative importance of each activity, a natural grouping would be the following hierarchy (with a decreasing degree of constraint and importance): subsistence (work, school, college), maintenance (shopping, personal, appointments), and leisure (visiting, free-time) activities. The models treated for selectivity are models of subsistence frequency, maintenance frequency, and leisure frequency, each considered separately. A fourth model representing the sum of all activities is also estimated to identify possible "loss" of information when usual trip generation models are formulated.

Information on residential relocation was also collected within the panel. The data analyzed in this paper are from 1,662 households, of which 1,313 (79 percent) participated in both panel waves and 349 (21 percent) participated in Wave 1 only. From among the 1,313 participants, 111 (8 percent) changed residential location between the two waves, whereas 1,202 (92 percent) did not.

EMPIRICAL EXAMPLE

An application of the double-selection model mentioned earlier is provided here to address two related issues. The first is with respect to potential sample biases in a Wave 2 sample emerging from selective attrition and possible selective residential relocation. Restoring representativeness in the PSTP can be performed using weights derived from the bivariate probit model (joint attrition and relocation) mentioned earlier. Sample weights for subsequent waves aim at recreating population representativeness in the panel. One can use the results up to this point as seen elsewhere (*9*) to create sample weights. The second is with respect to consistent parameter estimation for the regression equations representing activity participation in Wave 2. The sample in Wave 2 contains only partial information on the population because of the double selection with part of the observations containing incomplete classification (i.e., the dropouts cannot be classified into movers and stayers). This affects the expectation of the error term in Equation 3b. The Tunali double-correction terms can be used to gain coefficient consistency. The definition of variables, cell frequencies, and average characteristics per group are presented in Table 1. The average value for each variable used in the models is presented separately for each of the three groups considered in this paper.

The first model of interest is the bivariate probit model with selection. Table 2 contains the single equation results, that is, estimates of two independent univariate probit equations ($\rho = 0$) and the bivariate probit estimates ($\rho \neq 0$). Model specification was defined mainly on the basis of past results using a similar data set on attrition and relocation and indications from past literature.

The regression parameter estimates are consistent (in terms of signs and relative magnitude) in the two models. With respect to the attrition model, as expected, the results confirm previous research using a similar data set. Households with a higher car ownership level, higher employment, and longer duration of residence in Wave 1 are more likely to participate in both waves of the panel. Confirming the usual tendency reported in other surveys, low-income households, single-adult households, and childless households with relatively young household composition tend to drop out after the first panel wave. People recruited via random digit dialing (in the sample analyzed here 92 percent are recruited via random digit dialing and 8 percent by special choice-based methods) tend to stay in the panel. The relocation equation exhibits agreement between the single-equation estimation and bivariate probit estimates. The household life-cycle stage is an important determinant of relocation (that is, households at their earlier stages are more likely to move than at their later stages). This is reflected by the coefficients of the two variables representing the number of children in the household. An interesting result is that the residence tenure (the dummy variable associated with 5 years or more in the current residence) has a negative coefficient. This may be an indication that, as residence tenure increases, the household is less likely to move. All three indi-

TABLE 1 Definition of Variables and Sample Characteristics

Variable	Description
FEMALESx	Number of females in the household in wave x
DRIVERSx	Number of drivers in the household in wave x
WORKERSx	Number of workers in the household in wave x
KID(0-5)x	Number of children whose age is less than five years in wave x
MIDINCOMEx	Dummy variable =1 if annual household income is between $15,000 and $50,000 in wave x ; 0 otherwise
HIGHINCOMEx	Dummy variable =1 if annual household income is more than $ 50,000 in wave x ; 0 otherwise
SGLADULTx	Dummy variable =1 if household has only one adult less than 35 years and no children in wave x ; 0 otherwise
YNGADULTSx	Dummy variable =1 if household has two or more adult less than 35 years and no children in wave x ; 0 otherwise
MIDADULTSx	Dummy variable =1 if household has two or more adult aged 35-64 years and no children in wave x ; 0 otherwise
YRHOME(0-1)x	Dummy variable =1 if number of years in current residence is less than one year in wave x ;0 otherwise
YRHOME(1-5)x	Dummy variable =1 if number of years in current residence is between one and five years in wave x ;0 otherwise
YRHOME(5-10)x	Dummy variable =1 if number of years in current residence is between five and ten years in wave x ;0 otherwise
ONECARx	Dummy variable =1 if household owns one car in wave x ; 0 otherwise
TWOCARSx	Dummy variable =1 if household owns two cars in wave x ; 0 otherwise
MULTICARSx	Dummy variable =1 if household owns more than two cars in wave x; 0 otherwise
TELE-RDD	Dummy variable =1 if household recruited by telephone random digit dialing
HHLDSIZEx	Household size in wave x
KINGx	Dummy variable =1 if residence locate in King County in wave x ; 0 otherwise
PIERCEx	Dummy variable =1 if residence locate in Pierce County in wave x ; 0 otherwise
SNOHOMISx	Dummy variable =1 if residence locate in Snohomish County in wave x ; 0 otherwise
Relocation	Binary Choice Dependent Variable = 1 if household has moved in second wave in panel
Attrition	Binary Choice Dependent Variable = 1 if household continues to participate in second wave of panel

Note : x=1 and 2 in variables indicate wave 1 and wave 2.

(*continued on next page*)

TABLE 1 (*continued*)

	Sample Mean of Variables		
	Participants and stayers	*Participants and movers*	*Non-participants in Wave2*
FEMALES1	.975	.883	.966
DRIVERS1	1.735	1.523	1.653
WORKERS1	1.256	1.243	1.206
KID(0-5)1	.218	.297	.310
KID(0-5)2	.204	.288	
MIDINCOME1	.651	.685	.590
HIGHINCOME1	.194	.153	.198
MIDINCOME2	.523	.478	
HIGHINCOME2	.333	.396	
SGLADULTS1	.029	.117	.063
YNGADULTS1	.050	.153	.109
MIDADULTS1	.292	.207	.238
YRHOME(0-1)1	.122	.297	.241
YRHOME(1-5)1	.333	.469	.384
YRHOME(5-10)1	.546	.234	.375
ONECAR1	.229	.324	.264
TWOCARS1	.449	.414	.415
MULTICARS1	.289	.225	.255
ONECAR2	.216	.270	
TWOCARS2	.426	.297	
MULTICARS2	.297	.162	
TELE-RDD	.950	.793	.560
HHLDSIZE1	2.575	1.820	2.752
HHLDSIZE2	2.513	1.182	
KING1	.400	.541	.410
PIERCE1	.207	.109	.261
SNOHOMIS1	.262	.198	.249
Frequency	1202	111	349

cators of county of residence (King, Pierce, and Snohomish) show that the movers are more likely to be from the fourth county (Kitsap). The most important result here is the lack of significance (and relatively small magnitude) of the error correlation coefficient between relocation and attrition (ρ). (The use of this method provides for clearer indications about the relationship between relocation and attrition. The usual caveat on the estimated standard error of ρ applies as well.) Similar to previous results on attrition and mode choice (*9*) and based on this paper, attrition is not correlated with other choices households make.

The results here provide some guidance on sample weight creation procedures. The results also reinforce past approaches to "sequential" and independent weight creation, that is, deriving weights that transform the Wave 2 panel sample into a representative sample by sequentially applying single-source derived weights to account for each source-specific sample bias.

The estimated bivariate probit model is used to create consistent estimates for the λ's for two out of the four cells in Figure 1. The first, corresponding to ($Y_1 = 0$, $Y_2 = 1$), represents the panel participants in both waves who did not relocate (participant stayers) and the second, corresponding to ($Y_1 = 1$, $Y_2 = 1$), represents the panel participants in both waves who relocated (participant movers). Four models are presented here for Y_3. The first three, in Table 3, depict 2-day household activity participation frequencies for subsistence, maintenance, and leisure. The fourth model depicts the sum of subsistence, maintenance, and leisure (called the total frequency of household activity participation resembling a trip generation model).

Table 3 provides a comparison between OLS and the Tunali method. The specification of all the models is the same in an attempt to provide a common basis for comparison. Alternative specifications provided similar results and are not presented here. Some of these models are underspecified, and this has an effect on the significance of the correction terms (*11*).

The standard errors of the coefficient estimates reported here (denominators in the "t-stats") are also corrected for selection on the basis of the method reported in LIMDEP (*7*). This is the same method used by Tunali for the two groups analyzed here (*11*). A consistent estimator is used for the standard error of the regression equation (Equation 3b) and is based on the usual OLS residuals with a correction (*12*). Estimates for the error correlation coefficients (ρ_{13} and ρ_{23}) are obtained with algebraic manipulations that involve the coefficients of the correction terms, the correlation in the bivariate probit model, and the standard error of the regression in Equation 3b. Unfortunately, in practice, this may produce correlation coefficients that are not within the unit circle, posing great difficulties in interpreting the coefficients.

With respect to the subsistence equation, one can observe a general agreement in the signs and relative magnitudes of the coefficients between the OLS and the Tunali models for both groups,

TABLE 2 Residential Relocation and Panel Attrition Models

	Univariate Probit		Bivariate Probit	
	Coef.	"t-stat"	Coef.	"t-stat"
Relocation				
Constant	-.831	-3.680	-.959	-2.916
FEMALES1	.015	.121	.007	.067
DRIVERS1	-.212	-2.104	-.200	-2.028
KID(0-5)1	.118	1.398	.100	.965
MIDINCOME1	-.010	-.067	.009	.051
HIGHINCOME1	-.075	-.386	-.067	-.315
SGLADULT1	.565	2.572	.513	1.845
YNGADULTS1	.597	3.131	.533	2.027
MIDADULTS1	.087	.624	.083	.559
YRHOME(5-10)1	-.511	-4.289	-.478	-3.328
TELE-RDD	-.050	-1.469	-.046	-1.226
KING1	.021	.135	.031	.187
PIERCE1	-.383	-1.926	-.367	-1.708
SNOHOMIS1	-.121	-.682	-.113	-0.611
Attrition				
Constant	.869	4.431	.880	4.376
ONECAR1	.405	2.201	.396	2.095
TWOCARS1	.582	3.098	.569	2.963
MULTICARS1	.574	2.896	.558	2.785
WORKERS1	.135	2.534	.132	2.448
YRHOME(0-1)1	-.425	-4.065	-.435	-4.158
YRHOME(1-5)1	-.209	-2.489	-.204	-2.403
LOWINCOME1	-.188	-1.522	-.198	-1.580
HIGHINCOME1	-.116	-1.229	-.114	-1.211
SGLADULT1	-.457	-2.533	-.455	-2.519
YNGADULTS1	-.532	-3.614	-.526	-3.575
MIDADULTS1	-.211	-2.151	-.210	-2.207
HHLDSIZE1	-.174	-4.800	-.172	-4.811
TELE-RDD	.037	1.618	.037	1.497
ρ (1,2)			.310	.415

<u>***Goodness-of-fit Statistics***</u>

	Univariate Probit		Bivariate Probit
Relocation			
Log-likelihood	-343.37	Log-Likelihood	-1161.03
Restricted Log-likelihood	-380.40	Restricted Log-likelihood	-1234.57
Chi-Squared (df=13)	74.06	Chi-squared (df=27)	147.08
Attrition			
Log-likelihood	-817.82		
Restricted Log-likelihood	-854.17		
Chi-Squared (df=13)	72.70		

that is, participant stayers and participant movers. For the stayers, as car ownership increases, the households are more likely to participate more frequently in these activities. The movers provide the exact opposite relationship between car ownership and activity frequency (but with loss of significance). Higher-income households tend to have higher frequencies, and the presence of young children inhibits participation in these activities (presumably to school and college). As expected, as household size increases, subsistence frequency also increases. Household size may also capture the effect of employed people in the household. In the OLS model its associated coefficient is unity; this was increased by 25 percent when the regression was corrected for selectivity. In Equation 3b a variable X influences Y in two ways: directly via its associated β and indirectly through the correction terms (λ's), and this explains the difference between the two models. The significance of the γ's indicates substantial selectivity bias for the participant stayers, whereas this is not true for the participant movers.

The maintenance activity frequency provides similar indications to the subsistence model. An exception to this is the effect of income. It appears that lower-income households are more likely to engage in this type of activity than higher-income households. One correction is significant for the participant-stayer model, and none is significant for the participant-mover model. Evidence of selectivity is present or absent depending on the type of frequency examined. This is even clearer when one examines the results in the leisure frequency models. None of the correction terms is signifi-

TABLE 3 Activity Regression Models

	SUBSISTENCE ACTIVITY FREQUENCIES				MAINTENANCE ACTIVITY FREQUENCIES				LEISURE ACTIVITY FREQUENCIES			
	OLS (without correction)		Tunali method (with correction)		OLS (without correction)		Tunali method (with correction)		OLS (without correction)		Tunali method (with correction)	
	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"	Coef.	"t-stat"
Participants and stayers												
Constant	1.547	3.157	2.342	2.497	2.836	4.675	3.782	4.159	1.791	3.650	2.093	3.239
ONECAR2	-1.308	-2.590	-1.835	-2.791	-.724	-1.158	-.662	-.979	-.925	-1.830	-.908	-1.716
TWOCARS2	-1.145	-2.297	-1.742	-2.641	-.190	-.308	-.297	-.433	-1.009	-2.021	-1.045	-1.944
MULTICARS2	-.220	-.422	-.758	-1.118	-.194	-.301	-.442	-.621	-.462	-.886	-.543	-.969
HHLDSIZE2	1.005	9.255	1.252	8.751	1.372	10.204	1.251	7.983	1.322	12.161	1.285	10.282
KID(0-5)2	-1.151	-5.379	-1.341	-4.869	-.657	-2.481	-.317	-1.049	-1.396	-6.517	-1.289	-5.750
MIDINCOME2	.949	2.991	.773	2.342	-.287	-.730	-.103	-.261	.243	.764	.300	.940
HIGHINCOME2	2.208	6.438	2.106	5.710	-.091	-.215	.046	.108	.728	2.120	.771	2.245
R^2		*.159*		*.173*		*.122*		*.141*		*.153*		*.156*
λ_1			-6.415	-2.296			6.961	2.490			2.175	1.330
λ_2			-4.779	-2.148			.946	.465			.281	.204
ρ_{13}			-1.693				2.135				.599	
ρ_{23}			- .958				-.388				-.122	
σ_3			2.914				3.122				3.488	
Participants and movers												
Constant	-.306	-.189	-6.603	-.721	3.729	3.453	-7.806	-.585	4.090	3.617	-8.393	-.609
ONECAR2	-.592	-.390	-.015	-.008	- .843	- .833	.345	.138	- .858	- .810	.525	.205
TWOCARS2	-.906	-.562	1.756	.759	-2.725	-2.534	-.939	-.338	-1.120	-.994	.986	.346
MULTICARS2	-2.293	-1.263	-1.226	-.439	-1.222	-1.010	1.024	.291	-.060	-.047	2.708	.748
HHLDSIZE2	1.315	2.642	.906	1.180	.989	2.981	.226	.241	.684	1.969	-.151	-.155
KID(0-5)2	-.839	-1.055	-.450	-.371	.654	1.233	1.318	.816	-.305	-.548	.376	.227
MIDINCOME2	2.056	1.668	2.493	1.539	-1.964	-2.389	-1.019	.543	-2.051	-2.383	-.919	-.482
HIGHINCOME2	4.507	3.003	4.724	2.341	-.531	-.531	.099	.040	-1.025	-.978	-.168	-.066
R^2		*.199*		*.212*		*.249*		*.330*		*.138*		*.239*
λ_1			2.782	.689			4.720	.772			4.826	.763
λ_2			6.242	.585			13.105	.847			15.438	.961
ρ_{13}			0.152				.080				.005	
ρ_{23}			0.964				1.410				1.484	
σ_3			5.579				8.260				9.393	

cant in these models. The higher standard error of the regression equation for all the models of the participant movers indicates higher variation in activity participation when compared with the stayers. The possibility of this functioning as an indicator of misspecification is discarded mainly because of the lack of significance of the correction terms. Table 4 presents a model with dependent variable the sum of the three activities in Table 3. The results parallel the indications of the subsistence models (signs of coefficients and relative magnitude). In general, the coefficients are higher because of the higher values of the dependent variable. Unlike the subsistence model, the correction terms are not significant. This leads to the conclusion that the effects of selectivity can be better captured by considering frequency of activity types separately. A refinement of the method here is under way using more complete specifications for the regression models, for example, incorporating transportation system attributes and better descriptors of household composition. In addition, a sensitivity analysis of the method to the specification of the bivariate probit model is also needed. The residential relocation model needs to consider additional determinants of relocation. This is also left as a future task.

SUMMARY AND CONCLUSIONS

A method to account for the possible simultaneity of multiple selection in panel surveys is presented in this paper. Two sources of selectivity are considered together—residential relocation and panel attrition—using a bivariate probit model that considers the lack of observed residential relocation for the sample of the dropouts. The method can be applied to derive sample weights for subsequent panel waves and to create correction terms that can be used to obtain consistent estimates of activity participation equations.

In the first two waves of PSTP, residential relocation and attrition are not correlated. This supports the use of sequential weighting for the Wave 2 sample. The application of correction terms to regression models of activity participation provided many insights. The effect of selectivity on activity participation may depend strongly on the type of activity analyzed. When all the activity types are aggregated to form a single model of frequencies (e.g., a trip generation model) selectivity bias may appear to be absent. When activities are considered separately, selectivity bias is present in some equations.

Many extensions and improvements are needed in the method presented here. The models need to be specified in radically different ways and the results need to be compared with those from this paper. This will provide some guidance on the effects of misspecification on selectivity equations. The Tunali method provides consistent estimates but is not fully efficient. Efficiency loss is associated with the two steps involved. A full information maximum likelihood method would be a suitable alternative. The three activity equations in Table 3 have been considered separately. It is well known that participation in one type of activity influences participation in another. This can be easily modeled by creating a system of equations and applying the Tunali method to the system. During an earlier review it was suggested that potential improvements in the method here may emerge from alternate forms of the activity frequency equations. One could extend the method using a system of "TOBIT" models with

TABLE 4 Sum of Activity Frequencies (Trip Generation)

	OLS (without correction)		Tunali method (with correction)	
	Coef.	"t-stat"	Coef.	"t-stat"
Participants and stayers				
Constant	6.174	5.653	8.217	5.754
ONECAR2	-2.957	-2.627	-3.404	-2.898
TWOCARS2	-2.344	-2.110	-3.085	-2.578
MULTICARS2	-.876	-.754	-1.744	-1.397
HHLDSIZE2	3.699	15.283	3.788	13.517
KID(0-5)2	-3.203	-6.719	-2.947	-6.088
MIDINCOME2	2.905	1.280	.970	1.361
HIGHINCOME2	2.845	3.722	2.923	3.797
R^2		*.243*		*.249*
λ_1			2.722	.788
λ_2			-3.553	-1.160
ρ_{13}			.454	
ρ_{23}			-.522	
σ_3			8.426	
Participants and movers				
Constant	7.512	2.837	-22.802	-.656
ONECAR2	-2.294	-.925	.855	.132
TWOCARS2	-2.939	-1.114	1.802	.251
MULTICARS2	-3.456	-1.164	2.505	.275
HHLDSIZE2	2.989	3.673	.981	.403
KID(0-5)2	-.490	-.377	1.244	.297
MIDINCOME2	-1.958	-.972	.555	.115
HIGHINCOME2	2.951	1.203	4.655	.726
R^2		*.223*		*.319*
λ_1			12.328	.773
λ_2			34.785	.864
ρ_{13}			.072	
ρ_{23}			1.428	
σ_3			21.693	

double "Probit" selectivity. With respect to model specification, the method here can be improved by the inclusion of level-of-service variables that are currently created for PSTP. In addition, for the relocation model a more in-depth specification analysis is needed.

ACKNOWLEDGMENTS

The authors thank the Puget Sound Regional Council for providing the data. Funding was provided by the U.S. Department of Transportation Region 3 Mid-Atlantic Universities Transportation Center. Comments by anonymous reviewers are also acknowledged.

REFERENCES

1. Heckman, J. J. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement,* Vol. 5, No. 4, 1976, pp. 475–492.
2. Heckman, J. J. Sample Selection Bias as a Specification Error. *Econometrica,* Vol. 47, 1979, pp. 153–161.
3. Mannering, F. L. Selectivity Bias in Models with Discrete and Continuous Choice: An Empirical Analysis. In *Transportation Research Record 1085,* TRB, National Research Council, Washington, D.C. 1985, pp. 58–62.
4. Kitamura, R., and P. H. L. Bovy. Analysis of Attrition Biases and Trip Reporting Errors for Panel Data. *Transportation Research A,* Vol. 21, 1987, pp. 287–302.
5. Hensher, D. A., P. O. Bernard, N. C. Smith, and F. W. Milthorpe. *Modeling the Dynamics of Car Ownership and Use: A Methodological and Empirical Synthesis.* Working Paper 32. Transport Research Group, Macquarie University, Sydney, Australia, 1987.
6. Monzon, J., K. G. Goulias, and R. Kitamura. Trip Generation Models for Infrequent Trips. In *Transportation Research Record 1220,* TRB, National Research Council, Washington, D.C. 1989, pp. 40–46.
7. Maddala, G. S. *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, England, 1983.
8. Murakami, E., and W. T. Watterson. Developing a Household Travel Panel Survey for the Puget Sound Region. In *Transportation Research Record 1285,* TRB National Research Council, Washington, D.C., 1990, pp. 40–46.
9. Kitamura, R., R. M. Pendyala, and K. G. Goulias. Weighting Methods for Choice-Based Panels with Correlated Attrition and Initial Choice. In *Transportation and Traffic Theory* (C. F. Daganzo, ed.), Elsevier Science Publishers, Amsterdam, The Netherlands, 1993, pp. 275–294.
10. Amemiya, T. *Advanced Econometrics.* Harvard University Press, Cambridge, Mass, 1985.
11. Tunali, I. A General Structure for Models of Double-Selection and an Application to a Joint Migration/Earnings Process with Remigration. *Research in Labor Economics,* Vol. 8, Part B, 1986, pp. 235–282.
12. *LIMDEP User's Manual and Reference Guide,* Version 6., Econometric Software, Inc., New York, 1992.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Influence of Dutch Mobility Policy on Emancipation Process for Women and Men

MARIËTTE POL, DICK ZOUTENDIJK, AND URSULA BLOM

In the literature on emancipation and mobility it is often assumed that mobility policy impedes womens' chances acquiring equal opportunities in paid employment and activities outside the house. It is said that too much emphasis is placed on the use of the traffic and public transport system by men; therefore, the demands women have in using the transport system are not met. A literature research is conducted to determine the influence of the Dutch mobility policy on the emancipation process in The Netherlands. Further, the validity of this assumption is tested. Relevant policy measures with respect to mobility and traffic safety are evaluated on the basis of three emancipation indicators: (*a*) the increase in possibilities to participate in employment by women, (*b*) the increase in possibilities to do activities outside the house by women, and (*c*) the increase in housekeeping by men. One conclusion of this study is that mobility is just one of the many aspects influencing emancipation. Mobility can be seen as more or less facilitating emancipation, but never as decisive in itself. Another finding is that the effects of the mobility measures do not impede the emancipation process as is assumed in current literature. Most measures are rated neutral or slightly positive on the indicators. This can be explained by the fact that most policy measures tend to increase the attractiveness of modes of transport other than the car. Only few policy measures aim to decrease the use of cars directly.

In the last decades mobility in The Netherlands has taken flight. The most important reason for this growth is the general economic and population growth (*1*). Next to the increase in the economy, the massive growth in mobility is also influenced by social cultural shifts like the emancipation process. During the last decades the traditional sex role models have been changing, causing women to become more mobile (*2–4*).

It is obvious that the growth in mobility (and, specifically, road traffic) has negative consequences. The mobility policy of the Dutch Ministry of Transport is formulated in the Second Transport Structure Plan. The major goals of this plan are as follows; first to restrain the increase in road traffic for travel considered less necessary and second, to encourage people to switch from the car to alternative modes of transport. In addition to this mobility policy, the central Dutch government policy is to stimulate the emancipation process (*5*). Therefore, the mobility and emancipation policies are apparently conflicting issues. In current literature on emancipation and mobility it is often assumed that the mobility policy impedes women's chances of acquiring equal opportunities in paid employment and activities outside the house (*4,6–8*). It is said that too much emphasis is placed on the use of the road traffic and public transport system by men and, therefore, the traveling needs of women are not met. Therefore, policy makers at the Ministry of Transport are concerned about the aims of emancipation policy in relation to mobility.

M. Pol and D. Zoutendijk, Traffic Test bv, Landjuweel 22, 3905 PG Veenendaal, The Netherlands. U. Blom, Ministry of Transport, Public Works and Water Management, Transport Research Center P.O. Box 1031, 3000 BA Rotterdam, The Netherlands.

Because of the apparently conflicting interests between emancipation and mobility, the Ministry of Transport initiated a (pilot) study to uncover the relationship between the mobility policy and the emancipation process (*3*). In this paper the results of this study are summarized and interpreted. The goal is to establish the extent to which the mobility policy actually impedes or stimulates the emancipation process. First, three "measurement tools" (emancipation indicators), derived from targets set by the Dutch central government in its emancipation policy as formulated in the policy program, are defined. Using these indicators, the travel behavior of women and men that can be related to the emancipation process is measured. Next the differences between women and men in aspects of mobility behavior are discussed. These differences give some insight into differences in demands for traffic and transport facilities. Finally, the influence of the Dutch mobility policy on the emancipation process is evaluated.

DATA USED IN STUDY

Data that directly relate policy measures to observable (emancipated) behavior are, unfortunately, not yet available because most policy measures have only very recently been implemented or are still in the process of being implemented. Therefore, the conclusions of this study should be considered only preliminary.

The study is carried out on the basis of data on mobility behavior of women and men. For this reason, the differences in mobility behavior of women and men are formulated and used.

METHOD

Defining Emancipation Indicators

The first step in the evaluation of policy measures is to formulate emancipation indicators. With these indicators the possibilities of changing mobility behavior of women and men have to be scored because of changes involving more emancipated activity patterns. In this study it was decided that emancipation indicators should involve observable mobility behavior that can be directly "measured." Therefore, the emancipation indicators are to be found in activities or in patterns of activities. The central aim of the emancipation policy as formulated in the Dutch policy program reads as follows:

> To promote the transformation process in current society, in which the differences based on sex are still institutionalized to a large extent, to a multiform society in which everyone, regardless of sex or marital state, has the possibility of living independently and in which women and men can realize equal rights, opportunities, liberties and responsibilities.(*5*)

In this same document, this emancipation aim is "translated" to the various departmental areas. The goal of the Ministry of Transport and Public Works is as follows:

> Realizing an optimal situation for traffic and transport, thereby creating the conditions for equal participation of both women and men in public and, especially, economic activities and in unpaid work in private life.(*5*)

This is the Dutch formulation of the targets set for the emancipation process. In other (Western) countries similar policy targets have been set. For example, the Swedish policy aims to build a society in which each individual participates fully in four major roles of daily life: the household/family role, the work/career role, the interpersonal/social role, and the leisure/recreation role (*9*).

In Dutch society the following three emancipation indicators (fulfilling the prerequisite to be observable activities) can be derived as follows:

1. The goal of "equal participation of both women and men in . . . economic activities" leads to the first indicator: "the increase of possibilities of women to participate in employment";
2. The aim of "equal participation of both women and men in public . . . activities" leads to the second indicator: "the increase of possibilities of women to do activities outside the house"; and
3. The goal of "equal participation of both women and men in . . . unpaid work in private life" results in the third indicator: "the increase of housekeeping and caring tasks by men."

Recent Developments in Activity Patterns of Women and Men

The proportion of the working population of Dutch women increased from 5 percent in 1960 (*10*) to 34 percent in 1975 and to 50 percent in 1991 (*2*). Working (more than 1 hr week) women spent an average of 21 hr week at their jobs. For reference, 75 percent of the Dutch men participated (at a minimum of 1 hr/week) in paid labor and spent an average of 36 hr/week at their jobs, both in 1975 and 1990. No shift was observed between 1975 and 1990 in this respect (*2*). In this research conducted by Batenburg and Knulst, people are considered workers when they spend at least 1 hr of paid work in the week the research was conducted. Congruently, people are engaged in housekeeping and child care if they spend at least 1 hr on these caring tasks in the week the research was conducted. In 1975, 99 percent of the Dutch women spent an average of 28 hr per week on housekeeping and child care tasks [referred to as "caring tasks" in this paper, which include private housekeeping and child care–volunteer aid (paid and unpaid) and are considered activities outside the house]. In 1990, a slight decrease was observed to an average of 25 hr per week. Of the Dutch men in 1975, 85 percent spent an average of 9 hr/week on caring tasks, whereas in 1990 89 percent of the Dutch men spent on average 10 hr/week on caring tasks. From these data it can be concluded that the increase in the amount of women participating in paid jobs has not resulted in an equivalent decrease in the amount of time spent by women on housekeeping tasks, or an equivalent increase in the amount of time in which men participated in caring tasks (*2*).

Until the 1960s, Dutch women (and especially married women) traditionally did not participate in paid employment. Instead of paid employment they were engaged mainly in housekeeping, child care, and volunteer aid. Having a job is appreciated much more than housekeeping, child care, or volunteer aid. Traditionally the prime task of men in Dutch society is to participate in paid employment. Women, on the other hand, are primarily responsible for the less-appreciated activities. The high appreciation of paid jobs directs the emancipation process to a society in which women combine tasks. From this observation one can also conclude that the emancipation process has not been 'completed' yet because men did not emancipate at the same rate as women and are not yet taking over women's tasks (*2,9*). A significant body of evidence from the emancipation literature (*7,10–13*) indicates that more and more women need to combine paid work with caring tasks as a consequence of the emancipation process. Facilitating the combination of these tasks (for both women and men) is crucial to the realization of emancipation aims.

One of the main effects of combining tasks on mobility behavior is that people try to organize their commutes as efficiently as possible by combining different purposes into one trip. This phenomenon is called trip-chaining (*14,15*). An example of trip-chaining is escorting a child to school on the way to work. In Table 1 the percentages are presented of Dutch full-time employed mothers and fathers who combine trips while commuting. Far more mothers than fathers are making trip-chains. These findings are consistent with Dutch research carried out by Drooglever Fortuijn (*16*). She finds that women combine on the average 3.3 trips (round trip) during commuting, whereas men on the average combine 2.9 trips. Moreover, 84 percent of the purposes combined with commuting of women can be regarded as unpaid work (mostly caring tasks) compared with 79 percent of the men's trip-chains while commuting. Similar trends were found in the United States in 1989 and 1990 (*17*). In this study the authors found that women spend 1.3 times as much time combining household activity trips with commuting as did men.

TABLE 1 Percentage of Full-Time Employed Married Mothers and Full-Time Employed Married Fathers Who Make Trip-Chains While Commuting (*15*)

	Youngest Child Younger Than 6		Youngest Child Older Than 6	
	From Home to Work	From Work to Home	From Home to Work	From Work to Home
Married Mothers	28%	52%	23%	68%
Married Fathers	12%	24%	15%	12%

How Mobility Measures Affect Emancipation Indicators

The next step is to determine how relevant mobility measures described in the 'second transport structure plan' will influence the emancipation indicators and task combining. These measures can have a direct influence on activity patterns of women and men, thereby directly affecting the emancipation indicators. For example, decreasing the frequency of bus lines has a direct negative effect on the possibilities of activities outside the house for people depending on buses. Another (indirect) influence from these measures depends on the way the measures will affect trip-chaining, because trip-chaining is a fundamental need of people combining paid work and caring tasks. As a result, measures that impede the possibilities of trip-chaining will also decrease the possibilities of task combining. This, in turn, will impeded the possibility of women combining their caring tasks with paid work and the possibility of working men performing (more) caring tasks.

Differences in Travel Behavior Between Women and Men

In this section the focus is on the most important differences between women and men with respect to mobility behavior because these differences offer insight into differences between women and men in their demands for traffic and transport facilities.

1. Because women combine various tasks, they also combine different travel purposes in a single trip. This means that far more women than men are engaged in making trip-chains (*14,15,17*).
2. Typically, in The Netherlands, men commute during peak hours. Women also commute during off-peak hours because of part-time jobs. For the caring tasks (traditionally done by women), people travel often in off-peak hours (*2*). The number of trips for caring tasks during peak hours has increased in the last decades, but, at present, the largest part of these trips is still being made during off-peak hours in The Netherlands.
3. Although women and men make the same number of trips, the travel distances of men are much larger than those of women (*6,*18–20). In Table 2 travel distances and number of trips for the various modes of transport of Dutch women and men are presented.
4. Women and men use different modes of transport (Table 2). In The Netherlands most households own a single car that is used mainly by the man; far fewer woman than men use cars (*18*). The car is often used for commuting by men. Women, on the other hand, use more diverse modes of transport than men (*21*). Until recently, far fewer women than men had drivers' licenses (*18*).
5. Far more often than men, women are involved in caring tasks. Women escort (young) children to day-care centers, schools, doctors, hospitals, and other activities more frequently than do men (*4,6,14,21*).
6. As a result of their housekeeping tasks, women carry heavy shopping bags more often than men (*4*).
7. Far more women than men are involved in voluntary aid. Therefore, far more women than men escort elderly and sick people to doctors, hospitals, and other activities or just for a short walk. Voluntary aid is part of the indicator activities outside the house (*22*).
8. Women (far more than men) are concerned about their social safety. Their fear of assault, especially at night, restrains them in traveling (*21,23*).

Women and men differ in (almost) every aspect of mobility behavior. Having different mobility behavior implies that women have other kinds of travel demands than men. In this sense the assumption that men have other demands on transport than women is true. The assumption that too much emphasis is placed on the use of the transport system by men and that, therefore, demands women have are not met, has, however, not yet been proved. This different mobility behavior is for the greater part directly based on the traditional roles many Dutch women and men (still) have.

TABLE 2 Average and Percentage of Travel Distances and of Number of Trips per Day by Mode of Transport in 1992 of Men and Women 12 Years and Older

	Kilometers		Trips	
	Men	Women	Men	Women
Car driver	27,8 (61%)	8,6 (30%)	1,7 (47%)	0,8 (23%)
Car passenger	6,2 (14%)	10,5 (36%)	0,3 (8%)	0,6 (17%)
Public transport	5,9 (13%)	5,3 (18%)	0,2 (6%)	0,2 (6%)
Moped	0,4 (1%)	0,2 (1%)	0,05 (1%)	0,03 (1%)
Bicycle	3,4 (8%)	3,0 (10%)	0,9 (25%)	1,1 (31%)
Walking	0,9 (2%)	0,9 (3%)	0,5 (14%)	0,7 (20%)
Total	45,3 (99%)	28,8 (98%)	3,6 (101%)	3,5 (98%)

Note: Percentages do not add up to 100% because of rounding.

EFFECT OF MOBILITY MEASURES ON EMANCIPATION INDICATORS

In the Second Transport Structure Plan 35 policy areas are formulated to restrain the increase in road traffic. In the study by Pol and Zoutendijk (*3*) 12 relevant policy areas are evaluated. For the purposes of conciseness, the focus in this paper is on the three policy areas that have the strongest impact on the emancipation indicators:

1. Concentration of housing, employment, recreation, and other public facilities;
2. Urban remodeling schemes; and
3. Pricing policy.

Some of the policy measures have recently been implemented; others have plans in an advanced stage and some measures still do not have a (detailed) plan.

In the following section these three policy areas are evaluated with respect to the three emancipation indicators and use of the knowledge of the differences between women and men in aspects of mobility behavior.

Policy Area: Concentration Policy on Housing, Employment, Recreation, and Other Public Facilities

The policy on urban planning and (regional) transport aims to reduce the need to travel (by shortening travel-to-work distances) and promote public transport and cycling. The target scenario is as follows,

- Serving every major housing development from 1995 onwards by high-grade public transport; and
- Equipping residential areas with appropriate public transport services and bicycle infrastructure.

This concentration policy makes facilities more accessible for everyone traveling by bike or public transport. Those who will profit most are people who do not have access to a car. In the Netherlands far fewer women than men have access to a car (*18*).

This concentration policy also facilitates the combination of paid work and caring tasks for women and men. Community facilities such as day-care centers, schools, and shops are allocated to be situated in a way that optimizes the accessibility by bicycle and public transport. Consequently, no car is required and less time is needed to travel to these facilities. For a lot of married women with (young) children, participation in paid employment is highly dependent on the ability to combine work with caring tasks (*7,8,14*). This concentration policy has probably a bigger positive effect on the participation in paid employment by women than on the participation in caring tasks by men. This effect can be concluded from the fact that the increase in the number of women participating in paid jobs has not resulted in an equivalent increase in men participating in caring tasks.

Another objective of the employment location policy is improved accessibility of paid work with short commuter distances and commuter times. To what extent this last objective can be reached is not certain because (in The Netherlands) more and more people are willing to travel long distances to find a job that fits their education and ambitions.

An unwanted (by the Ministry of Transport) side effect can occur in households consisting of two adults who own a single car that is used for commuting by the man. When this man makes the transfer from commuting by car to commuting by bicycle or public transport the car will be available for the woman. This availability of the car will only positively influence the emancipation indicators if she uses the car in a way that alters her activity pattern, such as in participating in paid employment, increasing the amount of working hours per week, or becoming involved in (more) activities outside the house. On the other hand, if she uses the car for making caring tasks more convenient, it will not have a positive effect on the emancipation indicators because her activity pattern does not change.

Policy Area: Urban Remodeling Schemes

A reappraisal of the car's role in Dutch cities will lead to the target scenario of discouraging nonessential traffic in the inner city. The use of the car for short journeys will be discouraged by the implementation of a coarse-mesh structure for the road network using ring and loop systems in which urban and residential areas only can be reached by car from an external ring route. Bicycles, public transport, and delivery vehicles will be advanced through keeping open the internal routes for these modes of transport.

As a consequence of the urban remodeling policy, community facilities such as schools, day-care centers, and shops will be less accessible by cars and more accessible by walking, biking, or public transport. This can have some negative effect on the indicator "participation of women in employment" because the use of a car to escort very young children to day-care centers can be crucial. Mothers may consider the amount of organizing and time they need to bring and pick up their babies and toddlers to and from day-care centers too high a barrier to continue their work or to start a new job. This impeding effect can be overcome by an integration of this policy area with the policy area of "concentrating housing, employment, recreation, and other public facilities" and by improving the service of the public transport in off-peak hours.

It is possible that the travel behavior of men will be more constrained than that of women because far more men than women rely on car mobility (*18*). The mobility of women can be enhanced because they, more often than men, use a bicycle or public transport. The urban remodeling policy may have a positive effect for women participating in activities outside the house because the accessibility of these facilities will be improved also.

To require a modest role for the car as the mode of transport the notion of "low-car-density" residential areas with parking facilities on their fringes has emerged. This concept probably has the same negative effect as that described earlier because it impedes the escorting of young children to day-care centers by car.

A positive effect of low-car-density residential areas can be expected as well, however. Cars hinder children in their ability to play outdoors because of the risk of accidents and the parking space cars require. A prime advantage of low-car-density residential areas is that children can play outdoors without adult escort. This will ease the caring tasks of mothers who, in Dutch society, as said before, are often responsible for caring tasks. This might have a positive effect on the emancipation indicators depending on the personal interpretation of this free time by women. For instance, if this time is used for education or personal development, it will have a positive effect on the indicator "the increase of possibilities of women to do activities outside the house."

Policy Area: Pricing Policy

Pricing policy consists of the increase of the price of certain modes of transport. Measures will be taken to substantially increase the variable costs of the car, to introduce charges on certain sections of the road network at certain times, and to keep the costs of car use in line with the costs of public transport.

Pricing policy is one of the key tools to achieving the target scenario of deterring the use of cars on certain sections of the road network at certain times to reduce congestion problems. Raising fuel tax in combination with a cut in annual vehicle tax will deter car use in general. Tolls and peak-hour surcharge will deter car use in peak hours on congestion points.

The effect of pricing policy is uncertain because most of the pricing policy measures have not yet been introduced in The Netherlands. The pricing policy aims to stimulate car drivers into avoiding the expensive and busy rush hours. Further, it aims to stimulate car drivers to switch and use public transport or a bicycle.

An unwanted possibility of the pricing policy is that people might just pay the price and continue to use their cars. Increasing costs have to be paid by the households of these car drivers who do not change their behavior. These costs have to be compensated in households with a rather small income. One way of compensation is to reduce car use for the kinds of traveling considered less necessary. This will more often lead to reduced car use by women than by men because women are more willing to curtail their car use than are men. This may, therefore, have a negative effect on outdoor activities by women or even on participating in paid employment by women.

A lot of women have part-time jobs and, therefore, have small relative incomes. If women who work are restricted to commuting during peak hours, the pricing policy can have impeding effects on the participation of married women in employment. In this particular case, when the household is not dependent on the salary of the woman, the costs of traveling are not counterbalanced by income.

When car drivers do change their commuting behavior in transferring from car to other modes of transport, this often means that the car will be left available for the woman. As discussed before, the implications of the availability of the car will positively influence the emancipation indicators only if she uses the car in a way that alters her activity pattern. On the other hand, if she uses the car for making caring tasks more convenient, this will not have a positive effect on the emancipation indicators because her activity pattern does not change.

DISCUSSION

In another section, Effect of Mobility Measures on Emancipation Indicators, a description is given of the expected effects of the policy measures as presented in policy areas. This description is translated into scores on the three emancipation indicators presented in Table 3. The range of the scores in this table is from ++, meaning a very positive effect to −−, meaning a very negative effect. A 0 indicates that the expected effect of the policy area on the specific indicator is neutral.

The scores indicated in Table 3 suggest that mobility measures do have only a small influence on the emancipation process. On a scale from ++ to /−− the scores of the policy areas mostly have neutral (0) or slightly positive (0/+) or slightly negative (0/−) effects. It is true that aspects of transport such as supply, comfort, and price can certainly (indirectly) impede or stimulate participation in paid work and activities outside the house by women, but aspects of transport are only a decisive factor in a restricted number of cases.

The scores indicated in Table 3 also suggest that most measures set by the Dutch Ministry of Transport as formulated in the Second Transport Structure Plan have, in fact, a neutral or slightly positive effect on the realization of the aims set by the emancipation policy. Only a few measures tend to have a negative impact on the eman-

TABLE 3 Scores on Emancipation Indicators

Policy Area	Effect Measure	Paid Employment by Women	Activities Outside the House by Women	Caring Tasks by Men
Concentrating housing, employment and other public facilities	To reduce travel needs and to promote cycling and public transport	+	0/+	0/+
Urban modeling schemes	To discourage traffic in inner cities	0/-	0/-	0
	To advance cycling and public transport	0/+	0/+	0
	To create low-car-density residentials	0/-	0/+	0
Pricing policy	To restrain car use	0/-	0/-	0

cipation indicators. Especially the scores on the third emancipation indicator, "the increase of housekeeping and caring tasks by men," indicate that policy measures hardly have any impact on the emancipation process of men.

CONCLUSIONS

Used Methods

In this study, the aim is to evaluate the effects of policy measures on emancipated behavior in an objective way. For this purpose three emancipation indicators from the goals act by the central government on emancipation policy are derived. Using these indicators, the potential changes in emancipated activity induced by mobility policy can be "measured." Three policy areas are found suitable for evaluation using emancipation indicators: (*a*) the increase in possibilities to participate in employment by women, (*b*) the increase in possibilities to do activities outside the house by women; and (*c*) the increase of housekeeping by men. Unfortunately, there is still insufficient direct data about changes in mobility patterns. Consequently, the discussion and conclusions are based on current literature on the mobility behavior of women and men. The conclusions of this work should, therefore, be viewed as preliminary only. Future research should aim to acquire direct data on the consequences of mobility policy on emancipated behavior and to further explore these consequences in terms of the emancipation indicators developed in this paper. For each policy area the effects of the policy measures are described. This description is into scores on the three emancipation indicators.

Conclusions on Basis of Scores on Emancipation Indicators

From this survey the following conclusions are drawn. Mobility measures have only a modest influence on the emancipation indicators. It is true that aspects of transport certainly (indirectly) impede or stimulate participation in paid work and activities outside the house by women, but aspects of transport are decisive arguments in only a few cases. This can be explained by the fact that emancipated behavior requires a change of attitude. This change of attitude cannot be achieved solely by mobility policy. If, however, an attitude change were to occur, mobility policy would be able to both facilitate and to impede the emancipated behavior.

Most policy measures as formulated in the Second Transport Structure Plan can be concluded to have a neutral or slightly positive effect on the realization of the aims set by the Dutch emancipation policy. Few measures tend to have a negative impact on the formulated emancipation indicators. Thus, the commonly made assumption that mobility measures have an impeding effect on emancipation seems unwarranted. This finding can be explained by the fact that most policy measures are pull measures, that is, they reward the desired behavior by improving the attractiveness of alternative modes of transport. Far fewer policy measures are push measures for which there is a penalty for undesired behavior, such as decreasing the attractiveness of car use in pricing policy, whereas the expected effect of push measures is stronger, that is, to deter the increase in car use, than the expected effect of pull measures.

Another conclusion of this study is that almost none of the evaluated measures have any impact on the increase in caring and household tasks by men. This finding can also be explained by the fact that an attitude change by men is required, which cannot result from transport policy. Again, if an attitude change does take place, the measures that facilitate task combination will have a positive effect on men doing caring and household tasks and women participating in paid jobs.

In this study, the expected effects of the intended policy measures are discussed, as formulated in the policy program Second Transport Structure Plan on the emancipation process in The Netherlands. The effects of the current Dutch mobility situation on the emancipation process were not addressed. Further research evaluating the current mobility situation on the formulated emancipation indicators is needed. Potentially this research could give insight into the aspects of the current mobility situation that have impeding effects on the emancipation process and that may have stimulating effects on the emancipation process. With these results in combination with the results of the study discussed in this paper, the mobility policy can be reviewed to reach the aims set by the emancipation policy.

ACKNOWLEDGMENTS

The authors thank Ingeborg Oostlander of the Ministry of Transport and Public Works for initiating this study. The paper is sponsored by the Dutch Ministry of Transport and Public Works and Water Management.

REFERENCES

1. *Second Transport Structure Plan: Part D: Government Decision: Transport in a Sustainable Society.* Second Chamber of the States-General, Session 1989–1990, 20922, No. 16. SDU, The Hague, The Netherlands, 1990.
2. Batenburg, R. S., and W. P. Knulst. *Sociaal-culturele beweegredenen: onderzoek naar de invloed van veranderende leefpatronen op de mobiliteitsgroei sinds de jaren zeventig.* Sociaal en Cultureel Planbureau, Rijswijk, The Netherlands, 1993.
3. Pol, M., and D. C. Zoutendijk. *Emancipatie en Mobiliteit.* TT94-45. Traffic Test bv. Veenendaal, The Netherlands, 1994.
4. Van Schendelen, M. C. *Op weg naar een emancipatoir mobiliteitsbeleid: Een verkennende studie naar de invloed van emancipatie op de mobiliteit.* Emancipatieraad, The Hague, The Netherlands, 1992.
5. Tweede Kamer der Staten-Generaal. *Met het oog op 1995: Beleidsprogramma Emancipatie, Vergaderjaar 1992–1993.* 22913. Nos. 1 en 2. SDU, The Hague, The Netherlands, 1992.
6. Fox, M. B. Working Women and Travel. *Journal of the American Planning Association,* Vol. 49, No. 2, 1983, pp. 156–170.
7. Drooglever Fortuijn, J., S. Hietbrink, L. Karsten, and M. Rijkes. *Van hot naar her; Het combineren van taken in tijd en ruimte door tweeverdieners, alleenstaande ouders en alleenwonenden in Haarlem, Enkhuizen en Stede Broec.* Studierapport 38. Provinciale Planologische Dienst van Noord-Holland, Haarlem, The Netherlands, 1987.
8. Hendriks, G. *Ruim baan voor herintreedsters; (Ruimtelijke) Randvoorwaarden voor toetreding tot de arbeidsmarkt.* Vakgroeppublicatie 27. Katholieke Universiteit, Nijmegen, The Netherlands, 1989.
9. Hanson, S., and P. Hanson. The Impact of Married Women's Employment on Household Travel Patterns: A Swedish Example. *Transportation,* No. 10, 1981, pp. 165–183.
10. Hooghiemstra, B. T. J., and M. Niphuis-Nell. *Sociale atlas van de vrouw, Deel 2: Arbeid, inkomen en faciliteiten om werken en de zorg voor kinderen te combineren.* Sociaal en Cultureel Planbureau, Rijswijk, The Netherlands, 1993.
11. Raux, C., and S. Rosenbloom. *Employment, Childcare and Travel Behavior: France, The Netherlands, the United States.* In *Behavioral Research for Transport Policy,* VNU Science Press, Utrecht, The Netherlands, 1986.

12. Pickup, L. Combining Family and Work Roles: the Importance of Transport in Women's Job Chances. In *Interpreting and Valuing Transport's Role in Social Wellbeing.* An International Conference, Noordwijk, The Netherlands, April 1983.
13. De Vries, A. Milieu en emancipatie-strijd of coöperatie? *Socialisme en Democratie,* No. 3, Maart, 1994, pp. 155–161.
14. Rosenbloom, S. The Impact of Growing Children on Their Parent's Travel Behavior: a Comparative Analysis. In *Transportation Research Record 1135,* TRB, National Research Council, Washington, D.C., 1987.
15. Rosenbloom, S. Trip Chaining Behaviour: A Comparative and Cross-cultural Analysis of the Travel Patterns of Working Mothers. In *Gender, Transport and Employment* (M. Grieco, L. Pickup, and R. Whipp, eds.), 1989.
16. Drooglever Fortuijn, J. Een druk bestaan; Tijdsbesteding en ruimtegebrek van tweeverdieners met kinderen. *Academisch Proefschrift.* University Press, Amsterdam, 1993.
17. Niemeier, D. A., and J. Morita. Duration of Trip-Making Activities By Men and Women: a Survival Analysis. Presented at the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., 1994.
18. De mobiliteit van de Nederlandse bevolking 1992. CBS, Voorburg/Heerlen, The Netherlands, 1993.
19. Gordon, P., A. Kumar, and W. Richardson. Gender Differences in Metropolitan Travel Behaviour. *Regional Studies,* Vol. 23, No. 6, 1989, pp. 499–510.
20. Hanson, S., and I. Johnston. Gender Differences in Work-Trip Length: Explanations and Implications. *Urban Geography,* Vol. 6, No. 3, 1985, pp. 193–219.
21. Cassell, B., S. D. Lawson, A. Snelson, and H. T. Morris. *Women and the Importance of the Car.* PTRC, 1993.
22. De Boer, A. H., J. C. Hessing-Wagner, M. Mootz, and I. S. Schoemakers-Salkinoja. *Informele zorg—een verkenning van huidige en toekomstige ontwikkelingen.* Social en Cultureel Planbureau, Rijswijk, The Netherlands, 1994.
23. Hamilton, K., and L. Jenkins. *Women and Transport; Implications for Policymakers.* PTRC, 1993.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Effects of Increased Highway Capacity: Results of Household Travel Behavior Survey

RICHARD G. DOWLING AND STEVEN B. COLMAN

Travel behavior is likely to change when road congestion and travel times are improved as a result of new highway capacity. The behavioral change is complex and may manifest itself over both the short and long run. Short-term impacts may include changes in route choice, time of day that trips are made, mode choice, trip frequency, trip chaining, and destination choice. Longer-term impacts may include changes in automobile ownership, residential location, choice of workplace location, and land development patterns. These changes occur against a background of economic, demographic, and pricing changes affecting the population as a whole. A fresh approach is taken to illuminate the question of whether highway improvements induce new travel. The research has been framed in terms of relating the time "released" by a highway improvement to how households would use this time. The question then becomes, Do travelers use the time saved to make more (or longer) trips, or do they use it for other activities? To make the responses more realistic, respondents were asked to relate hypothetical changes in congestion levels to their previous day's travel and activity patterns. The results of a stated preference/activity survey of nearly 700 urban Californians indicate that congestion-relieving projects are likely to induce a small (3 to 5 percent) but not trivial increase in trip generation. This effect could be accounted for by modifications in the traditional "four-step" travel forecasting models, which gives transportation and air quality analysts a better sense of how to assess the potential induced travel impacts of new highway capacity.

Few current transportation issues engender more controversy than the effects of new highway capacity on traffic and travel demand. The purpose of adding highway capacity is to reduce traffic congestion and improve automobile travel times and, in some cases, air quality. These changes, in turn, affect travel behavior by affecting peoples' choice of modes of travel, their choice of destination, and their choice of travel route. Less well known is how travel time changes caused by capacity increases may affect total travel demand, especially trip generation (i.e., the number of vehicle trips made per person or per household). Estimating the magnitude of this effect on trip generation is particularly uncertain. A primary purpose of this project was to examine the effects of new capacity on trip generation, because in most North American travel forecasting models, trip generation is not sensitive to transportation supply variables.

IMPORTANCE TO CLEAN AIR AND TRANSPORTATION

Federal, state and local governments spend billion dollars a year on new road improvements to reduce congestion, improve safety, and provide for economic development. There is popular and some professional opinion that new capacity in urban areas is eventually swamped by new demand so that in the end motorists are no better off than they were before the improvement was made (*1,2*). Disagreements arise about whether this effect exists and, if it does, what its magnitude is. The issue has moved to center stage because the 1990 Clean Air Act Amendments prohibit recipients of federal transportation funds from constructing projects that worsen air quality in nonattainment areas.

A road improvement may improve air quality depending on whether a trip-inducing effect occurs. New road capacity, to the extent that it reduces speed variations (stop-and-go driving) and allows vehicles to travel a steady 30 to 45 mph (48 to 72 kph), improves air quality. This claim has been challenged by others, who maintain that any air quality benefit of new road capacity in the short term will be offset in the longer term by increased travel demand that will nullify any improvement in total emissions. Of course, the trip induction effects of new highway capacity do not have to be 0 for there to be a net air quality benefit, but they must be smaller than the increase in emissions per vehicle.

STUDY PURPOSE AND RESEARCH APPROACH

The purposes of this study were to answer two fundamental questions: Do capacity increases increase trip making? If so, what is the magnitude of this increase, if it exists? The overall research objectives were accomplished through a variety of means; this paper reports on the results of a household survey of traveler behavior conducted as part of the study. Past attempts to assess the travel impacts of new highway capacity have mostly relied on before-and-after traffic volume comparisons. In some cases traffic counts have been supplemented with roadside interview or home interview surveys. A few investigators have attempted to fit regression models for predicting regional vehicle kilometers of travel (VKT) increases that result from regional increases in highway capacity. However, this approach has generally not been fruitful because a variety of extraneous factors can affect the results, including the availability of alternative modes and routes in each corridor; the condition of the local economy (growing or stagnant); zoning; and natural constraints to development. These factors not only affect the conclusions but also limit the validity of extending these results to other situations and locations. Shortcomings of the case study approach are documented in the literature (*3,4*). A brief summary of the reasons for proposing an alternative approach follows.

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

Control of Exogenous Variables (Economic Conditions)

Transportation changes take place in a highly dynamic environment: household income, population, employment, fuel and parking prices, and other variables cannot be directly controlled for. A time series approach may not control for the distributional shifts in land use activities that transportation investments may induce if the area of analysis is limited. This creates a considerable problem in distinguishing between a shift along the demand curve (because of the reduced price of travel caused by added capacity), and a shift in the demand curve itself (see Figure 1). Demand curves may shift as a result of changes in income, tastes, and demographic factors. Point 1 represents an initial condition with a four-lane freeway; Point 2 is the result of a capacity increase (travel time reduction) and the associated movement along today's demand curve. Point 3 is purely the result of a demand curve shift, possibly caused by such factors as increased population or income but also possibly caused by reduced transit service, higher fares, or changes in taste. Point 4 is the final equilibrium—a combined result of capacity and demand increase.

Completeness of Data Sets

The data requirements of a case study approach include (as a minimum) annual traffic counts on the new facility and all paralleling routes along with good records of land use changes in the corridor. Local agencies often lack consistent annual count programs with counters at the correct locations to assess changes in corridor demand because of capacity changes. Even if all of the count data were available perfectly, the appropriate temporal resolution needed to assess the impacts of new capacity may be missing. Ideally, counts would be available at 15-min intervals to assess the impacts of temporal shifting in travel, and especially the "peak within the peak." Information needs to be available on all paralleling transit services; even then, one would not know what the changes in destination choices were (were people driving further because of the new capacity to reach a "better" destination?); or the shifts in land uses that took place over time.

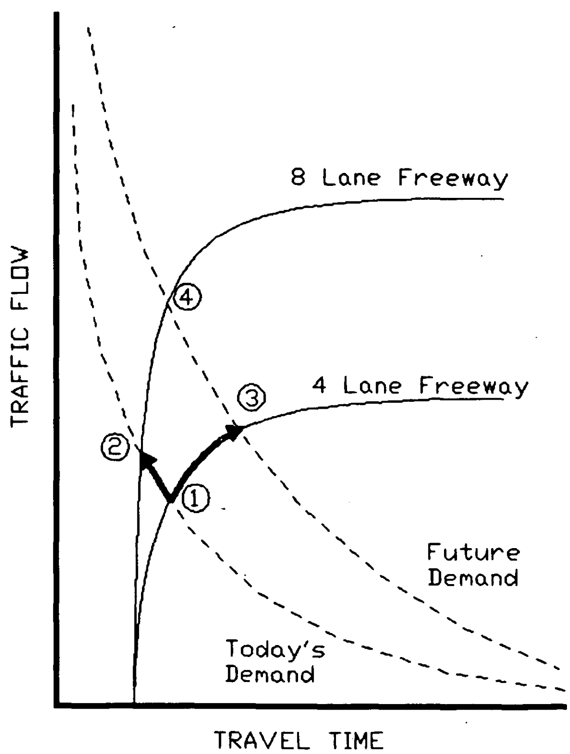


FIGURE 1 Demand versus capacity changes.

Differences in and Comparability of Data Collection Years

Traffic counts, income, and other demographic information typically are not available annually. Most agencies make projections at 5-year intervals, and generally traffic counts are made only at 2- or 3-year intervals (sometimes less often than that). This requires interpolating between demographic data, traffic count, and traffic forecast years. Increased real income and family size (lifecycle issues) typically result in higher levels of automobile ownership and a desire for more residential space. Detailed geographic information at the corridor level is usually available only from the U.S. Census, which is conducted too infrequently (every 10 years) to be useful.

Institutional Bias

Forecasts may contain an institutional bias, perhaps unconsciously. An agency may make reasonable assumptions within a "gray area" of discretion that favors the action that the constructing agency wishes to take. Biases can vary with time, place, and the individuals involved, but can all lead to forecasting errors. An agency could use optimistic or pessimistic views of the economy, of population growth, and so forth.

All of these considerations pointed toward the need for an approach that

- Considers trips in the context of the overall activity patterns of travelers,
- Considers a wider range of alternatives than would be possible to test with the case study approach, and
- Avoids the shortcomings of incomplete data sets, control of exogenous variables, and other limitations noted earlier.

RESULTS OF PREVIOUS RESEARCH

Increased highway capacity may affect travel in a number of ways. In urban areas, new capacity typically reduces congestion, resulting in shorter travel times during some or all of the day, and a less stressful driving experience. In rural areas and small cities, where congestion is minimal, new capacity may or may not change travel times. The literature (*5–8*) documents a strong relationship between reduced travel times and the following short-term effects:

- The choice of the route taken. This effect has been found to be consistently important in the literature. A major assumption underlying the conventional four-step travel forecasting process is that people seek routes that minimize travel time and cost.
- The scheduling of the trip (time of day the trip starts/ends). This effect also has been found to be consistently important in the literature; new highway capacity often has been found to cause

shifts from off-peak or "shoulder" transitional times to the "core" peak periods of travel. This effect was found in examining traffic count data before and after widening of CA-78 in San Diego, the Amsterdam M10 Orbital Motorway (*7*), and other locations.

- The choice of the travel mode used (e.g., carpool, transit, drive alone). This effect has been shown to have a much weaker impact than route and scheduling choice but is still important. The effect is probably more important in the longer term, as changes in automobile ownership and land use take place. Studies of the substantial and sudden capacity reductions caused by the 1989 Loma Prieta earthquake indicate substantial shifts to transit modes (*9*), with about a 10 to 15 percent reduction in the number of total daily person-trips (Markowitz, unpublished data). This reduction is modest compared with the large increase in travel time (often 50 to 100 percent) occasioned by many transbay travelers during the approximately 1-month period when the San Francisco–Oakland Bay Bridge was closed because of the Loma Prieta quake.
- The frequency the trip is made. The literature has been inconclusive on this topic, with some studies indicating significant impacts and others indicating little or no measurable impact. This impact was one of the primary concerns of this project.
- The linking of trips with several destinations together (sometimes known as "trip chaining" or "trip tours"). This appears to be an important impact but has proven difficult to measure and is generally outside the scope of this paper.
- A change in the choice of the destination of a trip; likewise, this impact has proven difficult to measure.

A study of disaggregate household vehicle trip generation rates as a function of proximity to freeway ramps (*10*), using distance as a proxy for accessibility to destinations in 24 urban California counties, was recently made of 6,200 randomly selected households. The study found no significant correlation between the two variables after controlling for other factors. However, this approach had limitations in that distance to the freeway could be measured only as distance to the census tract centroid because survey address records were destroyed (*11*). Furthermore, the results are complicated by the fact that the frequently found convergence of freeways near the core of central cities meant that lower-income residents were often the most proximate to one or more freeway interchanges.

Areawide models (derived by correlating VKT growth to highway system increases) seem more desirable than facility-specific studies because they eliminate the route choice effects by considering entire regions (*11,12*). They are also able to take into account long-term land use effects by extending the analysis over several decades. However, they focus on VKT instead of person-hours of travel and consequently confuse mode shift effects with true induced demand. These studies have been inconclusive about the elasticity of demand (VKT) with respect to new lane-miles of capacity; although all the reported results have been inelastic, they range from a very inelastic 0.1 to a much more elastic 0.8 (*8*).

Even the areawide studies suffer from several critical deficiencies: first, they use a single relatively simple measure of capacity increase (such as lane-kilometers or lane-miles) that is insensitive to the potentially significant different demand effects that would occur if the same investment were made in the center of the region versus the fringes. There are definitional problems in computing the denominator of the elasticity equation; the percentage increase in capacity must be estimated, meaning that a "base" capacity must be measured. Should the base capacity be measured at the corridor, county, primary metropolitan statistical area or consolidated metropolitan statistical area (CMSA) level? Economic theory, as well as experience with transportation and land use forecasting models, indicate that transportation supply cannot be treated as a homogeneous product (*13*).

Common sense suggests that new highway capacity has different impacts in an area that is already "built out" as opposed to one where much undeveloped land exists simultaneously with strong pressures for development. The costs of parcel assembly, structure demolition, and so forth, are simply too high. In most cases the structure built on a parcel of land in the United States is the only one that has ever occupied that piece of property (*14*).

Second, most areawide studies assume a constant elasticity of demand, probably because of the lack of enough data points. Intuition and economic theory suggest that elasticity is not necessarily constant but instead depends on the amount of current congestion and capacity of the system, the time frame involved (short-versus long-term), the trip purposes of road users, and possibly other factors. This issue requires further research.

Because of the problems associated with the case study before-and-after approach (facility specific or areawide), it was decided to use a survey of household travel behavior to isolate the various effects of new highway capacity and identify those effects not currently treated by conventional travel forecasting models. The travel survey and its results are described below.

TRAVEL BEHAVIOR SURVEY

A travel behavior survey was developed and administered to fill in the missing information from the case studies on the relative importance of the different effects of new highway capacity on travel behavior. Each potential effect (mode, time, destination, trip generation) would be identified and quantified for the purpose of determining its relative importance in estimating the total demand effects of new highway capacity.

Selection of Survey Approach

There are two general approaches to conducting behavioral surveys: stated preference (SP) and revealed preference (RP). Other references provide a comparison of these two methods (*15*); briefly, a stated preference survey poses various situations to the interview subject and asks How would you respond to the given situation given certain constraints? A revealed preference survey relies on measurement of actual responses to alternatives existing in the field. RP surveys can test only for the conditions that exist at the time of measurement, but an SP survey can explore behavioral changes because of a much wider range of options. RP surveys traditionally have been used to calibrate travel forecasting models. RP surveys provide information on the actual choices made by individuals in the face of two or more options. RP surveys have several limitations when applied to the problem of estimating the behavioral effects of new highway facilities. Critical shortcomings are the difficulty in avoiding bias in the selection of the survey sample and accounting for persons moving into and out of the presumed "impact" area of the new facility, and controlling for changes in background variables, such as economic and demographic changes.

The major difficulty in applying an SP survey to the research problem is that traditional SP surveys require that the respondent be offered a choice between trip or transportation system attributes that force a realistic trade-off by the user. In a classic SP survey, the respondent is offered a higher fare/shorter travel time option, and a lower fare/longer travel time option. With increased highway capacity/reduced congestion, such a trade-off is not possible because presumably everyone would prefer a shorter travel time. To make meaningful tradeoffs between alternatives, respondents were asked to describe all of their previous day's activities and then contemplate how they would alter them if more (or less) time were available on that day to perform those activities. Perhaps more precisely, it is how people would use "released" or "freed-up" time, if congestion-relief projects made such time available.

The survey also embodied concepts from the developing field of activity analysis (*16*). Within the survey instrument here, people were asked about all of the previous day's activities and then asked to respond to changes in travel and activity patterns given changes in travel time for trips made on the reference day. Although the 24 hr available each day is fixed for every individual, the allocation of time to each activity is not. The time and money allocated to travel is further subdivided among mandatory activities such as going to work, school, and so forth, and discretionary activities such as going to a movie. These various daily activities can be thought of as "goods" in the economic sense that people "purchase" by spending "time" and money on the activity. A 1987 survey (*17*) found that the average California adult spends 1.8 hr a day traveling, more than 10 percent of his or her waking hours.

Each survey respondent was told the following:

> We are trying to find out how traffic congestion affects what people do. I am going to describe what might happen if traffic congestion got better or worse, and ask you how you might change your activities or travel as a result. Please take some time to think carefully about what you might do.

The respondent was then read back all of the trips he or she made the previous day, and asked,

> Consider what you told me about what you did yesterday. For each trip I am going to ask you what you would have done if it had taken less time to make the trip. Consider your first trip yesterday. You started at . . . [time] and went to . . . [destination] by . . . [mode]. This trip took . . . [duration previously stated by respondent]. Now suppose that this trip took [randomized duration] less time to make. Please select one or more of these statements that best describe what you would have done.

Respondents were not asked about trips that were less than 10 min in duration, because the minimum travel time savings "offered" was 5 min, and it was thought that for trips of less than 10 min, a time savings of 50 percent or more would be unrealistic and unlikely to be achieved by any plausible capacity-increasing project and also because of the desire to offer travel time savings in increments of 5 minutes. In fact, one of the survey problems was that the total travel time change was independent of the individual's reported trips. Also the total released time during the day was not keyed to a specific hour, which some respondents indicated would condition their response of how the time would be used.

Survey Methodology

Adults over the age of 16 in the San Francisco and San Diego metropolitan areas were randomly selected; these two areas contain about 8.7 million people. Respondents were interviewed about their existing travel behavior, activity patterns, and hypothetical behavior under changes in travel time. "Number plus one" dialing was used to reach unlisted numbers. The Los Angeles area was excluded because the Northridge earthquake occurred shortly before the survey commenced and had dramatically affected travel patterns there. The survey was administered using computer-assisted telephone interviewing (CATI) because of the complex branching required in the survey. Interviews were conducted on Tuesday through Friday evenings and Saturday midday, with survey questions asked about the prior day's travel. Randomization techniques were used to ensure that the person who answered the phone was not necessarily the person interviewed.

After all trips were enumerated, the CATI program selected each trip made that was at least 10 min long. For trips between 10 and 15 min, a 5-minute reduction in travel was offered. For trips longer than 15 min, a randomized travel time savings of between 1 and 50 percent was offered; the randomized savings was a minimum of 5 min if the survey number was odd and 10 min if the survey number was even.

Survey respondents were given the following options: doing nothing differently; starting at the same time and arriving earlier; starting later and arriving at the same time; changing mode; changing trip destination; making an extra stop along the way; and "other." Only one additional "extra stop" was allowed for in the questionnaire, although in reality it is possible that some individuals might add two (or more) trips to their tour. The possibility of entirely new trips was allowed for at the end of this process by asking, Would you have left home again before the end of your day if you had [randomized time] minutes extra time? If the answer was yes, the respondents were asked where they would have gone, how much time they would have spent there, and for what purpose.

Survey Results

A total of 676 individuals over the age of 16 were interviewed in 676 households. They collectively made a total of 2,182 trips the previous day. The respondent demographics (age, income, educational achievement, and automobile ownership) were compared with those from the 1990 Census. The respondent pool was close to the state average, except that poor households (those earning under $15,000 per year) were somewhat underrepresented. About 90 percent of the respondents were willing to report their household income. Of those answering the question, 9.5 percent reported household incomes under $15,000 per year. The 1990 Census found the same group constituted 15.1 percent of the households in the San Francisco Bay Area (CMSA). Some of the difference can be accounted for by inflation between 1989 (the reference year for the census) and the year of the survey (1994).

Very-low-income groups tend to be underrepresented in most telephone surveys, but the importance of these households is mitigated by the fact that they produce a small percentage of VKT. The National Personal Transportation Survey (*18*) found that households with incomes under $10,000 generate VKT/household that is only 40 percent of the average rate for all households (using automobile driver miles as the measure). The 1990 Census found that these households represent about 15.5 percent of all households in the United States; therefore, it appears that they are responsible for somewhat over 6 percent of VKT.

The key results of the survey (Tables 1 and 2) were as follows:

TABLE 1 Responses of Travelers to Travel Time Savings for Each Trip

	Travel Time Savings due to Congestion Relief (minutes)				
Response	5	10	15	20+	All
No Change	46.5%	49.6%	35.1%	38.1%	46.5%
Arrive Earlier	34.9%	33.9%	40.5%	31.0%	34.6%
Leave Later	12.9%	12.5%	16.2%	23.8%	13.5%
Change Mode	0.4%	0.4%	2.7%	2.4%	0.6%
Change Destination	0.9%				0.5%
Make Extra Stop	2.9%	2.8%	5.4%	4.8%	3.1%
Other	1.5%	0.8%			1.1%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

- Over 35 percent of the trips made would be unaffected when the trip travel time increased or decreased by 15 min or less considering all trip purposes.
- Another 20 to 40 percent of trips made would change only to the extent that the respondent would arrive earlier or later at a destination and make no change to the departure time to compensate for the effect of the travel time change.
- About 10 percent to 15 percent of the trips would be rescheduled to compensate for or take advantage of the travel time change.
- A time savings of 5 min would generate extra stops for about 3 percent of the trips. This percentage increased to 5 percent when a 15-min time savings was offered. The average across all time savings offered was 3 percent.

The overall result is that 90 percent to 95 percent of the trips would be unchanged or would have schedule changes in response to travel time increases and reductions of 15 min or less. As expected, the greater the magnitude of the travel time change, the greater the traveler response. Interestingly, the results are not symmetrical: respondents tended to react slightly more strongly to increases than to decreases in travel time (see Figure 2). When faced with a travel time increase, respondents would try to adapt by changing mode, destination, and route for a higher percentage of the trips than if they were offered an equal amount of time decrease. Given the nature of the two metropolitan areas in which the survey was conducted, it is likely that more respondents had recent experience adjusting to travel time increases than decreases. Asymmetric behavior is probably not surprising; some gaming simulations have shown that even given the same actuarial odds (expected value), people are much more concerned with a possible loss of wealth than they are with a possible gain.

TABLE 2 Responses of Travelers to Travel Time Increases for Each Trip

	Travel Time Increase due to Congestion (minutes)				
Response	5	10	15	20+	All
No Change	53.5%	41.3%	38.6%	24.4%	45.7%
Arrive Later	22.1%	31.0%	38.6%	36.6%	27.8%
Leave Earlier	17.3%	17.6%	9.1%	24.4%	17.4%
Change Mode	1.2%	1.5%	4.5%	2.4%	1.6%
Change Destination	1.0%	0.4%	2.3%		0.7%
Make Extra Stop	0.2%	1.3%			0.7%
Other	4.6%	6.9%	6.8%	12.2%	6.1%
Total	100.0%	100.0%	100.0%	100.0%	100.0%

The respondents indicated that only approximately 1.6 percent of their trips would be susceptible to a modal change given increased travel time for a specific trip. Of these hypothetical "mode switchers," most (38 percent and 35 percent, respectively) said they would switch to driving alone or public transit. It was implicit in the survey that the travel time by alternative modes was not changed. Greater time increases and decreases had a greater effect on traveler responses than smaller amounts of time changes. However, given that only 13 percent of survey trips were greater than 30 min in length, it was not realistic to ask the majority of the respondents about time savings of greater than 15 min.

CONCLUSIONS AND RECOMMENDATIONS

Most previous investigations of the effects of new highway capacity have been facility-specific "before-and-after" studies. At first, this approach seems appealing and logical, but on reflection, it becomes clear that it is nearly impossible to isolate the effects of new highway capacity on induced trip making. There are too many extraneous factors that can affect the results, including the availability of alternative modes and routes in each corridor; the condition of the local economy; zoning; and natural constraints to development. These factors not only affect the conclusions but also limit the validity of extending these results to other situations and locations. These factors may have been responsible for the conflicting conclusions that other researchers frequently arrived at in the past.

The results of this survey must be qualified by its relatively small size (under 700 households) and limited geographic scope. However, the following are some of the indications from this survey:

- Current travel forecasting practice probably results in an underprediction of 3 to 5 percent in the number of trips that may be induced by major new highway capacity projects. Where a project is expected to yield travel time savings of more than 5 min for a large number of trips, adjusting travel demand upward to reflect induced travel is probably warranted.
- A key impact of new highway capacity is temporal shifts in demand (trips formerly made in the off-peak moving to the peak periods). From the highway user's perspective, this is not necessarily bad because it means that he or she can make a trip in response to personal needs rather than to traffic conditions. On the other hand, it will affect the congestion, speeds, and emission estimates produced by travel models. There is a strong need to develop better models to predict peak spreading/time of day of travel.

In the longer term, new highway capacity may influence decisions about automobile ownership, residential location, employment location, and the locations of expansion areas for businesses and government. These effects are important but are beyond the scope of this paper. Several of these effects cannot be addressed with a household travel behavior survey. However, some of these impacts are already accounted for in current transportation/land use forecasting practices in California's largest metropolitan areas, using models such as DRAM/EMPAL and POLIS.

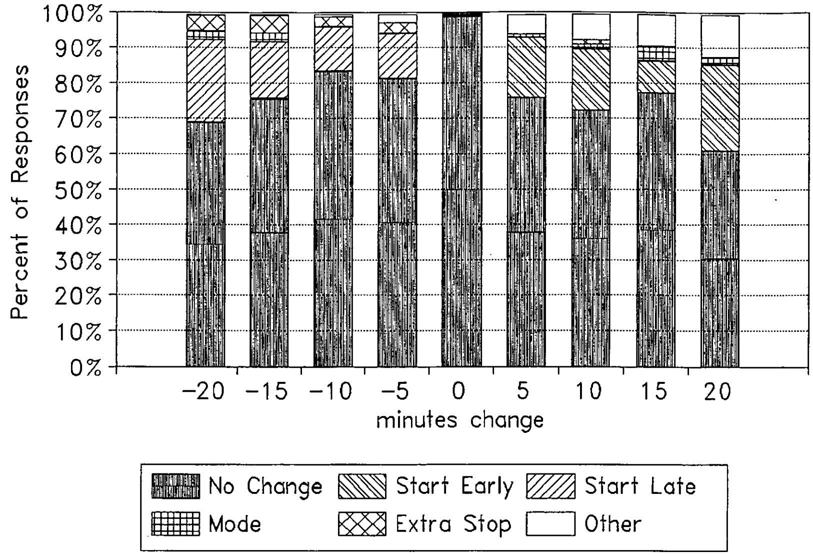


FIGURE 2 Response of travelers to hypothetical trip time changes.

Key Conclusions

Highway capacity changes influence travel behavior principally by affecting travel time and cost. The principal conclusions from the survey are as follows:

• The sample indicated definite preferences about how travelers would respond to changes in travel time. Their response preferences are in the following order:

- Change route (find a faster route if the current one becomes congested);
- Change schedule (find another time of day when congestion is lower);
- Consolidate trips (reduce number of daily trips by accomplishing more activities with a given trip);
- Change mode (switch to more convenient mode); and
- Change destination (find another location with similar services).

• Whether a person prefers to change mode over destination (or vice versa) may depend on the trip purpose, for example, a destination change is probably preferred over a mode change for most shopping trips.

• The order of preference responses appears to be similar for travel time increases and decreases, although the magnitude is different. Whether faced with travel time increase or decrease, both changes would result in the respondent preferring a different route or rescheduling the trip, rather than changing the trip mode or destination.

• Survey respondents indicated a high degree of resistance to change in their travel behavior when offered travel time savings of between 5 and 15 min per trip. A 5-min travel time savings (on average) resulted in a 3 percent increase in daily trips made per person and a 15-min time savings resulted in a 5 percent increase in trips per person per day.

Because most trips in metropolitan areas are less than 15 min long and realistic time savings on such short trips would rarely exceed 5 min, it is unlikely that adding new lanes to an existing highway would significantly reduce travel times for the majority of trips, although this general observation may not apply to new highways or to home–work (commute) trips. Commute-related trips are longer at an average of between 20 and 30 min and are more likely to encounter peak-period congestion. The commute trip also drives many other decisions, such as vehicle holdings and household location, and those considerations have a substantial influence on generation of short trips. Thus, there could be some important secondary impacts that are not accounted for here.

Recommendations for Future Research and Survey Improvement

There were questions that could not be answered in this study. They include assessing whether the results are transferable to other areas; how congestion affects interactions between household members; and how qualitative factors (such as stress) may influence travel behavior when congestion is reduced. It seems logical to presume that a 30-min drive in stop-and-go traffic would be perceived differently from a 30-min drive in free-flowing traffic, but the survey instrument was not able to distinguish between the two. A small sample of commuters in Orange County, California (*19*), found that most, but not all, drivers perceived commuting in congested traffic as more stressful than commuting in uncongested traffic. To the extent that this is true, it suggests that the results of the travel survey conducted here could underestimate the true effects on tripmaking of reduced congestion.

It is recommended that the following steps be taken to improve the understanding of the effects of increased highway capacity on travel behavior and to improve the ability to forecast these effects at the regional level. Repeating the behavioral survey in other metropolitan, and possibly rural, areas to determine whether the survey results can be reliably extrapolated to all travelers would be desirable. A larger survey sample would also yield more information on the effect of new highway capacity on various trip types and purposes.

The wording of survey questions and presentation of alternatives are critical in most SP surveys and are among the known weaknesses of the method. Some respondents were confused about whether a visit to a different location meant a different location for the same purpose or a different location for a different or additional purpose. For some respondents who made fairly short trips, the total travel time savings presented was near or greater than the amount of time the respondent had reported in travel. Some respondents who realized this were confused.

This survey did not allow for the possibility that people could save a trip time reduction over a week, and "spend" it as a block. The survey approach was thought to be appropriate since, unlike money, time is not easily "banked." However, the authors recognize that the greater an individual's flexibility in allocating time, the more likely that travel time savings should be investigated using a week as the reference period (rather than 24 hr). Nonworkers or those working part time would appear to have the greatest flexibility in this regard (the increasing use of 4-day work weeks may also be important).

It would be useful to use other research approaches to corroborate the results of this survey. One is activity gaming and simulation, which allows researchers to better understand the intrahousehold allocation of travel and other activities. This study made only a rudimentary attempt to consider how one household member's travel time changes might affect the travel and activity patterns of other members of the household. Another approach would be to collect detailed information on the before-and-after effects of those living in a corridor where travel times are improved. Recently developed automatic vehicle location technology, using cellular phone technology, would allow detailed multiday travel diaries to be analyzed without the tedium and error associated with the traditional manually kept diaries.

ACKNOWLEDGMENTS

This paper was prepared on the basis of work prepared under the sponsorship of the California Air Resources Board. The authors thank Fereidun Fezollahi, Anne Geraghty, and Doug Thompson of the California Air Resources Board for their guidance. Ryuichi Kitamura developed the concept of the stated preference and prior day's activity survey along with David Reinke. Peter R. Stopher made many helpful comments on the survey, as well as the overall study process. The authors thank them all for their thoughtful assistance.

REFERENCES

1. Bass, T. Road to Ruin. *Discover Magazine,* May 1992. pp. 56–60.
2. Downs, A. The Law of Peak-Hour Expressway Congestion. *Traffic Quarterly,* Vol. 16, No. 3, July 1962.
3. Evaluation of the Accuracy of Past Urban Transportation Forecasts. *ITE Journal,* Feb. 1980, pp. 24–34. Institute of Transportation Engineers, Committee 6F-13, Washington, D.C.
4. Stopher, P. R. Travel and Locational Impacts of Added Transportation Capacity: Experimental Designs. Resource Paper Presented at Conference on Travel Effects of Added Transportation Capacity, Bethesda, Md., Dec. 1991; FHWA, U.S. Department of Transportation, pp. 47–54.
5. Jorgensen, R. E. Influence of Expressways in Diverting Traffic from Alternate Routes and Generating New Traffic. *HRB Proc.* Vol. 27, 1947.
6. Pells, S. R. *User Response to New Road Capacity: A Review of Published Evidence* Institute for Transport Studies Working Paper 283. University of Leeds, England, 1989.
7. *Effects of the Opening of the Amsterdam Orbital Motorway.* Final Report, Phase 1. Rijkswaterstaat, Transportation and Traffic Research Division, Rotterdam, The Netherlands, 1992.
8. Dobbins, A., M. Hansen, D. Gillen, Y. Huang, and M. Puvathingal. *Air Quality Impacts of Urban Highway Capacity Expansion: Traffic Generation and Land Use Impacts.* Draft Report UCB-ITS-RR-93. Institute of Transportation Studies, University of California, Berkeley, April 1993.
9. Homburger, W. S. The Loma Prieta Earthquake: What Happened to the Trip Makers? A Preliminary Report. In *Compendium of Technical Papers,* Institute of Transportation Engineers (District 6) 1990 Annual Meeting, Boise, Idaho., pp. 1–6.
10. Rothblatt, D. R., S. B. Colman, and E. Bossard. *Effects of Residential Density on Transit Usage and Residential Trip Generation.* Prepared for the California Department of Transportation Research and Training Program by Institute for Metropolitan Studies, San Jose State University, Calif., 1994.
11. Garrison, W. L., and R. D. Worrall. *Monitoring Urban Travel.* Final report NCHRP Project 2-8: Estimation and Evaluation of Diverted and Generated (Induced) Traffic, HRB, National Research Council, Washington, D.C., 1966.
12. Ruiter, E. R., W. R. Loudon, et al. NCHRP Project 8-19: The Relationship of Changes in Urban Highway Supply to Vehicle Miles of Travel. In *NCHRP Research Results Digest 127.* Washington, D.C., TRB, National Research Council, 1980.
13. *Assessing the Future: A Sensitivity Analysis of Highway and Road Improvements on Growth in the San Francisco Bay Area.* Working paper 91-4. Association of Bay Area Governments, Oakland, Calif., 1991.
14. Meyer, J. R., and J. A. Gomez-Ibañez. *Autos Transit and Cities.* A 20th Century Fund Report. Harvard University Press, Cambridge, Mass., 1981.
15. Ben-Akiva, M., T. Morikawa, and F. Shiroshi. Analysis of the Reliability of Stated Preference Data in Estimating Mode Choice Models. In *Transport Policy, Management and Technology Towards 2001,* Western Periodicals, Ventura, Calif., 1989.
16. Kitamura, R. An Evaluation of Activity-Based Travel Analysis. *Transportation,* Vol. 15, 1988, pp. 9–34.
17. Wiley, J. A. *Activity Patterns of California Residents.* Contract A6-177-33. California Air Resources Board Research Division, Sacramento, 1991.
18. *1990 NPTS Databook—Nationwide Personal Transportation Survey, Vol. I.* Report FHWA-PL-94-010A. FHWA, U.S. Department of Transportation, 1993.
19. Novaco, R. W., D. Stokols, and L. Milanesi. *Objective and Subjective Dimensions of Travel Impedance as Determinants of Commuting Stress.* Reprint 30, University of California Transportation Center, Berkeley, 1990.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Travel-Time Uncertainty, Departure Time Choice, and the Cost of Morning Commutes

ROBERT B. NOLAND AND KENNETH A. SMALL

Existing models of the commuting time-of-day choice were used to analyze the effect of uncertain travel times. Travel time included a time-varying congestion component and a random element specified by a probability distribution. The results from the uniform and exponential probability distributions were compared and the optimal "head-start" time that the commuter chooses to account for travel time variability, that is, a safety margin that determines the probability of arriving late for work, was derived. The model includes a one-time lateness penalty for arriving late as well as the per-minute penalties for early and late arrival that are included by other investigators. It also generalizes earlier work by accounting for the time variation in the predictable component of congestion, which interacts with uncertainty in interesting ways. A brief numerical analysis of the model reveals that uncertainty can account for a large proportion of the costs of the morning commute.

The choice of home departure time for commuters is an important element in determining how congestion levels will vary during morning peak travel. This choice has been related empirically to the cost of early or late arrival relative to some preferred work arrival time (*1,2*). The planning of on-time arrivals is, however, complicated by the presence of uncertainty in actual travel times.

This paper describes a model in which commuters simultaneously trade off costs of inconvenient schedules, lateness penalties, and the desire to minimize time spent in congested traffic. Like Gaver (*3*) and Polak (*4*), the authors assumed that commuters face a probabilistic distribution of travel times and choose departure time to minimize an expected cost function. In contrast to these authors, the cost function includes a discrete lateness penalty as well as per-minute penalties for both early and late arrival; it also accounts for variation over time in the predictable component of congestion. Furthermore, the optimized expected cost function (i.e., the costs resulting after an optimal departure time is chosen) is derived analytically. This is done for both a uniform and an exponential distribution for uncertain travel time.

The results show how changes in the uncertainty of travel time affect both the departure time decision and the resulting expected costs. For example, as uncertainty increases, commuters shift their departure schedules to earlier hours to compensate for the increased probability of late arrival; in some cases they overcompensate in the sense that the probability of late arrival decreases as uncertainty increases. As for the resulting expected costs, the functional relationship that is derived by relating costs to the underlying parameters of the model is of great interest for empirical studies of traveler behavior under uncertainty (*5–7*). For example, only when lateness penalties are disregarded is that functional relationship linear in the standard deviation of travel time, as is frequently assumed.

Department of Economics, School of Social Sciences, University of California, Irvine, Calif. 92717. Current address, R. B. Noland: Environmental Protection Agency, 401 M Street, S.W., Mail Code 2126, Washington, D.C. 20460.

Changes in the level of congestion over the course of the peak period also play an important role in commuter decisions. Rapidly rising congestion shifts the commuter to earlier departure times but also lowers the probability of late arrival. The opposite is true when congestion levels are falling. These types of trade-offs are fully accounted for in the model.

The paper begins with a review of the literature on departure time and route choice, especially previous work dealing with uncertain travel times. The analytical model is then presented and solved. Some numerical examples that provide quantitative information about the possible importance of various components of the model are given. Implications for both research and policy are discussed in the conclusion.

LITERATURE REVIEW

The reliability of arriving at a destination on time is a key component in the decisions made by commuters for their morning trips. Prashker (*8*) attempts to classify some perceived components of reliability into a measurable framework using factor analysis. More recently, researchers have produced direct empirical estimates of how travelers respond to reliability (*5–7*). Much of this work has been aided by the development of stated preference survey techniques.

It is useful to begin with an understanding of how travelers choose departure time choice under certainty. Most research has focused on schedule delay, defined as the difference between the actual time of arrival and some ideal time, usually identified with an official work start time. Typically the commuter is assumed to receive some disutility from schedule delay as well as from travel time (*1,2,9*). In Small's specification (*2*) this disutility is piecewise linear in schedule delay, that is, disutility rises linearly in either the early or late direction. In addition, there is a discrete penalty for being late. In all these studies scheduling disutility is traded off against the possible advantages, caused by variation in congestion over the rush hour, of shifting one's schedule to take advantage of lower congestion. In Cosslett's (*1*) continuous model, this tradeoff appears as a maximization condition involving the slope of the congestion function.

Scheduling models such as these have been incorporated into equilibrium analyses of congestion formation. Basic models for a single link (*10–15*) have been extended to a variety of circumstances including elastic demand (*16,17*); networks (*16,18,19*); heterogeneous commuters, including arbitrary population distributions for desired arrival times (*19,20*); and uncertain capacity or demand (*21,22;* Arnott et al., unpublished data). Small gives a more complete review (*23*). Although most of these analyses use deterministic models of the traveler's choice of departure time, a few (*16,24*) use a discrete-choice model analogous to that of Small (*2*).

Other researchers have incorporated a simpler version of this utility specification into models analyzing uncertain travel times. Gaver (*3*), Polak (*4*), and Bates (*5*) all consider the piecewise linear disutility specification when travel time is uncertain, but none considers congestion that varies over the rush hour. Hence they examine only the trade-offs inherent in trying to minimize the expected disutility from given arrival times given the randomness in travel times.

Mahmassani and associates (*22,25–27*) simulate time-of-day departure choices using hypothetical data collected from actual commuters and fed through a traffic simulation model. These papers focus on day-to-day variations in travel time as commuters gain experience with the system. Although travel times may be uncertain, these simulations emphasize how people learn about the shape of the congestion profile as opposed to uncertainties caused by nonrecurrent events.

Mannering (*28*) and Abdel-Aty et al. (*7*) investigate how likely commuters are to make changes in their departure time or route choices, or both. Mannering finds that those commuters with longer travel times are more likely to make changes and speculates that these trips may have larger variances. His results also indicate that nonrecurrent events may not allow a steady-state equilibrium to evolve, which may have implications for simulating traffic congestion.

Mannering and Hamed (*29*) find empirical evidence that work-to-home departure decisions are influenced by similar factors. Such decisions may not be independent of home-to-work departure decisions: for example, some commuters may delay the morning departure with the intent of staying at work until evening congestion levels have fallen. Neither the model in this paper nor any other one known to the authors attempts to deal with this dependence.

Mahmassani and Herman (*25*) and Mahmassani et al. (*30*) show that commuters tend to adjust departure times more readily than they do routes. In fact, route switches tend to occur when commuters are continually dissatisfied with the outcomes from departure time switches alone (*27,30*). The lower likelihood of route switching adds credibility to models that examine only the choice of departure time, which can have important impacts on the development and timing of peak congestion levels.

ANALYTICAL DERIVATION OF MODEL

A model is described that explains how uncertainty in travel time affects the expected cost of the morning commute. First, the basic components of the cost model, including how changes in congestion levels are accounted for are specified. Then the commuter's scheduling problem is formulated and solved using both a uniform and an exponential probability distribution. The solution is then inserted into the expected cost function to determine how total commuting cost depends on the parameters describing the commuter's travel environment. This cost consists of various components that offer a better understanding of how significant unreliability is as a contribution to travel cost.

Cost Model

The following cost function for the morning commute is assumed:

$$C = \alpha T + \beta\,(\mathrm{SDE}) + \gamma\,(\mathrm{SDL}) + \Theta D_L \tag{1}$$

where

T = travel time;
SDE, SDL = schedule delay early and late, respectively (defined later),
D_L = 1 when SDL > 0 and 0 otherwise;
α = cost of travel time;
β, γ = costs per minute of arriving early and late, respectively; and
Θ = discrete lateness penalty.

The variables SDE and SDL are defined with respect to the official work start time, t_w, and the home departure time, t_h. Let $\mathrm{SD} \equiv t_h + T - t_w$ be "schedule delay," the difference between actual arrival time and official work start time. Define

$$\mathrm{SDL} = \begin{cases} \mathrm{SD} & \text{if } \mathrm{SD} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

$$\mathrm{SDE} = \begin{cases} -\mathrm{SD} & \text{if } \mathrm{SD} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

This formulation of costs is that of Small (*2*) Table 2, Model 1. It could result if pay is docked for late arrival, or if in some other way the frequency and magnitude of late arrival are costly to one's career. Many analyses of time-of-day decisions have used the first three terms of Equation 1; others have implicitly added the fourth term with Θ set to infinity by excluding the possibility of late arrivals. A more complex model formulation could also vary the amount of time spent at work and could thus account for evening travel conditions as additional determinants of the morning commute decision.

The total commute time, T, consists of three elements. T_f is the free-flow travel time when there is no congestion. T_x is the extra travel time caused by congestion, which the traveler is sure to encounter; it is a function of t_h, the home departure time. T_r is the extra travel time caused by nonrecurrent congestion and is modeled formally as a random variable. Following the standard classification of congestion delays into recurrent and incident-related delays (*31,32*), T_x is "recurrent delay" and T_r is "incident delay."

For simplicity it is assumed that the probability distribution of T_r is independent of recurrent congestion and of the time of day of travel. This assumption has the advantage that it enables one to isolate the impact of exogenous changes in travel time uncertainty. Although the assumption may appear unrealistic, there is a surprising absence of clear-cut empirical evidence for alternative assumptions. Satterthwaite (*33*), in a review, finds no reported relations between congested traffic and accidents (which are a primary cause of nonrecurrent congestion). Hendrickson et al. (*34*) analyzed data in Pittsburgh and concluded that variance of travel times is independent of departure times. Richardson and Taylor (*35*) posit a relationship between congested traffic and increases in travel time variability but do not derive an explicit relationship.

To simplify the analysis, define the variable T_e to be the amount one would arrive early if there were no incident-related delays:

$$T_e = t_w - t_h - T_f - T_x \tag{4}$$

As defined by Gaver (*3*), T_e is the "head start" time. Polak's (*4*) "safety margin" is equal to $T_e - E(T_r)$, where $E(T_r)$ denotes the expected incident delay. Note that $T_e > 0$ implies the possibility of early arrival (if recurrent congestion turns out to be nil), whereas $T_e < 0$ implies certain late arrival. Schedule delay can now be written as SD $= T_r - T_e$ and the lateness dummy, D_L, is equal to 1 if $T_r > T_e$ and 0 otherwise.

These definitions enable the cost function to be written as follows:

$$C(T_r) = \alpha[T_f + T_x + T_r] + \beta(1 - D_L)[T_e - T_r] + \gamma D_L[T_r - T_e] + \Theta D_L \qquad (5)$$

Two alternative probability distribution functions for T_r are specified. The first uses a uniform distribution, which assumes that the likelihood of a delay is equal for any level in the domain; the second is an exponential distribution, as in Gaver (*3*), which allows lower levels of delay to have a greater likelihood than longer levels of delay. Many authors, including Richardson and Taylor (*35*), have fit log normal curves to travel time variance data; Giuliano (*36*) has found specifically that nonrecurrent congestion follows a log normal distribution. Unfortunately the log normal distribution is found to be intractable in this model, so it is not pursued here.

Changes in Congestion Levels

Before proceeding with the derivation of expected cost functions, it is convenient to describe how congestion levels change with the choice of departure time, t_h. First, it is possible to describe the commuter's choice of departure time by head start time, T_e, instead of departure time, t_h. To do this, one assumes that T_x, the travel time associated with congestion, is a differentiable function of t_h, $T_x(t_h)$. Differentiating the implicit definition $t_h = t_w - T_f - T_x(T_h) - T_e$, one finds that

$$\frac{dt_h}{dT_e} = -\left(\frac{dT_x}{dt_h}\right)\cdot\left(\frac{dt_h}{dT_e}\right) - 1 \qquad (6)$$

or, solving

$$\frac{dt_h}{dT_e} = \frac{-1}{(1 + T_x')} \qquad (7)$$

where $T_x' \equiv dT_x/dt_h$. The requirement $T_x' > -1$ is imposed to rule out "overtaking," in which a person can arrive earlier by leaving later (*23,37*). This condition guarantees that Equation 7 is well defined and negative. Using Equation 7, the functional relationship between T_x and T_e, defined by $T_x[t_h(T_e)]$, has total derivative

$$\frac{dT_x}{dT_e} \equiv T_x' \cdot \left(\frac{dt_h}{dT_e}\right) = \frac{-T_x'}{(1 + T_x')} \equiv -\Delta \qquad (8)$$

The quantity Δ is a measure of how steeply congestion increases if departure is delayed; more precisely, Δ is the rate at which congestion increases as the "planned" arrival time, $t_h + T_f + T_x \equiv t_w - T_e$, is made later. It has the same sign as T_x'. If $\Delta > 0$, conditions worsen as planned arrival time is delayed, thus favoring earlier schedules; whereas $\Delta < 0$ favors later schedules. Note that the restriction $T_x' > -1$ implies $\Delta < 1$.

Henceforth T_x is regarded as a function of T_e, with well-defined derivative $-\Delta$. As it turns out, making T_x a function of traffic volume at T_e rather than that at t_h is necessary for consistency in an important equilibrium model of endogenous scheduling choice associated with Henderson (*10,11*); see work by Chu (*37*) for a demonstration. If T_x has a kink so that Δ is undefined, corner solutions in addition to those described below become possible.

It is now possible to solve the model for two alternative probability distributions for T_r. In each case the expected cost given scheduling choice T_e is computed; then the choice of T_e that minimizes the expected cost is computed and this chosen value is inserted into the expected cost equation. The resulting expected cost is then a function solely of those parameters that the commuter faces in choosing the schedule for a morning commute trip.

Uniform Distribution

A uniform probability distribution is defined for the domain $[0, T_m]$. The probability density function is defined as $f(T_r) = 1/T_m$ for $0 \leq T_r \leq T_m$, and 0 otherwise. The mean of T_r is $\frac{1}{2} T_m$, and its standard deviation is $T_m / \sqrt{12}$. The mean and standard deviation for the total travel time are $T_f + T_x + \frac{1}{2} T_m$ and $T_f + T_x + (T_m/\sqrt{12})$, respectively.

The expected cost for the morning commute is

$$\text{EC} = \frac{1}{T_m}\int_0^{T_m} C(T_r)dT_r \qquad (9)$$

Substituting Equation 5 into Equation 9, there are three possible cases: (*a*) $0 < T_e < T_m$; (*b*) $T_e \geq T_m$; and (*c*) $T_e \leq 0$. For Case *a*, the chosen departure time can lead to either early or late arrival, depending on the realization of the random variable T_r; Equation 9 becomes

$$\text{EC} = \alpha\left(T_f + T_x + \frac{T_m}{2}\right) + \frac{1}{T_m}\int_0^{T_e} \beta(T_e - T_r)dT_r + \frac{1}{T_m}\int_{T_e}^{T_m}[\gamma(T_r - T_e) + \Theta]\,dT_r \qquad (10)$$

$$= \alpha\left[T_f + T_x + \frac{T_m}{2}\right] + \frac{1}{T_m}[\Theta(T_m - T_e)] + \frac{1}{2T_m}[\beta T_e^2 + \gamma(T_m - T_e)^2] \qquad (11a)$$

$$= \alpha E(T) + \Theta P_L + \beta E(\text{SDE}) + \gamma E(\text{SDL}) \qquad (12)$$

In Equation 11a the first term is merely the expected travel time multiplied by its cost. The second term is the probability of arriving late, P_L, multiplied by the lateness penalty, Θ. The last two terms are the expected cost associated with the amounts of early and late schedule delays.

The other cases result in simple modifications of Equations 10 and 11a. For Case *b*, where $T_e \geq T_m$ (implying the commuter is early with a probability of 1), the limit of integration T_e is replaced by T_m in Equation 10; the result is

$$EC = \alpha\left[T_f + T_x + \frac{T_m}{2}\right] + \beta\left[T_e - \frac{T_m}{2}\right] \qquad (11b)$$

For Case *c*, where $T_e \leq 0$ (implying the commuter is late with a probability of 1), then T_e is replaced by 0 as a limit of integration in Equation 10 and the result is as follows:

$$EC = \alpha\left[T_f + T_x + \frac{T_m}{2}\right] + \Theta + \gamma\left[\frac{T_m}{2} - T_e\right] \qquad (11c)$$

In Cases *b* and *c* the per-minute scheduling cost is simply that associated with the expected arrival time because there is no uncertainty about whether the commuter will arrive late. Equation 12 continues to apply, with appropriately modified expressions for the probability P_L and for the expectations of SDE and SDL. As will be seen, Cases *b* and *c* can occur when the cost parameters and the rate of change in the level of congestion have specified ranges; for example, if Θ is very large or if congestion is increasing rapidly in departure time, one may choose to always arrive early (Case *b*).

The value of T_e that minimizes the expected cost can now be calculated. For Case *a*, the derivative of Equation 11*a* is set to 0, while regarding T_x as a function of T_e as in Equation 8. Solving for T_e gives the following result:

$$T_e^* = \frac{1}{(\beta + \gamma)} (\Theta + \gamma T_m + \alpha \Delta^* T_m) \tag{13}$$

where $\Delta^* = - dT_x/dT_e$ evaluated at T_e^*. The second-order condition requires that $d\Delta/dT_e < (\beta + \gamma)/(\alpha T_m)$, which may be interpreted as requiring that congestion be convex, or at least not too strongly concave, in planned arrival time $(t_w - T_e)$. If T_x is a concave function of $(t_w - T_e)$, then $d^2T_x/dT_e^2 < 0$, that is, $\Delta \equiv - dT_x/dT_e$ is increasing in T_e. This solution is valid only if it is consistent with Case *a* as an interior solution, which requires that $0 < T_e^* < T_m$, that is, $- \gamma T_m < (\Theta + \alpha \Delta^* T_m) < \beta T_m$.

To evaluate the expected cost when T_e is chosen optimally, Equation 13 is substituted into Equation 11a, yielding the following:

$$EC^* = \alpha E(T^*) + \Theta P_L^* + C_S^* \tag{14}$$

where

$$E(T^*) = T_f + T_x[t_h(T_e^*)] + \frac{1}{2} T_m \tag{15}$$

$$P_L^* = \frac{T_m - T_e^*}{T_m} = \frac{\left(\beta - \alpha \Delta - \frac{\Theta}{T_m}\right)}{(\beta + \gamma)} \tag{16}$$

$$C_S^* = \frac{1}{2} \delta T_m + \frac{(\Theta + \alpha \Delta T_m)^2}{2(\beta + \gamma) T_m} \tag{17}$$

and

$$\delta = \frac{\beta\gamma}{(\beta + \gamma)} \tag{18}$$

When $\Theta = \Delta = 0$, Equations 14 through 17 are especially easy to interpret. The probability of being late is then chosen independently of travel time variance and is decreasing in γ/β. In addition, the uncertainty of travel time creates a cost $C_S^* = \frac{1}{2} \delta T_m$ which is proportional to the standard deviation $(T_m/\sqrt{12})$ of travel time and also to the coefficient δ, which is a kind of geometric average of the two schedule delay cost parameters; this cost arises because the commuter is unable to eliminate the likelihood of some schedule delay, either early or late. When $\Theta = \Delta = 0$, the probability of being early is $1 - P_L^* = \gamma / (\beta + \gamma)$ in agreement with Gaver (*3*) Equation 2.3; Polak (*4*) Equation 3.8 (with notation $c_E = \beta$ and $c_L = \gamma$); and Bates (*6*) Equation 17 (with notation $l = \gamma$ and $e - h = \beta$).

The last term in C_S^* may be regarded as the scheduling-cost consequences of shifts in T_e that are made to reduce congestion (if $\Delta \neq 0$) or to reduce the likelihood of a discrete lateness penalty (if $\Theta > 0$). For example, when $\Delta \neq 0$, indicating that some congestion can be avoided by changing the head start, the commuter does so; expected travel time is thereby reduced and C_S^* increased. When $\Theta > 0$, indicating an extra penalty for being late by any amount, T_e^* is increased so as to reduce P_L^*; C_S^* will go up unless a negative Δ was already causing a tendency toward lateness.

Consider now Case *b* of an individual who arrives early with a probability of 1; this occurs if, in Equation 13, $T_e^* > T_m$, that is, if

$$\alpha\Delta^* \geq \beta - \frac{\Theta}{T_m} \tag{19}$$

This case can occur when Θ is high or when congestion is increasing at a rapid rate. In this case, the commuter seeking to minimize cost will choose T_e to minimize Equation 11b. An interior solution occurs when

$$\alpha\Delta = \beta \tag{20}$$

which requires $\Delta > 0$; the second-order condition requires that $d\Delta/dT_e < 0$. Hence the congestion function must have a region where it is a rising convex function of planned arrival time $t_w - T_e$. At Solution 20 the consumer trades off the extra schedule-delay costs of still-earlier arrival (βdT_e) against the saving in travel time cost caused by less congestion $(\alpha\Delta dT_e)$; this is the same tradeoff that forms the basis for determination of early-side arrival times in the models of Vickrey (*12*), Cosslett (*1*), Fargier (*13*), Hendrickson and Kocur (*14*), Arnott et al. (*15*), and others. Alternatively, Case *b* may lead to the corner solution $T_e = T_m$. This will occur if Equation 19 is satisfied but Equation 20 cannot be, as could easily happen if Θ/T_m is large. The interpretation here is that the discrete lateness penalty is large enough for the commuter to eliminate entirely the possibility of late arrival, but variation in congestion, Δ, is not large enough to cause a desire for still earlier planned arrivals.

Consider finally Case *c* of an individual who decides to arrive late with a probability of 1, that is, someone who chooses $T_e \leq 0$. This occurs if $T_e^* \leq 0$ in Equation 13, if

$$\alpha \Delta^* \leq -\left[\gamma + \frac{\Theta}{T_m}\right] \tag{21}$$

This requires that Δ^* be negative, that is, congestion is decreasing and also that neither γ nor Θ be too large. In such a situation, the commuter chooses to incur the relatively mild lateness penalties to take advantage of lessening congestion. Expected cost (Equation 11c) has a local minimum where

$$\alpha \Delta = -\gamma \tag{22}$$

provided again that $d\Delta/dT_e < 0$ (convex congestion function). Again, there could also be a corner solution $T_e = 0$. Note that Equations 21 and 22 are compatible only if Δ changes considerably over the range of possible values of T_e. This could happen if, for example, the interval $[t_w - T_m, t_w]$ occurs near the end of the rush hour, so that Δ^* is strongly negative (representing rapidly falling congestion at T_e^*); the commuter may then choose a later time than T_e^* when both congestion T_x and its rate of change, Δ, are smaller in magnitude, making Equation 22 possible. In fact, if Δ^* is strongly negative there must be a later region where $|\Delta|$ is smaller because T_x cannot fall below 0.

A practical difficulty is to find a reasonable congestion profile that allows one to solve these equations for the optimal head start. A linear congestion profile will work for Equation 13 but not for Equations 20 and 22. Conversely, other functional forms work for Equations 20 and 22 but will not give analytic solutions for Equation 13. An explicit congestion profile is not defined; additional research is examining simulations that endogenously generate congestion profiles (*38*).

Exponential Distribution

The exponential distribution for T_r is defined by the probability density function,

$$f(T_r) = \frac{1}{b} e^{(-T_r/b)} \tag{23}$$

which applies for $0 \geq T_r$. The parameter b is the mean and the standard deviation of the distribution (this differs from the uniform distribution in which the mean is $\sqrt{3}$ times larger than the standard deviation). The exponential distribution more accurately reflects the actual probability of the occurrence of an incident by allowing short delays to have a higher probability of occurrence than longer delays.

Following the same procedures as those described earlier yields an expected cost for the exponential distribution. Assuming that $T_e > 0$, to guarantee an interior solution,

$$\begin{aligned} \text{EC} = & \frac{1}{b}\int_0^\infty \alpha\,(T_f + T_x + T_r)e^{-T_r/b}\,dT_r \\ & + \frac{1}{b}\int_0^{T_e} \beta\,(T_e - T_r)\,e^{-T_r/b}\,dT_r \\ & + \frac{1}{b}\int_{T_e}^\infty [\gamma(T_r - T_e) + \Theta\,]\,e^{-T_r/b}\,dT_r \end{aligned} \tag{24}$$

Note that it is now possible to specify an infinite range for the distribution function. Integration yields the following result:

$$\text{EC} = \alpha(T_f + T_x + b) + \beta\,(T_e - \text{b}) + e^{-T_e/b}\,(\Theta + b\beta + b\gamma) \tag{25}$$

which can be rewritten as

$$\text{EC} = \alpha(T_f + T_x + b) + \beta\,(T_e - P_E b) + P_L(\Theta + b\gamma) \tag{26}$$

where $P_L = e^{-T_e/b}$ is defined as the probability of arriving late, and $P_E = 1 - P_L$ is the probability of arriving early, given $T_e > 0$. This can again be put in the form of Equation 12, where in this case $E(T) = T_f + T_x + b$, $E(\text{SDE}) = T_e - P_E b$, and $E(\text{SDL}) = bP_L$. These expectations can be verified by direct calculations from Equations 2 through 4.

The value of T_e that minimizes expected cost can now be calculated. Taking the derivative of Equation 25 with respect to T_e, setting it equal to 0, and solving for T_e^* gives the following result:

$$T_e^* = b \cdot \ln\left[\frac{\Theta + b(\beta + \gamma)}{b(\beta - \alpha\Delta)}\right] \tag{27}$$

where ln denotes the natural logarithm. When Θ and $\Delta = 0$, implying no late penalty and no change in congestion levels, this formula corresponds to that of Gaver (*3*), Equation 2.5. The second-order condition requires that $d\Delta/dT_e < -1/\alpha b^2 \cdot \exp(-T_e/b) \cdot [\Theta + b(\beta + \gamma)]$, which can simplify to $d\Delta/dT_e < (\alpha\Delta - \beta)/\alpha b$. The probability of being late, $P_L^* = e^{-T_e^*/b}$, can be rewritten as

$$P_L^* = \frac{b(\beta - \alpha\Delta)}{(\Theta + b\beta + b\gamma)} \tag{28}$$

Lateness is favored by small values of θ and γ and by a large negative slope to the congestion function. Equation 27 will have no solution where $\alpha\Delta \geq \beta$, but this is not a problem because if Δ is large enough for this to occur at some head start, then the commuter will seek larger head starts and must eventually find a region where Δ is small. Such a region must exist because T_x cannot be negative.

The interior solution of Equation 27 is valid only when it is compatible with $T_e \geq 0$, the range under which it was derived. That condition is violated if the term in square brackets is ≤ 1, i.e., if

$$\alpha\Delta^* \leq -\left(\gamma + \frac{\Theta}{b}\right) \tag{29}$$

This condition is the same as that in Equation 21, except that T_m is replaced by b (recall that the standard deviation in the uniform distribution is $T_m/\sqrt{12}$, whereas for the exponential distribution it is equal to b). If it holds, the commuter chooses to always be late; expected cost is found by replacing T_e by 0 in the limits of integration in Equation 24, resulting in the following

$$\text{EC} = \alpha(T_f + T_x + b) + \Theta + \gamma(b - T_e) \tag{30}$$

which is equivalent to Equation 11c for the uniform distribution. Head start, T_e, would be chosen either at the corner solution, $T_e = 0$, or at a point where $\alpha\Delta = -\gamma$, just as in Equation 22. This is analogous to Case *c* of the uniform distribution; there is nothing analogous to Case b because the exponential distribution has no upper limit and therefore there is no way to set T_e so that one always arrives early.

Returning to the interior solution (Equation 27), the optimized value of expected cost can be calculated by substituting Equation 27 into Equation 25:

$$\text{EC}^* = \alpha(T_f + T_x + b) - b\alpha\Delta + b\beta \cdot \ln\left[\frac{\Theta + b(\beta + \gamma)}{b(\beta - \alpha\Delta)}\right] \tag{31}$$

The first term is the expected cost of travel time. This can be rewritten to compare with Equation 14:

$$\text{EC}^* = \alpha(T_f + T_x + b) + \Theta P_L^* + C_s^* \tag{32}$$

where P_L^* is given by Equation 28 and

$$C_s^* = b\left\{\beta \cdot \ln\left[\frac{\Theta + b(\beta + \gamma)}{b(\beta - \alpha\Delta)}\right] - \frac{\Theta(\beta - \alpha\Delta)}{\Theta + b(\beta + \gamma)} - \alpha\Delta\right\} \tag{33}$$

The equations derived above describe the expected cost functions associated with uncertainty in travel times. These can be used to evaluate the relative proportion of expected cost associated with travel time uncertainty. The analyses in the next section provide some useful examples showing the relative importance of travel time variance for the cost of commuting.

NUMERICAL EXAMPLES

To analyze the head start times and expected costs associated with travel variance, estimates of the cost coefficients in the models are

TABLE 1 Head Start Times by Standard Deviation and Change in Congestion

	Uniform Distribution: T_e^* (in minutes)		
$T_m / \sqrt{12}$ = Std. Dev.	Δ = -0.1	Δ = 0	Δ = 0.1
5	15.03	15.61	16.19
10	28.23	29.39	30.55
15	41.44	43.18	44.92
20	54.64	56.96	59.28
30	81.05	84.54	88.02
	Exponential Distribution: T_e^* (in minutes)		
b = Std. Dev.	Δ = -0.1	Δ = 0	Δ = 0.1
5	8.74	9.50	10.40
10	16.05	17.57	19.36
15	23.28	25.56	28.25
20	30.49	33.53	37.11
30	44.89	49.44	54.82

needed. Empirical estimates by Small (*2*) of the ratios β/α and γ/α are used in combination with a value of time of \$6.40/hr. These values are also used by Arnott et al. (*15*). The result, using α = 6.40/hr, is β = 3.90/hr and γ = 15.21/hr (rescaled to minutes for these calculations). The authors also use Θ = 0.58 from Small (*2*).

Table 1 shows the values of T_e^* for standard deviations of travel time between 5 and 30 min and for the congestion slopes, Δ, between -0.1 and 0.1. The optimal head start time is always larger with the uniform distribution than with the exponential distribution; this is because of its higher probability weighting for large delays. The head start is larger (earlier departure) when the standard deviation is larger and when congestion is increasing. Table 2 shows the corresponding optimal values of P_L^*, the probability of arriving late, which is smaller when congestion is increasing.

If a hypothetical commuter has scheduling flexibility, then it is possible to assume that $\beta = \gamma$, that is, the commuter is indifferent between schedule delay early and schedule delay late. In addition, this hypothetical commuter would have no lateness penalty, Θ. This can be considered a form of flextime. A commuter with flextime may still have some preferred arrival time, perhaps determined by constraints on the work departure time or personal preferences, such that β and γ are not 0. Table 3 indicates the head start times chosen by such a commuter (with $\beta = \gamma = 3.90$). In all cases the commuter still desires a head start time to avoid congestion, although these values are significantly less than those in Table 1. Note that the head start times increase linearly with respect to the standard deviation because Θ = 0. In the case with no change in congestion levels, $T_e = \sqrt{3} \cdot b$ with the uniform distribution, and $T_e = b \cdot \ln(2)$ with the exponential distribution.

Our analytical derivations have separated the costs associated with travel time, $E(T^*)$, from those associated with the uncertainty of travel time, $C_s^* + \Theta P_L^*$. How important are the relative contributions made by these terms toward the total expected cost of travel, EC*? Table 4 provides a breakdown for each distribution for different levels of travel time uncertainty and Table 5 provides a breakdown for different levels of Δ, excluding the cost of certain travel time, $\alpha (T_f + T_x)$. The total EC* does not differ much between the two distributions, the largest difference being about \$0.73 (when SD = 30). However, C_s^*, the expected cost of schedule delay, is much larger under the exponential distribution than the uniform distribution. For large standard deviations of travel time, C_s^* from the uniform distribution becomes virtually insignificant regardless of

TABLE 2 Optimal Probability of Being Late by Standard Deviation and Change in Congestion

	Uniform Distribution: P_L^*		
$T_m / \sqrt{12}$ = Std. Dev.	Δ = -0.1	Δ = 0	Δ = 0.1
5	13.24%	9.89%	6.55%
10	18.50%	15.15%	11.80%
15	20.25%	16.90%	13.55%
20	21.13%	17.78%	14.43%
30	22.00%	18.66%	15.31%
	Exponential Distribution: P_L^*		
b = Std. Dev.	Δ = -0.1	Δ = 0	Δ = 0.1
5	17.41%	14.96%	12.50%
10	20.10%	17.26%	14.43%
15	21.19%	18.20%	15.21%
20	21.77%	18.71%	15.64%
30	22.40%	19.24%	16.08%

TABLE 3 Head Start Times by Standard Deviation and Change in Congestion with Flex Time

	Uniform Distribution: T_e^* (in minutes)		
$T_m / \sqrt{12}$ = Std. Dev.	Δ = -0.1	Δ = 0	Δ = 0.1
5	7.239	8.660	10.081
10	14.478	17.321	20.163
15	21.717	25.981	30.244
20	28.956	34.641	40.326
30	43.435	51.962	60.489
	Exponential Distribution: T_e^* (in minutes)		
b = Std. Dev.	Δ = -0.1	Δ = 0	Δ = 0.1
5	2.706	3.466	4.362
10	5.412	6.931	8.724
15	8.118	10.397	13.086
20	10.824	13.863	17.448
30	16.236	20.794	26.172

the level or direction of changes in congestion. However, under the exponential distribution, the proportion of expected costs attributable to C_s^* remains relatively stable at about 46 to 48 percent of the total expected costs for each level of standard deviation. This is about the same contribution made by the expected value of uncertain travel time, αb or $\frac{1}{2} \alpha T_m$, which in the case of the uniform distribution accounts for virtually all of the expected costs of commuting. In both distributions the proportion of expected cost associated with the probability of arriving late, ΘP_L^*, decreases as the standard deviation increases; apparently the shifts in head start time shown in Table 1 more than compensate for the increases in standard deviation.

CONCLUSIONS

This research has analyzed the costs associated with uncertain travel times. The work of Gaver (*3*) and Polak (*4*) has been followed but with some new contributions, focusing primarily on scheduling considerations. The effects of congestion that the commuter encounters every day have been explicitly separated from the non-recurrent congestion that accounts for day-to-day variability in travel times. In addition, a discrete lateness penalty, which was originally detected empirically by Small (*2*) has also been introduced.

Using one set of empirically estimated parameters, the expected cost of schedule delay is found to be a relatively minor component of costs when the uniform distribution is used but quite large when the exponential distribution is assumed. In both cases the residual probability of being late is set by the commuter at a small enough value that the expected discrete lateness penalty is only a small fraction of the total costs.

The model described in this paper enables the analyst to predict the expected cost of schedule delay, including penalties for lateness, taking into account how the traveler adjusts the trip schedule in response to the travel environment. Our numerical example suggests costs of several dollars per trip, arising from standard devi-

TABLE 4 Expected Costs of Scheduling and Incident Delay with Uncertain Travel Time (Δ = 0)

	Uniform Distribution				
$T_m / \sqrt{12}$ = Std. Dev.	EC*	C_s^*	%	θP_L^*	%
5	1.0375	0.0564	5.43%	0.0574	5.53%
10	1.9765	0.0411	2.08%	0.0879	4.45%
15	2.9054	0.0360	1.24%	0.0980	3.37%
20	3.8317	0.0335	0.87%	0.1031	2.69%
30	5.6817	0.0309	0.54%	0.1082	1.90%
	Exponential Distribution				
b = Std. Dev.	EC*	C_s^*	%	θP_L^*	%
5	1.1508	0.5307	46.11%	0.0868	7.54%
10	2.2084	1.0416	47.17%	0.1001	4.53%
15	3.2612	1.5557	47.70%	0.1056	3.24%
20	4.3126	2.0708	48.02%	0.1085	2.52%
30	6.4139	3.1023	48.37%	0.1116	1.74%

Note: Costs in dollars per morning commute.

TABLE 5 Expected Costs of Scheduling and Incident Delay with Uncertain Travel Time (SD = 10)

	Uniform Distribution				
Δ	EC*	C_S^*	%	θP_L^*	%
-0.1	1.9827	0.0279	1.41%	0.1073	5.41%
0	1.9765	0.0411	2.08%	0.0879	4.45%
0.1	1.9827	0.0667	3.37%	0.0685	3.45%
	Exponential Distribution				
Δ	EC*	C_S^*	%	θP_L^*	%
-0.1	2.2163	1.0331	46.61%	0.1166	5.26%
0	2.2084	1.0416	47.17%	0.1001	4.53%
0.1	2.2183	1.0679	48.14%	0.0837	3.77%

Note: Costs in dollars per morning commute.

ations of travel time varying from 10 to 30 min. Furthermore, if the exponential distribution applies, about half this cost is due purely to the variance of travel times (the other half being the expected value of the incident delay).

If the expected cost of schedule delay is indeed a major cost of unreliable commute trips, as this suggests, then policies that reduce travel time variance may be preferable in many cases to policies that reduce travel times, especially when the latter are costly. Policies that decrease the response time needed to clear incidents, for example, may be much cheaper than and provide cost savings comparable to capacity expansion.

The information the commuter has about congestion will influence the departure time decision and ultimately the expected cost of commuting. Future work will analyze the impact of providing commuters with accurate information about changes in congestion levels and travel time variance. For example, how will changing information affect head start times when combined with a supply-side congestion model? What are the impacts on congestion when commuters have varying degrees of information about both congestion and reliability? The model presented here offers a starting point for addressing such questions, which are central to the evaluation of intelligent transportation systems.

ACKNOWLEDGMENTS

This research was funded by California Partners for Advanced Transit and Highways and the California Department of Transportation, whose support the authors gratefully acknowledge.

REFERENCES

1. Cosslett, S. The Trip Timing Decision for Travel to Work by Automobile. In *Demand Model Estimation and Validation,* (D. McFadden, A. P. Talvitie, and Associates), Special Report UCB-ITS-SR-77-9. Institute of Transportation Studies, University of California, Berkeley, 1977.
2. Small, K. A. The Scheduling of Consumer Activities: Work Trips. *American Economic Review,* Vol. 72, 1982, pp. 467–479.
3. Gaver, D. P. Headstart Strategies for Combating Congestion. *Transportation Science,* Vol. 2, 1968, pp. 172–181.
4. Polak, J. *Travel Time Variability and Departure Time Choice: A Utility Theoretic Approach.* Discussion Paper 15. Transport Studies Group, Polytechnic of Central London, England, 1987.
5. Bates, J. Modifying Generalised Cost to Deal with Uncertain Travel Times. Presented at 17th Colloquium Vervoersplanologisch, Speurwerk, The Hague, The Netherlands, Nov. 1990.
6. Black, I. G., and J. G. Towriss. *Demand Effects of Travel Time Reliability.* U.K. Department of Transport, London, 1993.
7. Abdel-Aty, M. A., R. Kitamura, P. P. Jovanis, and K. M. Vaughn. *Understanding Commuter's Attitudes, Uncertainties, and Decision-Making and Their Implications for Route Choice.* Draft Research Report MOU 56. Partners for Advanced Transit and Highways, California, Feb. 1994.
8. Prashker, J. N. Direct Analysis of the Perceived Importance of Attributes of Reliability of Travel Modes in Urban Travel. *Transportation,* Vol. 8, 1979, pp. 329–346.
9. Hendrickson, C., and E. Plank. The Flexibility of Departure Times for Work Trips. *Transportation Research,* Vol. 18A, 1984, pp. 25–36.
10. Henderson, J. V. Road Congestion: A Reconsideration of Pricing Theory. *Journal of Urban Economics,* Vol. 1, 1974, pp. 346–365.
11. Henderson, J. V. The Economics of Staggered Work Hours. *Journal of Urban Economics,* Vol. 9, 1981, pp. 349–364.
12. Vickrey, W. S. Congestion Theory and Transport Investment. *American Economic Review, Papers and Proceedings,* Vol. 59, 1969, pp. 251–260.
13. Fargier, P. Effects of the Choice of Departure Time on Road Traffic Congestion. *Proc. 8th International Symposium on Transportation and Traffic Theory,* (V. F. Hurdle, E. Hauer, and G. N. Steuart, eds.), Ontario, Canada, University of Toronto Press, 1983, pp. 223–263.
14. Hendrickson, C., and G. Kocur. Schedule Delay and Departure Time Decisions in a Deterministic Model. *Transportation Science,* Vol. 15, 1981, pp. 62–77.
15. Arnott, R., A. de Palma, and R. Lindsey. Economics of a Bottleneck. *Journal of Urban Economics,* Vol. 27, 1990, pp. 111–130.
16. Ben-Akiva, M., A. de Palma, and P. Kanaroglou. Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates. *Transportation Science,* Vol. 20, No. 2, 1986, pp. 164–181.
17. Arnott, R., A. de Palma, and R. Lindsey. A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand. *American Economic Review,* Vol. 83, 1993, pp. 161–179.
18. Vythoulkas, P. C. Two Models for Predicting Dynamic Stochastic Equilibria in Urban Transportation Networks. In *Transportation and Traffic Theory* (M. Koshi, ed.), Elsevier, Amsterdam, The Netherlands, 1990, pp. 253–272.
19. Arnott, R., A. de Palma, and R. Lindsey. Route Choice with Heterogeneous Drivers and Group-Specific Congestion Costs. *Regional Science and Urban Economics,* Vol. 22, 1992, pp. 71–102.
20. Newell, G. F. The Morning Commute for Nonidentical Travelers. *Transportation Science,* Vol. 21, 1987, pp. 74–88.
21. Arnott, R., A. de Palma, and R. Lindsey. Does Providing Information to Drivers Reduce Traffic Congestion? *Transportation Research,* Vol. 25A, 1991, pp. 309–318.
22. Mahmassani, H., and C. C. Tong. Availability of Information and Dynamics of Departure Time Choice: Experimental Investigation. In *Transportation Research Record 1085,* TRB, National Research Council, Washington, D.C., 1986, pp. 33–49.

23. Small, K. A. *Urban Transportation Economics.* Harwood Academic Publishers, Chur, Switzerland, 1992.
24. De Palma, A., M. Ben-Akiva, C. Lefevre, and N. Litinas. Stochastic Equilibrium Model of Peak Period Traffic Congestion. *Transportation Science,* Vol. 17, 1983, pp. 430–453.
25. Mahmassani, H., and R. Herman. Interactive Experiments for the Study of Tripmaker Behaviour Dynamics in Congested Commuting Systems. Presented at the Oxford Conference on Travel Behavior, New Developments in Dynamic and Activity-Based Approaches, England, 1989.
26. Chang, G., and H. Mahmassani. Travel Time Prediction and Departure Time Adjustment Behavior Dynamics in a Congested Traffic System. *Transportation Research,* Vol. 22B, No. 3, 1988, pp. 217–232.
27. Mahmassani, H., and D. G. Stephan. Experimental Investigation of Route and Departure Time Choice Dynamics of Urban Commuters. In *Transportation Research Record 1203,* TRB, National Research Council, Washington, D.C., 1988, pp. 69–84.
28. Mannering, F. Poisson Analysis of Commuter Flexibility in Changing Routes and Departure Times. *Transportation Research,* Vol. 23B, No. 1, 1989, pp. 53–60.
29. Mannering, F., and M. M. Hamed. Occurrence, Frequency, and Duration of Commuters' Work-to-Home Departure Delay. *Transportation Research,* Vol. 24B, No. 2, 1990, pp. 99–109.
30. Mahmassani, H. S., G. S. Hatcher, and C. G. Caplice. Daily Variation of Trip Chaining, Scheduling, and Path Selection Behavior of Work Commuters. *Proc., 6th International Conference on Travel Behavior,* IATB, Quebec, Canada, 1991.
31. Lindley, J. A. Urban Freeway Congestion: Quantification of the Problem and Effectiveness of Potential Solutions. *Journal of the Institute of Transportation Engineers,* Vol. 57, No. 1, Jan. 1987, pp. 27–32.
32. Schrank, D. L., S. M. Turner, and T. J. Lomax. *Estimates of Urban Roadway Congestion—1990.* U.S. Department of Transportation Report FHWA/TX-90/1131-5. Texas Transportation Institute, Texas A&M University, College Station, 1993.
33. Satterthwaite, S. P. *A Survey of Research into Relationships Between Traffic Accidents and Traffic Volumes,* TRRL Supplementary Report 692. U.K. Transport and Road Research Laboratory, Crowthorne, Berkshire England, 1981.
34. Hendrickson, C., D. Nagin, and E. Plank. Characteristics of Travel Time and Dynamic User Equilibrium for Travel-to-Work. *Proc., 8th International Symposium on Transportation and Traffic Theory* (V. F. Hurdle, E. Hauer, and G. N. Steuart, eds.) University of Toronto Press, Ontario, Canada, 1983, pp. 321–347.
35. Richardson, A. J., and M. A. P. Taylor. Travel Time Variability on Commuter Journeys. *High-Speed Ground Transportation,* Vol. 6, 1978, pp. 77–99.
36. Giuliano, G. Incident Characteristics, Frequency, and Duration on a High-Volume Urban Freeway. *Transportation Research,* Vol. 23A, No. 5, 1989, pp. 387–396.
37. Chu, X. Endogenous Trip Scheduling: The Henderson Approach Reformulated and Compared with the Vickrey Approach. *Journal of Urban Economics,* Vol. 37, 1995.
38. Noland, R. B. Commuter Responses to Travel Time Uncertainty Under Congested Conditions and the Effects of Information. Presented at 1995 Western Regional Science Association Meeting, San Diego, Calif., Feb. 1995.

All responsibility for the contents of this paper remains with the authors.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

New Approach to Route Choice Data Collection: Multiphase, Computer-Aided Telephone Interview Panel Surveys Using Geographic Information Systems Data Base

MOHAMED A. ABDEL-ATY, RYUICHI KITAMURA, PAUL P. JOVANIS, PRASUNA REDDY, AND KENNETH M. VAUGHN

The survey approach is often used in studying drivers' route choice behavior. Surveys enable the researcher to analyze route choice behavior and the effects of traffic information directly from reported behavior and perceptions of the respondent. A sample that represents well the population in the study area could facilitate better understanding of actual drivers' behavior and decision processes. Three route choice surveys targeting a random sample of commuters in the Los Angeles area are presented. The first two surveys were 1 year apart, and the third survey was a follow-up mail questionnaire. The surveys involved two innovative techniques that achieve the data collection required for the analyses of route choice, traffic information acquisition, and commuters' potential use of advanced traveler information systems. The first technique is using computer-aided telephone interviews, and the second utilizes geographic information systems capabilities.

The problem of route choice faced by an automobile driver is complex because of the large number of possible alternative routes through the road networks and the complex patterns of overlap between the various route alternatives. There have been several empirical studies of the factors affecting drivers' route choice. In the urban context the situation is not clear; some researchers have concluded that time minimization is the dominant criterion, whereas others have noted the importance of aspects such as road type (*1,2*), avoidance of congestion (*1*), and avoidance of stops and traffic signals (*3*).

The objective of an ongoing Partners for Advanced Transit and Highways project at the University of California is to understand the factors that influence drivers' route choice and route diversion. It is also to investigate the effect of travel time uncertainty and traffic information on route choice and the potential interplay between travel time reliability, advanced traveler information systems (ATIS), and route choice.

To gather the information needed to achieve these objectives, including capturing each respondent's exact commute route(s) by segment, three route choice surveys were designed and conducted. The first two waves utilized computer-aided telephone interviewing (CATI) techniques to capture all the branchings necessary for the surveys' design, and to be able to collect the high level of detailed information, which varies from one respondent to another, in an efficient manner (the first and second CATI surveys were conducted in May and June 1992 and May 1993, respectively). The third survey was a follow-up questionnaire to the second CATI and consisted of a mail questionnaire (conducted in October 1993), which involved a high level of customization and used information collected in the preceding two surveys, such as the exact commute route by segment, travel time, and an optimal route generated according to each respondent's origin/destination using a geographic information system (GIS) and network data bases of the study area. These survey techniques enabled gathering the data needed to perform the required analyses to the network level in an efficient and unprecedented manner.

This paper presents the design and administration of the three surveys, together with a discussion of each survey's objectives and data collected.

LITERATURE REVIEW

Surveys have been used in several studies with the aim of determining respondents' route choice behavior (and traffic information use). A large-scale survey could achieve a sample size that adequately supports quantitative modeling and forecasting and constitutes a data base for a better understanding of drivers' behavior and decision processes.

Haselkorn et al. (*4,5*) used a large-scale, on-road, mail-back survey that targeted a specific freeway corridor (I-5) in the state of Washington. Mannering et al. (*6*) used the same data set to investigate commuters' route, mode, and departure time flexibility and the influence of traffic information. Khattak et al. (*7*) used mail-back questionnaires to evaluate the effect of traffic reports on commuters' route and departure time changes. The questionnaires were distributed at downtown parking facilities. In a study by Hatcher and Mahmassani (*8*) to observe route and trip scheduling decisions for evening commuters, a mail survey was conducted in two stages. An initial short screening survey and a second stage survey sent to 331 selected first-phase respondents consisted of detailed diaries of actual departure times, route description, and intermediate stops (trip-chaining) information.

To investigate commuters' flexibility in changing routes and departure times Mannering (*9*) surveyed 117 commuters by telephone. Ullman et al. (*10*) also surveyed 44 subjects by telephone to

M. A. Abdel-Aty, Department of Civil and Environmental Engineering, University of Central Florida, P.O. Box 162450, Orlando, Fla. 32816. R. Kitamura, P. P. Jovanis, P. Reddy, and K.M. Vaughn, Institute of Transportation Studies, 2028 Academic Surge, University of California, Davis, Calif. 95616.

study the effect of a freeway corridor's attributes on motorist diversion responses to travel time information.

Polydoropoulou et al. (*11*) used a mail survey and trip diaries to collect data from MIT commuters to model the influence of traffic information on drivers' route choice behavior. Khattak et al. (*12*) used a stated preference approach to evaluate the effects of real-time traffic information along with driver, roadway, and incident characteristics on drivers' willingness to divert. Khattak et al. (*13*) studied stated and reported route diversion behavior and its implications on the benefits of ATIS. Questionnaires were distributed to peak-period commuters crossing the Golden Gate Bridge (San Francisco) during both morning and afternoon rush hours.

All but two of the studies cited used the mail survey design in one way or another. Mail surveys, in general, yield low response rates and do not provide interaction between the interviewer and the respondent. Possibly because of the limitation of mail surveys, none of these surveys has examined the exact routes taken by the drivers. Mannering (*9*) and Ullman et al. (*10*) surveyed the respondents by telephone. However, the administration did not involve any computer-aided structure that enables the survey's design to include branchings to account for differences between commuters in an efficient and timely manner and keeps the survey time to a minimal while gathering all the desired information. Also the computer software checks for errors and inconsistent responses and enables the interviewer to correct them. The random dialing performed in this study achieved a sample that well represents the population in the study area, which is missing from most of the studies cited above, for example, Haselkorn et al. (*4,5*), Khattak et al. (*7,12,13*), Ullman et al. (*10*), Polydoropoulou et al. (*11*), Huchingson et al. (*3*), which targeted either a specific corridor, central business district, or employees of a specific agency. Although Khattak (*12,13*) used stated preference approaches to investigate route switching and information use, neither of his studies accomplished the high level of customization, which was used in both the revealed and stated preference sections of the third survey. Finally, the third and last survey proposes a new application of GIS and data bases in surveying route choice behavior including a high level of customization. This approach was never used in any survey design to study route choice behavior.

RESEARCH DESIGN

The three surveys presented in this paper are designed to collect detailed data on commuters' route choice, information acquisition, and the interplay between route choice and traffic information including the potential effect of ATIS on route choice. The exact commute routes by segment and possible alternative routes were also sought to perform route choice analysis to the network level. The amount and complexity of the data required made it impossible to collect the data in one survey. In addition, there was a need to notice any changes in the routes used during a period of 1 year to investigate the reasons for these changes. Therefore there was a need to conduct a second CATI 1 year after the first one to investigate route changes as well as to gather more detailed data about commuters' attitudes and perceptions. The third survey was a follow-up to the second CATI. The survey design required introducing to the respondents alternative routes generated by a GIS and network data bases of the study area and stated preference customized scenarios that could be achieved only using mail-back questionnaires.

CATI ROUTE CHOICE SURVEY: PHASE 1

A route choice survey was developed targeting Los Angeles-area morning commuters. A mail-out/mail-back survey instrument was initially designed to gather detailed information on commuters' main and alternate routes, to determine the level of information commuters have about these routes, to measure commuters' attitudes toward, and perceptions of, these routes, and to determine how existing traffic information affects their route choice behavior. The mail survey instrument required several branchings, increasing its level of complexity, potentially jeopardizing the response rate and response accuracy. Therefore, it was decided to perform a CATI survey. A CATI survey allows interviewer/respondent interaction and automatically handles branchings with complete reliability and lower interviewer error.

Sampling Procedure

The survey targeted a random sample of households located in the area covered by the South Coast Air Quality Management District, which includes most of the contiguously populated areas of Los Angeles, Orange, San Bernardino, and Riverside counties. The sampling, based on a Mitofsky-Waksberg cluster sampling design (*14*), covered both listed and unlisted numbers. The Mitofsky-Waksberg sampling reduces the number of unproductive dialings and improves efficiency. It was estimated that the Mitofsky-Waksberg design showed an increase of 8 percent efficiency when compared with simple random digit dialing (*15*).

Respondents were limited to adults over 17 years of age who had worked at a fixed location outside their homes at least 1 day in the previous week. For households with more than one qualifying member, the targeted respondent was either the full-time worker (> 20 hrs/week) who answers the phone or who was present at home at that time (this could introduce bias if there is more than one full-time worker but is not considered crucial for this study); a randomly selected part-time worker if there are no full-time workers in the household; or the lone part-time worker. Interviews were performed during weekday evenings and on weekend days. Two callbacks were attempted before a sample telephone number was abandoned. Increasing the number of callbacks is desirable to eliminate nonresponse bias. However past experience has indicated that the effectiveness of additional callbacks diminishes, and the marginal cost per completed interview rapidly increases after two callbacks. Considering available budgets, it was determined to limit callbacks to two.

Survey Content

The survey yielded 944 completed interviews contacted between mid-May and early June 1992. The following information was obtained from each respondent:

- Identification of specific primary commute route by segment (each different road/freeway in sequence for the whole commute route);
- Availability of alternate commute routes and identification of secondary route by segment;
- Detailed information on both primary and secondary routes, including perceived traffic conditions;

- Individual's perception of the severity of different types of delays and other problems;
- Information that respondents receive before and during the commute and its effect on their behavior and awareness of the highway/street network; and
- Demographic and socioeconomic data, including household income, gender, employment status, and education level.

As mentioned earlier, one of the main objectives of this survey was to collect the exact commute route(s) of the respondents (i.e., collect each segment of the route), and to capture the respondents' perceptions and knowledge about these routes. The design of the survey needed to be flexible enough to allow respondents to describe the number of routes they used, the number of segments on each route and the name or number of the street or freeway, and the traffic conditions on each segment, requiring several branchings and interviewer/respondent interaction, which could be achieved only by computer. Figure 1 illustrates a flow chart of the survey, which shows the branchings performed to collect the required data.

Description of Sample

Summary statistics for the sample are presented in Table 1. To test the representativeness of the sample, several socioeconomic and commute characteristics were compared with, and statistically tested against, the 1990 Census (*16*), the 1991 California Statewide Travel Survey results (*17*), and the 1990 California Statistical Abstract (*18*). In most cases the null hypothesis that the values from the route choice survey are not different from the corresponding statistical sources was not rejected at the 0.05 level of significance, implying that the sample well represents the population in the study area. A research report (*19*) illustrates the tests performed with the three cited data bases (among the variables tested with these three data bases are income, mode split, home ownership, and gender) across the four counties. Table 2 shows examples of the comparisons performed for income and mode split.

CATI ROUTE CHOICE SURVEY: PHASE 2

The second route choice survey was developed and targeted the same individuals of 1992, which consisted of the Los Angeles-area morning commuters. Because the survey design also required many branchings, it was decided to also perform a CATI.

Survey Content

The survey was designed to

- Measure any changes within the last year, including home and work locations and primary and secondary routes [in the case of any changes the exact primary route (and secondary route if the respondents use one) is identified by segment];
- Gain an in-depth understanding of commuters' perceptions and decisions of various commute characteristics and problems; and
- Study the effect of travel time uncertainty on route choice using a simple stated preference choice set.

This survey design did not only require capturing the exact commute route and its segments as the CATI I survey, but it required capturing any changes, which complicated the design to a large extent. For example, some commuters changed their origin, destination, primary commute route, secondary commute route, or mode of travel, or a combination of these cases. The survey design had to follow each path of questions according to each commuter's circumstances. Again the CATI design would be the only method to achieve this objective efficiently and promptly. Figure 2 illustrates a flow chart of the survey, which shows the branchings performed to collect the required data.

Description of Sample

A maximum of 10 callbacks was attempted before abandoning a respondent's number, which yielded 564 interviews completed (about 60 percent response rate) in May 1993 (1 year after the first survey of May/early June 1992). Table 3 shows the breakout of the contacted and noncontacted first year's respondents. About 26.5 percent of the respondents either had a disconnected telephone or moved, or had the telephone for less than 1 year, which shows the very high degree of mobility that people in southern California have. The other reasons for not participating in the second survey wave (e.g., refused to participate, had a Fax machine) account for only 13 percent of the respondents. This indicates that increasing the number of call backs would not have increased the response rate (10 callbacks are already very extensive).

Attrition Model

To identify the factors that lead a commuter to participate in the second wave of the survey, a binomial probit model was developed. An attrition model also can be used to develop weights to adjust the sample for further analysis.

The model shown in Table 4 illustrates several socioeconomic factors that increase the probability of the household participation in the second survey. High-income households (income of at least $75,000) and highly educated respondents (college graduates) were more likely to participate. The number of years living at the present address and home ownership affected positively the likelihood of participation. Also whether the respondent in the previous survey was a woman increased the probability of participation in the second year.

Use of the likelihood ratio to test the null hypothesis that all the coefficients are 0 except the alternative specific constant was rejected $\{-2[L(C) - L(0)] = 25.9, df = 5$—number of constrained coefficients, which shows that the model is significant, although the likelihood ratio index is considerably lower. However, this is a good sign because this probably means that attrition is largely random.

Changes in Route Choice

One of the main objectives of the 1993 survey was to measure any changes in the commute routes and to understand the reasons for this change. The survey showed that 195 respondents changed their primary commute routes. Of these, 50 changed their home location and 89 changed their job location. The rest (56 respondents) changed their primary route as a result of factors related to the route itself, and this could be classified as follows:

- Changed route to avoid congestion: 10 respondents;

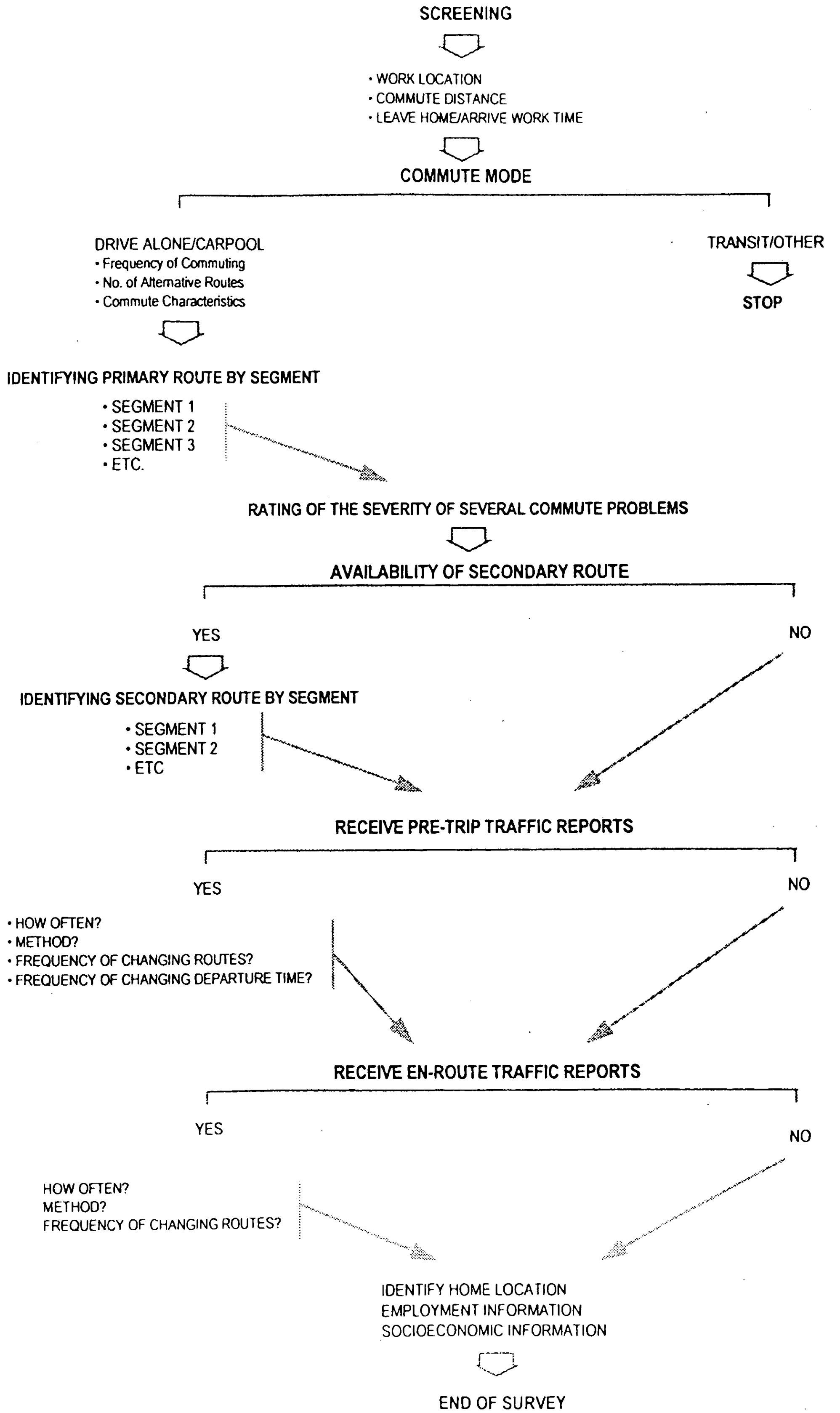


FIGURE 1 Flow chart of CATI I survey.

TABLE 1 Sample Summary Statistics (Averages Unless Noted)

Statistic	Value
Commute distance on usual route (miles)	12.75
Travel time on usual route (minutes)	28.14
Trip duration (including stops)	31.9
Percent of respondents commuting in single-occupant autos/carpool/public transit	78.8/14.6/4.9
Percent receiving pre-trip traffic reports	36.5
Percent receiving en route traffic reports	51.25
Percent of respondents with flexible/ somewhat flexible / fixed work starting time	24.4/30.4/45.2
Percent male/female	51.3/48.7
No. of household cars	2.31
No. of years at present address	7.24
No. of years at present job location	5.52
Percent own/rent their homes	59/41
Household income	38,750
Percent of college graduates	43.8
Think traffic congestion is a problem or major problem (percent)	61.3
Think trip time uncertainty is a problem or major problem (percent)	31.9

TABLE 2 Comparison of Income and Mode Share for the Sample, California Statewide Travel Survey, California Statistical Abstract, and 1990 Census

County	Average Household Income			Median Income
	Survey	CA Statewide Travel Survey 1991 using only study area residents	CA Statistical Abstract (1990)	Census (1990)
Los Angeles	32,500	32,750	38,138	34,965
Orange	43,250	40,655	36,151	45,922
San Bernardino/ Riverside	33,500	28,805	35,004	33,081
Overall Sample	38,750			

County	Percent of Drive alone, Carpool and Public Transit Users					
	Survey			Census 1990		
	Drive Alone	Carpool	Public Transit	Drive Alone	Carpool	Public Transit
Los Angeles	79.1	15.2	5.7	85.6	15.5	6.5
Orange	83.2	14.9	1.8	90.4	13.7	2.5
San Bernardino/ Riverside	82.2	15.6	2.2	91.8	17.3	0.8

Note: Totals add up to more than 100% in the 1990 Census because it accounts for multiple mode users. Statistically testing if the percent of carpoolers is not different from expected values from the 1990 census, is not rejected.

- Found a faster route: 7 respondents;
- Work-related reasons (final destination is the same but intermediate stops change, for example, construction worker in different sites): 16 respondents;
- Change mode (e.g., change from drive alone to carpooling): 8 respondents;
- Road construction: 4 respondents;
- Change route to drive more on freeways: 1 respondent;
- Change to be able to use freeway: 1 respondent;
- Opening a new on-ramp enabled freeway use: 1 respondent;
- Avoid traffic signals: 1 respondent;
- Avoid a particular roadway segment with bad pavement condition: 1 respondent; and
- No specific reason (e.g., mood): 6 respondents.

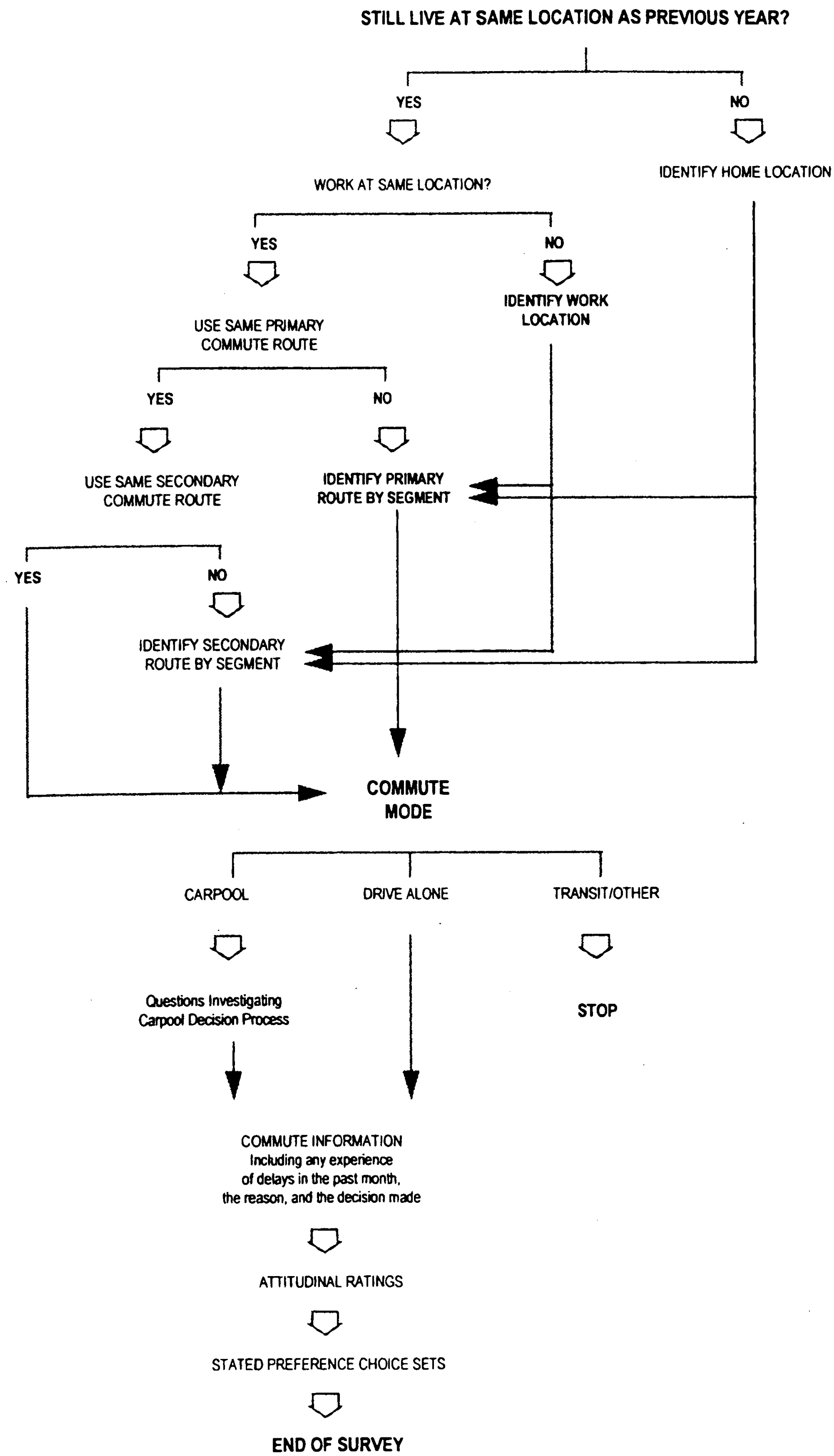


FIGURE 2 Flow chart of CATI II survey.

TABLE 3 Breakout of Frequency of Each Category of Noncontacted Respondents

Case	Number of first year's respondents
- Complete	564 (59.75%)
- Incomplete	
• Disconnected telephone no.	82 (8.70%)
• No one by the required name	59 (6.25%)
• Had the phone for less than 1 year	58 (6.14%)
• Moved from this address	51 (5.40%)
• Answering machine	41 (4.30%)
• No answer and reached maximum no. of call backs	30 (3.20%)
• Refused to participate	28 (3.00%)
• Reached maximum no. of call backs without being able to interview the respondent (e.g someone else answers the phone)	22 (2.30%)
• Fax machine	9 (1.00%)
Total	944 (100%)

TABLE 4 Binomial Probit Model Estimating Whether Respondents Continue To Participate in Second Wave of Survey

	Coef.	t-stat.
Constant	-0.0682	-0.737
X_1 Income dummy (1 if income ≥ $75,000, 0 otherwise)	0.2978	2.309
X_2 Level of education dummy (1 if respondent is a college grad., 0 otherwise)	0.1354	1.378
X_3 Female dummy (1 if female, 0 otherwise)	0.2129	2.265
X_4 No. of years living at present address	0.0105	1.538
X_5 Home ownership dummy (1 if respondent owns his home, 0 otherwise)	0.1380	1.293
Summary Statistics		
$L(0) = -529.564$		
$L(C) = -508.160$		
$L(\beta) = -495.232$		
Number of observations = 764		

Note: Variables' coefficients are defined for participating in the second survey

These responses show that the main reasons for changing the commute routes since the first wave of the survey were changing home or work location, avoiding congestion, and discovering a faster route.

Investigating Effect of Travel Time Variation on Route Choice

To investigate the effect of travel time variation on route choice, it was decided to include repeated hypothetical choice sets in the CATI survey. A major concern was that the design of the stated preference (SP) choices could be complicated to achieve the trade-offs between reliable and the unreliable routes, while trying to make the design of the choice sets as easy as possible to be understood on the telephone. The degree of travel time variation needed also to be as realistic as possible, which rules out a design that includes choice set with a large variation on one of the routes. If the hypothetical commutes were posed in a form that cannot be thought of as an actual commute, then one would have a reason to suspect whether the respondent's hypothetical choices would relate to actual ones.

Therefore, the SP choices were designed to be as simple as possible, so that respondents could comprehend and answer the choice sets on the telephone. Five SP choices are included in the survey. In each choice the respondent is asked to choose between two hypothetical routes. The first route has always fixed travel time every day (5 days a week), whereas the second route has a possibility that the travel time increases on some day(s): for example, Route 1—travel time 30 min everyday; Route 2—travel time 20 min 4 days per week and 40 min 1 day per week. In this case the respondent is informed that a choice of Route 1 will ensure that travel time will be 30 min every day, but a choice of Route 2 means the possibility that on any 1 day of the week travel time could be 40 min and on the other 4 days it could be 20 min.

The choices are designed such that the travel time on the first route is longer and certain, whereas that of the second route is shorter but uncertain. Each respondent is presented with five choices, in which Route 1 is certain and longer, whereas Route 2 is shorter and has different levels of variation. The mean travel time on the Route 2 changes and reaches in some choices the mean of Route 1. The average travel time on Route 2 ranged between 24 and 30 min (which is equal to the mean of Route 1). The standard deviation ranged between about 5 min and about 33 min. The sequence of the choices are randomized (different from one respondent to another) to avoid any ordering bias. The objective of this part of the survey is to measure and investigate whether commuters choose longer certain routes or shorter uncertain routes, and if so, to what extent the uncertainty is that will cause them to choose the route with the fixed travel time. A companion paper presents in detail the design of these SP choice sets together with the model estimated using these data (binary logit model with normal mixing distribution to account for the correlation of disturbances resulting from using multiple observations) (See another paper by Abdel-Aty et al. in this Record).

Turning to the frequency of choices for each case, it is clear that in Cases 2, 4, and 5 the majority of the respondents had chosen Route 1. These cases have the largest standard deviations on Route 2 (> 10 min), and also the mean travel time on Route 2 is either 28 or 30 min (the mean on Route 1 is always 30 min). In Case 1 both routes were chosen almost equally; the mean and standard deviation on Route 2 are 24 and about 9 min, respectively. In Case 3, where the standard deviation is the least and the mean is 24 min, Route 2 was chosen by the majority of the respondents. This means that the respondents correctly recognized the time savings and degree of variation and were willing to tolerate travel time variation to a certain limit; after that limit they were more likely to use the certain (although slightly longer) route. Analysis of these data underscored the significant effect of travel time variation on route choice. The results showed that the disutility of 1 min standard deviation on a route is exactly equivalent to a savings of 1 min of travel time (Abdel-Aty et al. in another paper in this Record).

CATI ROUTE CHOICE SURVEY: PHASE 3

A route choice survey was developed targeting a subsample of the respondents interviewed in the second CATI survey.

Survey Objectives

The survey was designed to obtain the following information:

- Which route attributes are considered important by the individual in the decision process that leads to the choice of a route;
- Commuters' willingness to use ATIS; and
- The effect of advanced traffic information on route choice.

Response Rate

The number of targeted respondents was restricted by the availability of their addresses and the success in geocoding their origins and destinations using the GIS. Therefore, 263 respondents' origins and destinations (O-D) were successfully geocoded and their addresses were available (they agreed to provide the address during the second CATI survey). The 263 questionnaires were customized according to each respondent's origin, destination, primary route, and travel time. The questionnaire included each respondent's primary route (from the CATI surveys), an optimal route generated using O-D information and customized SP choice scenarios using primary route and actual travel time data. The questionnaires were sent to the respondents along with a postage-paid return envelope and an incentive of $2.00. A total of 143 respondents completed and returned the questionnaires (54.4 percent response rate), which is considered a very good response rate for a mail-back survey.

Survey Design

As mentioned, the questionnaires were customized for each respondent. Each questionnaire consisted of two main parts. The first is a revealed preference (RP) section, whereas the second is an SP section.

Revealed Preference Section

The main objective of the RP section is to understand why commuters choose a particular route (in this case their primary route); why they do not necessarily use the optimal route; how they perceive both primary and optimal routes; how familiar they are with their streets/highways network; and how willing they are to use and accept the advice of ATIS.

The primary route for each respondent is identified from the previous CATI surveys. If respondents stated in the second CATI that they did not change their primary route then this route is captured from the first CATI; if they stated that they did change their primary route, then this route is captured from the second CATI. Each segment of the primary route is presented to the respondent in a table; then the respondent is asked to rate a series of subjectively measured route attributes related to his primary route.

Given each respondent's origin (home) and destination (work), and using GIS capabilities, the commercial navigation data bases are used to generate minimum path routes. These data bases have details that include all the highways/streets network in the study area. The experience of a large number of drivers that are acquainted with the area indicates that according to their chosen routes each route is assigned with a weight that also enters into the algorithm calculating the optimal (fastest) route. Figure 3 shows the fastest route by segment as presented in the questionnaire. The fastest route is followed by several questions that measure the respondent's familiarity with this route, the willingness to use an ATIS system, and the rating of a series of route attributes related to the route.

The RP data will support developing a route choice model (the choice set is binary: the primary and GIS-based routes) using both subjectively and objectively measured variables. This model also can be combined either sequentially or simultaneously with a route choice model based on the SP data, including the effect of traffic information.

The following route was generated by the computer as an alternative route from your home to work. The questions below are about this alternate route.

POSSIBLE ALTERNATE ROUTE		
Seg #	Road Segment	Distance (miles)
1	S WESTGATE AVE	0.1
2	WILSHIRE BLVD	0.8
3	I-405 SAN DIEGO FWY S	0.7
4	SANTA MONICA BLVD/CA-2 HWY	2.0
5	AVENUE OF THE STARS	0.4

10. Assuming that you use this route from your home to your work in typical traffic conditions, what would be your estimation of the travel time? _______ (minutes)

11. To what extent do you consider yourself familiar with this route?

☐1 Extremely familiar
☐2 Very familiar
☐3 Somewhat familiar
☐4 Not very familiar
☐5 Not at all familiar

12. Have you ever used this alternate route shown on page 3?

☐1 Yes ☐3 No
☐2 Used a part (or parts) of the route

FIGURE 3 Example of optimal route.

SP Section

The main objective of this section is to investigate the effect of ATIS together with road type, travel time, and familiarity with a particular route, on the route choice. SP methods become an attractive option in transportation research when RP methods cannot be used in a direct way to evaluate the effect or demand for nonexisting services (e.g., ATIS). SP methods are easier to control, more flexible, and cheaper to apply (as each respondent provides multiple observations for variations in the explanatory variables).

In this survey, respondents are provided with three scenarios; in each, they have to choose between two routes and indicate their departure times (Figure 4 shows an example of one of the scenarios). The choices are binary: Route 1 is customized for each respondent so that the SP design would be as realistic as possible, whereas Route 2 is hypothetical. For Route 1 it is stated: "Your primary route using . . ." and then a segment of the respondent's actual route is written. The travel time of Route 1 is the respondent's actual commute time as stated in the CATI surveys, and the road type is the actual route type of his or her primary route (mainly freeway, mainly surface streets, or freeway/surface streets). The objective here is to use the route that the respondent is familiar with, and make the SP design realistic. The road type of Route 2 is either mainly freeway, mainly surface streets, or freeway/surface streets.

For the travel time on the alternative route to be as realistic as possible, and because both routes have the same origin and destination,

PART II

On the following 2 pages, we are asking you to choose from among two routes, the first is similar to your primary route, while the second is a hypothetical route.

Suppose one day you are choosing between the following two routes from your home to work

	Route 1 Your primary route using OHIO ST	Route 2
1. Road type	Surface streets	Mainly Freeway
2. Normal Travel Time	15 minutes	13 minutes
3. Traffic Information		
• Estimated travel time on this day	Not available	13 minutes
• Information on the cause of the delay	----	----

24. Given these choices, which route would you choose on this particular day?

☐ 1 Route 1 ☐ 2 Route 2

25. When would you leave home on that day? _________ AM

FIGURE 4 Example of route choice question.

the travel time on both routes is likely to be close to a great extent. Therefore, normal travel time on Route 2 is one of the following:

0.9 * (Normal travel time on Route 1)

1.0 * (Normal travel time on Route 1)

1.1 * (Normal travel time on Route 1)

Traffic information is available on either Route 1 or Route 2, but not both. If traffic information is available then it gives an estimation of the travel time on that day, which is one of the following values:

0.9 * (normal travel time on the same route)

1.0 * (normal travel time on the same route)

1.1 * (normal travel time on the same route)

1.2 * (normal travel time on the same route)

1.4 * (normal travel time on the same route)

These values are chosen to be as realistic as possible to represent light and usual traffic conditions (Factors of 0.9 to 1.1), mild traffic conditions (factor of 1.2), and heavy traffic conditions that might be caused because of an accident (factor of 1.4), for example.

If the information system estimates a travel time above normal, the cause of the delay is given to the respondent. The cause of the delay is either accident, maintenance, stalled vehicle, or regular congestion. An ATIS was defined to the respondents as a system that can offer personalized information about a trip and offer advice about other routes while considering current traffic conditions.

All possible combinations of the previous cases are considered, after excluding the obvious choices (e.g., if Route 1 is faster and has information that predicts no delays). In all, 68 different combinations were used, 3 for each respondent randomly.

CONCLUSIONS

In this paper two innovative techniques in developing route choice surveys are introduced. The paper addresses the use of computer software, GIS applications, and network data bases in designing and undertaking route choice surveys, which yield data for modeling route choice decision making and for network analysis. The work introduces a new application of computers and GIS in transportation engineering. Also the SP techniques presented enabled the collection of data for analyzing the effect of travel time variations on commuters' route choice (which would be difficult to observe because it is time consuming to collect data that support the analysis), and the evaluation of the effect of a nonexisting service (ATIS) on route choice. The potential of these methods in collecting detailed information on commuters' routes are discussed. Analyses of the data collected from the three surveys proved the viability of these methods [see, for example, work by Abdel-Aty et al. (*19–22*) and in a paper in this Record]. In general, these suggested techniques in surveying commuters' route choice behavior could be extended to study different aspects of drivers' behavior and transportation planning.

ACKNOWLEDGMENTS

The authors acknowledge Navigation Technologies for providing the algorithm and the data bases used in developing the mail survey and Strategic Consulting Research for conducting the two CATI surveys. They also thank Thomas Golob, of the University of California at Irvine, for his significant contribution to the design of the first CATI survey. Special thanks to the California Department of Transportation and the Partners for Advanced Transit and Highways for funding this research.

REFERENCES

1. Wachs, M. Relationships Between Drivers' Attitudes Toward Alternate Routes and Driver and Route Characteristics. In *Highway Research Record 197,* HRB, National Research Council, Washington, D.C., 1967, pp. 70–87.
2. Ben-Akiva, M., M. Bergman, A. Daly, and R. Ramaswamy. Modelling Inter Urban Route Choice Behavior. *Proc., 9th International Symposium on Transportation and Traffic Theory,* Delft, The Netherlands, July 1984.
3. Huchingson, R., R. McNees, and C. Dudek. Survey of Motorist Route-Selection Criteria. In *Transportation Research Record 643,* TRB, National Research Council, Washington, D.C., 1977, pp. 45–48.
4. Haselkorn, M., J. Spyridakis, and W. Barfield. *Improving Motorist Information Systems: Towards a User-Based Motorist Information System for the Puget Sound Area.* Final Report. Washington State Transportation Center, University of Washington, Seattle, March 1990.
5. Haselkorn, M., J. Spyridakis, and W. Barfield. Surveying Commuters to Obtain Functional Requirements for the Design of a Graphic-Based Traffic Information System. *Proc., Vehicle Navigation and Information Systems Conference, Part 2,* SAE, Warrendale, Pa., Oct. 1991, pp. 1041–1044.
6. Mannering, F., S.-G. Kim., W. Barfield, and L. Ng. Statistical Analysis of Commuters' Route, Mode and Departure Time Flexibility and the Influence of Traffic Information. *Transportation Research,* Vol. 2C, No. 1, 1994, pp. 35–47.
7. Khattak, A., J. Schofer, and F. Koppelman. Effect of Traffic Reports on Commuters' Route and Departure Time Changes. *Proc., Vehicle Navigation and Information Systems Conference, Proc.,* Part 2, SAE, Warrendale, Pa., Oct. 1991, pp. 669–679.
8. Hatcher, G., and H. Mahmassani. Daily Variability of Route and Trip Scheduling Decisions for the Evening Commute. In *Transportation Research Record, 1357,* TRB, National Research Council, Washington, D.C., 1992.
9. Mannering, F. Poisson Analysis of Commuter Flexibility in Changing Routes and Departure Times. *Transportation Research B,* Vol. 23B, No. 1, 1989, pp. 53–60.
10. Ullman, G., C. Dudek, and K. Balke. Effect of Freeway Corridor Attributes upon Motorist Diversion Responses to Travel-Time Information. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1994.
11. Polydoropoulou, A., M. Ben-Akiva, and I. Kaysi. Revealed Preference Models of the Influence of Traffic Information on Drivers' Route Choice Behavior. Presented at 73rd Annual Meeting of the *Transportation Research Board,* Washington, D.C., Jan. 1994.
12. Khattak, A., J. Schofer, and F. Koppleman. Factors Influencing Commuters' En Route Diversion Behavior in Response to Delay. In *Transportation Research Record 1318,* TRB, National Research Council, Washington, D.C., 1992.
13. Khattak, A., A. Kanafani, and E. Colletter. Stated and Reported Route Diversion Behavior: Implications of the Benefits of ATIS. Presented at 73rd Annual Meeting of the Transportation Research Board, Washington, D.C., Jan. 1994.
14. Groves, R., P. Biemer, L. Lyberg, J. Massey, W. Nicolls, and J. Waksberg. *Telephone Survey Methodology.* John Wiley and Sons, New York, 1988.
15. Frey, J. *Survey Research By Telephone.* Sage Publications, Beverly Hills, Calif. 1983.
16. *1990 Census of Population and Housing, California.* Bureau of the Census, U.S. Department of Commerce, June 1992.
17. *1991 Statewide Travel Survey: Summary of Findings.* California Department of Transportation, Sacramento, Nov. 1992.
18. *California Statistical Abstract 1990.* Department of Finance, Sacramento, Calif., Oct. 1990.
19. Abdel-Aty, M., K. Vaughn, R. Kitamura, and P. Jovanis. *Survey of Route Choice Behavior: Empirical Results from Southern California and Their Implications for Advanced Information systems.* Research Report UCD-ITS-RR-93-12. Institute of Transportation Studies, University of California, Davis, 1993.
20. Abdel-Aty, M., K. Vaughn, R. Kitamura, P. Jovanis, and F. Mannering. Models of Commuters' Information Use and Route Choice: Initial Results Based on a Southern California Commuter Route Choice Survey. In *Transportation Research Record 1453,* TRB, National Research, Council, Washington, D.C., 1994.
21. Abdel-Aty, M., K. Vaughn, P. Jovanis, R. Kitamura, and F. Mannering. Impact of Traffic Information on Commuters' Behavior: Empirical Results from Southern California and Their Implications for ATIS. *Proc., 4th Annual Meeting of IVHS America,* Atlanta, Ga., 1994.
22. Abdel-Aty, M., R. Kitamura, P. Jovanis, and K. Vaughn. *Investigation of Criteria Influencing Route Choice: Initial Analysis Using Revealed and Stated Preference Data.* Research Report UCD-ITS-RR-94-12. Institute of Transportation Studies, University of California, Davis, 1994.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Household Travel Survey Nonresponse Estimates: The Chicago Experience

ASHISH SEN, SIIM SÖÖT, LIDAN YANG, AND ED CHRISTOPHER

Because response rates vary by household type and by neighborhood, certain groups can be underrepresented. Factoring can rectify this situation somewhat for data used for descriptive purposes, but assumptions underlying many model estimation procedures are violated if factored data are used. Perhaps the only practical solution is to increase the sampling rate of underreporting groups. Because sampling rates should be proportional to reciprocals of response rates, a model to estimate response rates is presented. Such a model could be of value for implementing future surveys. A logit regression model was constructed with the demographic data as independent variables. Observations on the dependent variable, response rate, were obtained from a large-scale household travel survey conducted in the Chicago metropolitan area by the Chicago Area Transportation Study.

Understanding traveler behavior is critical to urban transportation planning and modeling. For this reason, travel surveys are conducted periodically, but their response rates are usually fairly low. Written surveys tend to be cost effective but usually produce low response rates (*1*). Even for oral surveys (including telephone surveys or home interview surveys) where these rates are frequently higher, if one considers people without phones and various prescreening processes, response rates are rarely high (*2*).

A low response rate by itself is not much of a problem. The key difficulty is that these rates vary over different groups. Although in descriptive use of data obtained from such surveys, corrections can be made by means of factoring (*3,4*) this avenue is not always available for modeling uses of these data. For example, consider logit modal split models estimated by maximum likelihood methods. The application of maximum likelihood requires assumptions about the distributions of the number of travelers by mode, and this distribution is usually taken to be multinomial. If one scales up the number of travelers taking each mode by some factor, the resultant products will not have a multinomial distribution. Thus, the use of factored data violates an underlying assumption of the procedure used in estimating the model.

Another example is the usual procedure used for estimating gravity-type trip distribution models, which consists of equating estimated and observed origin trip totals, destination trip totals, and frequencies of trip travel times. This procedure is also a maximum likelihood procedure, and the same comments apply to it as for modal split models. Even for trip generation models, the situation is similar (*4*). Whether one applies a categorical method or a "regression" approach, one is using a linear model, and the assumptions used in estimating it are violated if factored data are used.

Thus, in general, factored data cannot be used for model building and they do not need to be, because assumptions underlying the procedures used to estimate the models do not require the data to have been gathered through random sampling. However, the fact still remains that the travel behavior of groups who are underrepresented in the sample remain ill represented in the final model.

One partial solution to this problem is to increase the sampling rate for such groups—that is, to sample a larger number of households from such groups so that, even if they respond at a lower rate, the group is better represented in the survey. In fact, this is seen as the only solution. Fortunately, because the estimation procedures for the models do not require complete randomness, the level of oversampling can be somewhat rough.

To appropriately oversample from underrepresented groups, one needs to identify groups that have low response rates and to estimate response rates for these groups. Because this information is needed before the survey is conducted it must be obtained from previous surveys. However, because previous surveys would have occurred over a different period and perhaps even a different geographical area, it is likely that the estimates would not be precise. Thus, it is not suggested that an improved sampling scheme would render factoring unnecessary for descriptive (as opposed to model building) uses of travel data. However, selective over- and undersampling will improve the quality of the ultimate data set.

More precise knowledge of nonresponse effects can also lead to other benefits. Special efforts could be made to elicit a higher level of response from underresponding groups. In this latter context, this paper describes in passing one such effort with the Hispanic community and how it paid response dividends. But the key reason for estimating response levels for different groups is for model estimation.

In this study, response rates observed in a large-scale mail survey—the Chicago Area Transportation Study (CATS) Household Travel Survey (HHTS)—were used to estimate a regression model that identifies response rates of key population subgroups. The model is described in the next section. The sections that follow it will be devoted to describing the data on which the model is based and the steps used in constructing the model.

The model was based on Chicago data and no claim is made here about the universality of the findings. Given the importance of the use of judicious over- and undersampling, it is hoped that similar models would be constructed elsewhere. Nevertheless, it is guessed that, in most major U. S. cities, adjustment of sampling rates using the results of this model would be better than no adjustment at all.

A. Sen and S. Sööt, Urban Transportation Center, University of Illinois at Chicago, M/C 357, 1033 West Van Buren Street, Suite 700 South, Chicago, Ill. 60607. L. Yang, Metra, 547 West Jackson Boulevard, Chicago, Ill. 60661. E. Christopher, Chicago Area Transportation Study, 300 West Adams Street, Chicago, Ill. 60606.

RESPONSE RATE MODEL

Model Estimates

The final response rate model is a logit-type model with 12 independent variables. It is of the form

$$\text{RATE}_i = \frac{\exp(Z_i)}{1 + \exp(Z_i)} \quad (1)$$

where RATE_i represents the response rate of zone i, and Z_i is estimated by

$$\begin{aligned}\tilde{Z}_i = 0.38 &- 2.89PMACH_i - 1.70\text{PLBR}_i - 1.14\text{PTECH}_i \\ &- 1.40\text{PH1}_i - 0.53\ \text{PH4}_i \\ &+ 0.79PW3_i \\ &- 0.30S1_i - 0.26S3_i \\ &+ 2.83\text{PC0}_i - 1.16\ \text{PC02}_i \\ &- 1.09\ \text{PAFRO}_i - 0.58\text{POTHER}_i \quad (2)\end{aligned}$$

The independent variables in the model were obtained from the Census Transportation Planning Package (CTPP) and are represented by boldfacing in Table 1. The table also presents all variables

TABLE 1 Description of Variables in Initial Model

Variable	Description
PEXEC	Proportion of the labor force with executive, administrative and managerial occupations
PPROF *	Proportion with professional occupations
PMACH	Proportion with machine operator occupations
PLBR	Proportion with transportation and material moving, machine handlers, helpers and labors, household service, and service occupations
PTECH	Proportion with technicians, administrative clerical occupations
POTHOCC	Proportion with other occupations, like arm force or farmers
PH1	Proportion of households with 1 member
PH2	Proportion of households with 2 members
PH3 *	Proportion of households with 3 members
PH4	Proportion of households with 4 or more members
PW0	Proportion of households with no worker
PW1	Proportion of households with 1 worker
PW2 *	Proportion of households with 2 workers
PW3	Proportion of households with 3 or more workers
S1	S1 = 1 if mean household income is less than \$30,000, S1 = 0 otherwise
S0 *	S0 = 1 if mean household income is between \$30,000 and \$60,000, S0 = 0 otherwise
S2	S2 = 1 if mean household income is between \$60,000 and \$100,000, S2 = 0 otherwise
S3	S3 = 1 if mean household income is over \$100,000 for this section, S3 = 0 otherwise
PC0	Proportion of households with no vehicle
PC02	Square of PC0
PC1	Proportion of households with 1 vehicle
PC2 *	Proportion of households with 2 vehicles
PC3	Proportion of households with 3 or more vehicles
PWHITE *	Proportion white
PAFRO	Proportion African American
POTHER	Proportion other races
PEMPLY *	Proportion of the employed to the total labor force
PUNEMP	Proportion of the unemployed to the total labor force
PHISPN*	Proportion Hispanic origin
PNHISP	Proportion non-Hispanic origin

1). All observations are based on square mile micro-zones.

2). The boldfaced variables were those included in the final model. See Equation(2).

3). Other variables are base group variables; those variables with * are initially considered to describe the base group.

initially considered in this study. Those variables were chosen on the basis of previous studies, which revealed that nonresponse usually results in underrepresentation of households with low incomes and low education levels (3). Therefore, such variables as occupation, household size, number of workers in the household, household income, vehicle ownership, and race were considered. However, data availability restrained the choice of variables. For example, the CTPP data do not report educational attainment. But because these data often reflect the degree of literacy and civic consciousness, they may have an influence on the response rates. Therefore occupation was considered as a surrogate for education.

Logit Transformation

In the final model, the variables in each class (e.g., occupation) sum to 1. For example, the sum of PEXEC, PPROF, PMACH, PLBR, PTECH, and POTHOCC is 1. If all these variables were left in the model, an overspecified model with consequent acute multicollinearity or singularity would result because the independent variables including the intercept would then be linearly dependent. The remedy used was to exclude one variable from each variable class. Those variables are indicated by an asterisk in the table and constitute what is called a base group.

When every independent variable is zero valued, all households in the zone have the characteristics of the base group (given in Table 1 without boldfacing). The base household group is white, has two or three members, and owns at least one vehicle. It has one or two people working in a professional, managerial, executive or administrative occupation; the annual household income is in the range of $30,000 to $100,000. In this case, because all independent variables are zeros, $\hat{Z}_i$ equals the intercept $\beta_0 = 0.38$ (Equation 2) and the response rate RATE_i could be estimated to be 0.59 [from Equation (1)].

In the process of variable selection, it was found that many coefficient estimates were close to 0, implying that response rates of population subgroups with correspondent characteristics are close to that of the base group. For instance, people with executive, administrative, and managerial occupations had response rates similar to those of the base population, which is professional. The variables corresponding to these coefficients were excluded from the analysis for the sake of parsimony. In the same way variables with like coefficients were candidates for consolidation. The variable selection process and diagnostics will be further discussed later.

The independent variables used in the model were obtained from the CTPP. Therefore, instead of focusing on individuals or households the focus was on areas of residence. This focus would have been problematic if there had been large demographic variations within zones. It was not a problem in this study (3), and since a very good fit was obtained (subject to the caveats mentioned later), the results appear useful.

When the variable being predicted is a proportion—response rate in this case—the logit model is frequently used. Apart from the fact that it makes no predictions that are larger than 1 or less than 0, which is clearly most appropriate in this context, other reasons have been cited about the value of the model for proportions (5). Further discussion of the interpretation of the model appears elsewhere in this paper.

Although in the transportation literature maximum likelihood methods typically are used for estimating logit models, linear least-squares (LS) methods can be used under certain circumstances, after the dependent variable is transformed and appropriate weights are used (as described later). Indeed, LS methods were used in the earliest applications of the logit models (5). The main advantage of linear LS is the wide availability of diagnostics; moreover, the economical usage of computer time makes it possible to experiment with various variable combinations to find the best model fit.

From Equation 1, the following is obtained:

$$\log \{E(\text{RATE}_i/[1 - E(\text{RATE}_i)]\} = Z_i \quad (3)$$

where $\text{RATE}_i = m_i/n_i$, is the ratio of completed surveys to the total number mailed out to zone i. When n_i is a fixed number, and the values of both m_i and (n_i and ($n_i - m_i$) are large enough, the following transformation (6, p. 188) is used:

$$\log(m_i + 0.5) - \log(n_i - m_i + 0.5) = Z_i + \epsilon_i \quad (4)$$

This transformed function can then be estimated by linear LS. Because the variance of the function on the left side is approximately equal to $\{n_i\, E(\text{RATE}_i)\, [1 - E(\text{RATE}_i)]\}^{-1}$, its reciprocal

$$w_i = n_i\, E(\text{RATE}_i)[\, 1 - E(\text{RATE}_i)] \quad (5)$$

would be the weights for the linear LS. Because the expected value of response rate, $E(\text{RATE}_i)$, is included in the weight and is unknown, an iteratively reweighted LS estimate is needed. Such a procedure is often carried out using a nonlinear LS program (6, pp. 298–318). In this work the SAS nonlinear LS program, PROC NLIN was used. Once they were computed, weights were inserted into a weighted linear LS procedure to take advantage of the diagnostic methods.

Standard errors and t-values of parameter estimates are given in Table 2. Although the resulting R^2 of 0.38 appears to be low, the fit is good as seen from the following observation. The value of

$$s = \sqrt{\sum_{i=1}^{n} e_i^2 \,/\, (n - k - 1)} \quad (6)$$

called root mean squares error in SAS is 1.2. The n is the number of observations, k is the number of independent variables, and e_i's are residuals. The s^2 is an estimate of the variance of the appropriately

TABLE 2 Results of LS Estimates on Travel Survey Response Rate

Variable	b_j	$s.e.(b_j)$	$t(b_j)$
INTERCEP	0.38	0.1465	2.623
PMACH	-2.89	0.3978	-7.272
PLBR	-1.70	0.2335	-7.290
PTECH	-1.14	0.2069	-5.520
PH1	-1.40	0.2304	-6.076
PH4	-0.53	0.2074	-2.571
PW3	0.79	0.2794	2.820
S1	-0.30	0.0595	-5.063
S3	-0.26	0.0793	-3.273
PC0	2.83	0.3402	8.321
PC02	-1.16	0.5329	-2.180
PAFRO	-1.09	0.0720	-15.097
POTHER	-0.58	0.1879	-3.081

($R^2 = 0.38$, $R_{adj.}{}^2 = 0.37$, $s = 1.20$)

weighted residuals. Because the weight is approximately the reciprocal of the standard deviation of Equation 3 as seen earlier, $s^2 \sim 1$ when the model is well specified. For a variety of reasons, the theoretical minimum of 1 is difficult to achieve. Thus, the fit obtained here has to be regarded as excellent.

Model Application

The estimated coefficients of the variables indicate that the independent variables shown in Table 2 have significantly different effects on response rates compared with the base group. As a class of variables, occupation seems to have the greatest effect on response rates. Also household size and vehicle ownership are key variables. However, the coefficients of the variables representing unemployment and households of Hispanic origin are not statistically significant, implying similarity with the base group.

The following example illustrates the use of this model, supposing that there is a diversified midincome zone i, in which

- 40 percent of the labor force is employed as professionals and managers and 21 percent is employed as machine operators; 5 percent have transportation, material moving, machine handling, or service occupations; and the remaining 34 percent are technicians or clerks ($PMACH_i = 0.21$, $PLBR_i = 0.05$, $PTECH_i = 0.34$);
- 64 percent of the households have four or more members; another 27 percent of the households have two to three members; and the remaining 9 percent are single-member households ($PH1_i = 0.09$, $PH4_i = 0.64$);
- All of the households have one or two workers ($PW3_i = 0$);
- The mean household income of the zone is $58,000 ($S1 = 0$, $S2 = 0$);
- 95 percent of the households have at least one vehicle ($PCO_i = 0.05$, $PCO2_i = 0.0025$); and
- 12 percent of the households are African-Americans; 10 percent of the households belong to other minority groups; and the rest of them are nonminority whites ($PAFRO_i = 0.12$, $POTHER_i = 0.10$);

To estimate the response rate of this zone i, Equation 2 may be used obtaining $Z_i \approx -1.21$. The response rate $RATE_{:iti}$ can then be obtained by Equation 1, and its estimate would be 0.23. This information would be helpful in deriving the sampling rate. If the target number of respondents of this zone is 100, then the sample drawn from this zone should be $100/0.23 \approx 435$. More precise knowledge of the nonresponse leads to more effective targeting of the survey distribution and thus helps in obtaining better survey results.

Discussion of Results

This model suggests that individuals with managerial and professional occupations have higher response rates than blue collar workers. Because occupation reflects education, this is not a surprising finding.

Household size is another class of key variables. It is widely believed that larger households are less likely to respond to travel surveys. Because the CATS survey requested that each household member over 14 report all of his or her trips made during a given weekday, large households were candidates for nonresponse. This study suggests that households with four or more members have lower response rates. This low response rate can be partly offset when there are three or more workers in the household. However, it also appears that single-member households have much lower response rates, which may reflect their lifestyle or attitude toward surveying.

Perhaps because of the existence of multicollinearity, household income plays a minor role in this model. The highest response rates come from the middle-income households, namely, $30,000 to $100,000, which is specified as the base group. The low response rates for households with lower income might be because those groups are not comfortable with written surveys: lower income usually is associated with lower education. Household members with high incomes might just be too busy to respond.

In contrast to some previous studies (*7*), this model suggests that households without vehicles are more likely to respond. One reason is that the nonresponse effect is represented by the coefficients of occupation, household size, and income. Second, this is not too relevant for suburban households, where vehicle ownership rates reach 100 percent. Still, the higher response rates partly suggest that people without vehicles have a stronger tendency to respond to the survey. Because of their mobility dependence on public transportation, they could be more sensitive to transportation issues.

Chicago is a socially diverse metropolitan area with large minority neighborhoods. It was anticipated that response rates from some communities would be low. Accordingly, CATS made an effort to approach the Hispanic population to improve their response rates. From the beginning of the survey design, special attention was given to neighborhoods with large Hispanic percentages. Survey subjects with Hispanic surnames received a Spanish-language insert, which explained the importance of completing the survey form and provided a toll-free telephone number for assistance. Almost 100 calls were received, and most were given specific instructions on how to complete the survey. These efforts resulted in an improved response, which is also demonstrated by the model estimates. This effort was feasible because the agency could target Hispanic surnames, but was not practical with the African-American community—another area where lower response rates were anticipated. The model suggests that the response rates from African-American communities are lower than they are for the white population. For the "other" minority category, which includes Asian-Americans and Native Americans, the average response rates are also lower.

DATA AND METHODS

Data and Survey Methodology

The CATS HHTS is a travel diary-type survey conducted in a period from 1988 to 1991, using a mail-out/mail-back format (*8*). This survey format proved to be an effective means of collecting travel data.

The survey area encompassed seven counties in the Chicago metropolitan area. A total of 79,346 of the 2.8 million households in the region received the survey instruments. The 19,314 completed and usable questionnaires resulted in an average return rate of 24 percent. However, response rates varied by area. As indicated in Figure 1, they ranged from 13 percent for the city of Chicago outside of the central business district (CBD) to 34 percent for Kendall County in the far southwest suburb. This corresponds to the widespread belief that nonresponse for mail surveys is greater for low-income and less-educated households, many of which reside in

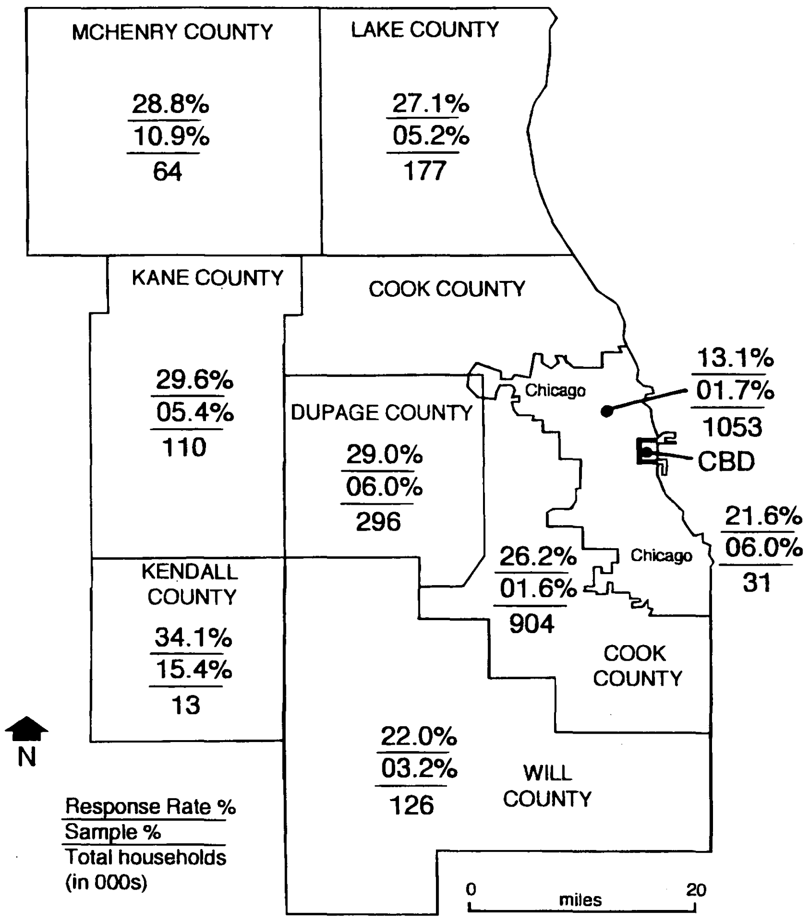


FIGURE 1 CATS household travel survey response rates and sampling sizes.

Chicago. Figure 2 further indicates a wide range of response rates by square-mile zones within the city of Chicago. Response rates were found to be high in the northwest and southwest corners and low west of the CBD.

The sample households were drawn from the local electric company, Commonwealth Edison, records. The number of surveys to be mailed per zone (square-mile area) corresponded to the number of electric meters adjusted by an educated guess of the number of potential respondents. This is the area in which the return rate estimates developed in this paper would have been useful.

Demographic data obtained through the CTPP were used to estimate response rates for population subgroups. These 1990 data were aggregated to a square-mile zone level, defined by the township and range system, to obtain a uniform geographic system with the HHTS. The advantage of the township and range system is that the zones are, for the most part, defined by major arterials resulting in largely homogeneous zones (*9*).

Model Estimate Diagnosis: Outliers and Influential Points

In regression model estimation, it is usual that some observations have large residuals. Sometimes, they occur when some observations reflect conditions or situations different from those under which other observations were obtained. When a few observations with high absolute values of studentized residuals ($|e^*| \geq 3$) were flagged for scrutiny, over half of the zones appeared to belong to Kendall County. This is a suburban county located southwest of Chicago. A rural area recently added to metropolitan Chicago, it consists of two subareas: large residential properties and farms. With a county population of only 38,000, it is vastly different from the neighborhoods throughout the rest of the study area. Therefore, the zones in Kendall County were all excluded.

Two other observations ultimately were deleted. One was a zone in southwest suburban Du Page County. Its response rate (88 percent) was considerably higher than was estimated in the model. A closer study of the zone revealed that this was an area near an employment center. Many residents were professional or skilled workers, either singles or working couples with no children. This group of people was likely to be different from typical urban young professionals, which contributed to a difference in the survey response rate. It was not possible to set a special variable that indicated the difference between urban or suburban young professionals. Besides, the large number of mailed-out surveys resulted in a large weight, which made the point influential. Therefore this data point was excluded from the model estimate. The second case was a zone near downtown Chicago in which there were many large minority households with extremely low incomes. There were also many zero-worker households implying high unemployment rates, which was not revealed in the census

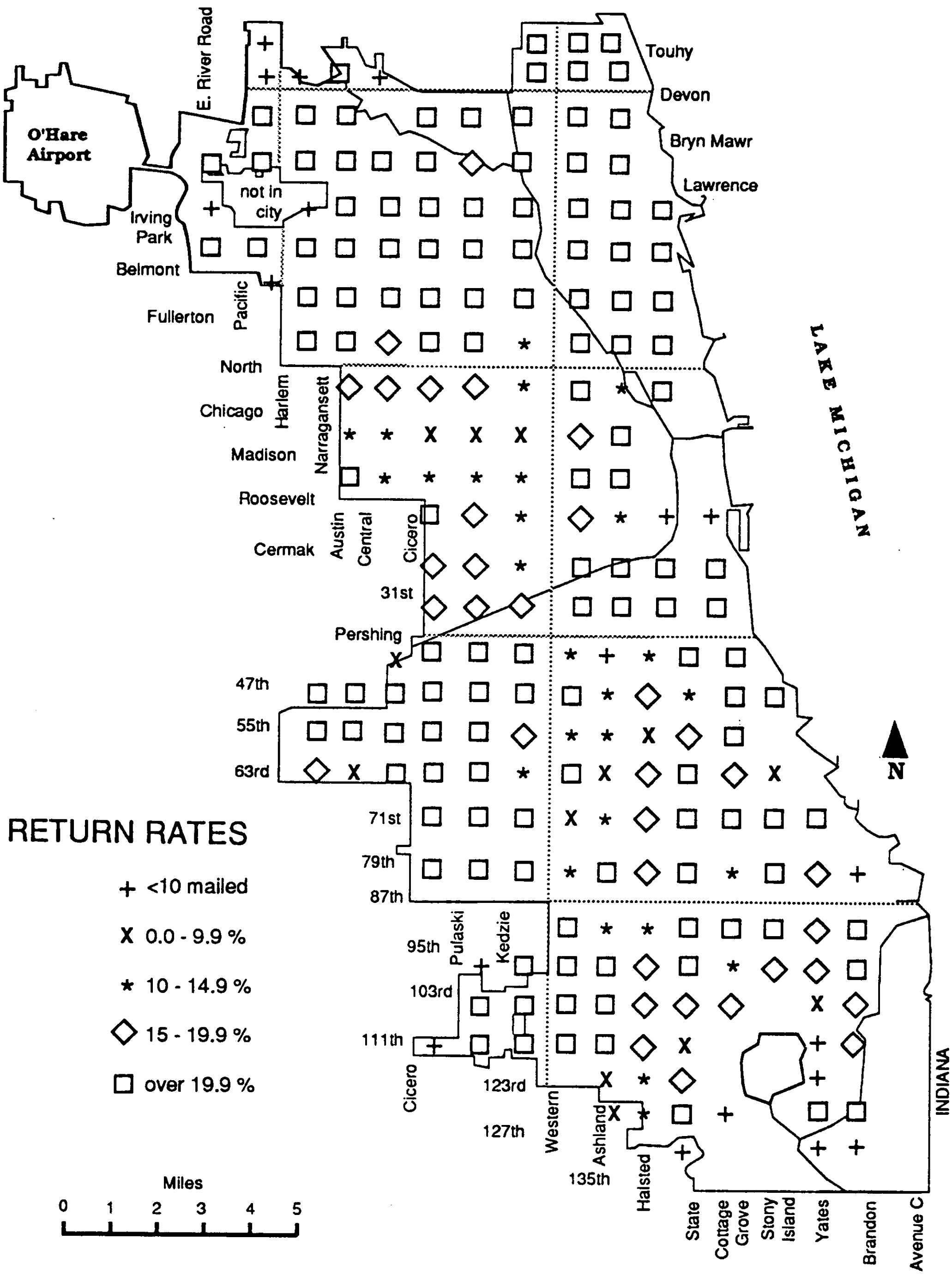


FIGURE 2 **City of Chicago travel survey response rates by square-mile zones.**

unemployment variable. The large mail-out size gave it a large weight and made it an influential point.

Ideally $|e^*| \geq 2$ occurs for about 5 percent of the observations. In the final model (after discarding Kendall County and the other two observations), there are 77 such data points whose e^* exceeds the critical level of $|e^*| \geq 2$. They are approximately 5.3 percent of the total 1,450 observations, further strengthening the earlier conclusion about a good fit of the model.

Plots of residuals against predicted values and each independent value were also carefully examined to check for the existence of unequal variances and the need for transformations. Figure 3 displays a residual plot. Because in this and other plots no obvious patterns were found, the model appeared acceptable.

Variable Selection and Multicollinearity

Variable selection is a critical process in model building. Given the large number of variables used, multicollinearity is likely present. For instance, low income is usually associated with minority and

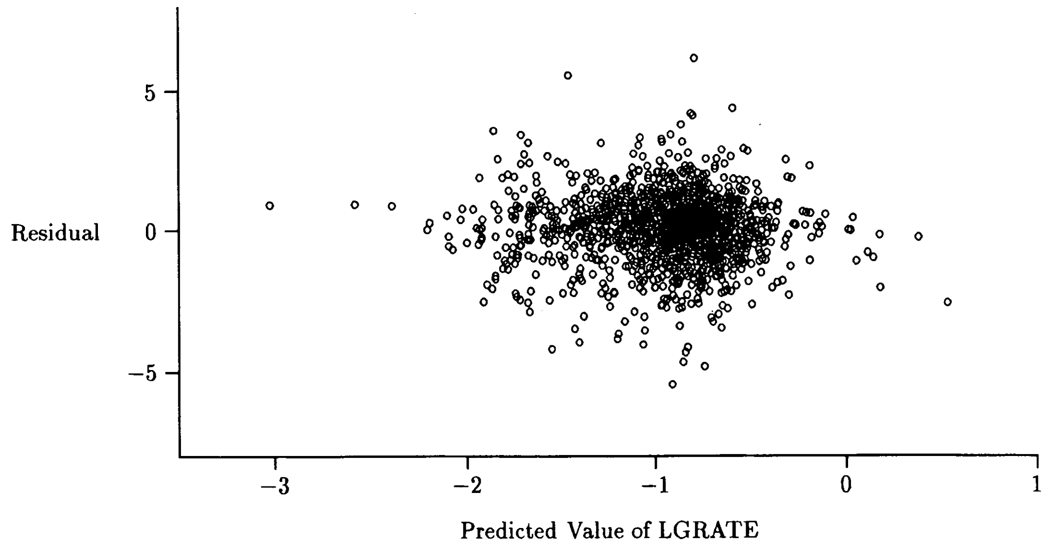


FIGURE 3 Residual versus predicted plot for response rate model.

blue collar occupations; therefore collinearity probably exists among race, income, occupation, and vehicle ownership. In addition, from the collinearity diagnostics, multicollinearity was also found among PH1, PH2, and PH4 (the household-size variables; see Table 1 for definitions). Because the model being constructed is for predictive purposes, multicollinearity is not as much of a problem as for studies in which the significance of variables is the key object—as long as one can reasonably conjecture that the structure of the multicollinearity will be similar for the place and time when the model is applied.

However, parsimonious models are easier to work with and inspire greater confidence because they are easier to interpret. Hence an all-possible-subset-search variable selection procedure was conducted using SAS-weighted PROC RSQUARE. Assuming that the initial model with all variables included is not biased, usually dropping some independent variables will cause bias in the parameter estimates left in the model, except when the values of the deleted variables are orthogonal to those of the remaining variables. To examine the variables, the value of C_p is a good indicator of the presence of noticeable bias. When $C_p \approx p$ (p is the number of independent variables plus one), the bias in the predicteds introduced by dropping variables is usually negligible (*6*, pp. 234–235). Also available are other indicators, such as s^2 and R^2, which estimate the goodness of fit. The combination of those indicators helps in the selection of a concise model with a reasonable level of goodness of fit and low bias. In the all-possible-subsets search, the C_p value began to approach p for $(p - 1) \geq 12$. This indicated that little bias would occur with those sets of suggested 12 independent variables. Their corresponding R^2's were around 0.38. The difference with the highest R^2 occurred only at the third decimal point, which was encouraging.

When variables in one class are similar and their b_j coefficients are alike then consolidation is appropriate. For instance, those occupations requiring analogous skills that also have similar coefficients, such as transportation and material movers, machine handlers, helpers, laborers, and household service and service workers, were combined into one summary variable: PLBR. Also it was seen that response rate estimates for households with one vehicle and households with several vehicles were approximately the same, suggesting that only two variables could be used: households with and without vehicles.

CONCLUSIONS

In a mail-out/mail-back survey, the lower-income, less-educated households are usually underrepresented. Because those population subgroups are usually mobility disadvantaged, it is particularly important to properly estimate and address their needs in transportation planning and policy. However, in many modeling procedures, factoring is not appropriate. Thus, to achieve a desirable number of responses, a carefully designed sampling scheme is necessary in surveys to under- and oversample in subareas.

In this study, such a survey response model was estimated. The model was estimated by linear LS with a logit transformation. Because of the linear LS approach, it was possible to apply a wide range of statistical diagnostic procedures.

The model results do not reveal surprises about response rates. It does, however, provide a means of estimating response rates in urban areas in which considerable demographic variations exist. Although this may not be critical for survey data used for descriptive purposes, it is important if the data are used for modeling.

ACKNOWLEDGMENTS

The authors thank the Illinois Department of Transportation and the Chicago Area Transportation Study for the financial and moral support to undertake this study. In particular the authors express their gratitude to Ed Christopher for his cooperation and assistance in the preparation of this manuscript.

REFERENCES

1. Hassounah, M. I., L. S. Cheah, and G. N. Steuart. Underreporting of Trips in Telephone Interview Travel Surveys. In *Transportation Research Record 1412,* TRB, National Research Council, Washington, D.C., 1993, pp. 90–94.

2. Cohen, B. A., A. P. Lobo, E. L. Fielding, and L. S. Yang. *Analysis of Households Without Phones: Impact for the NPTS 1995.* Office of Highway Information Management, FHWA, U.S. Department of Transportation, 1993.
3. Kim, H. J., J. Li, S. Roodman, A. Sen, and S. Sööt. Factoring Household Travel Surveys. In *Transportation Research Record 1412,* TRB, National Research Council, Washington, D.C., 1993, pp. 17–22.
4. Thakuriah, P., A. Sen, S. Sööt, and E. Christopher. Non-response Bias and Trip Generation Models. In *Transportation Research Record 1412,* TRB, National Research Council, Washington, D. C., 1993, pp. 64–70.
5. Ashton, W. D. *The Logit Transformation with Special Reference to Its Uses in Bioassay.* Hafner Publishing Company, New York, 1972.
6. Sen, A. K., and M. Srivastava. *Regression Analysis, Theory, Methods, and Applications.* Springer-Verlag, New York, 1990.
7. Clark, A. C., and C. Goldstucker. Mail-Out/Mail-Back Travel Survey in Houston, Texas. In *Transportation Research Record 1097,* TRB, National Research Council, Washington, D.C., 1986, pp. 13–19.
8. Ghislandi, A. C., A. R. Fijal, and E. J. Christopher. *CATS 1990 Household Travel Survey: A Methodological Overview.* Working Paper 94-05. Information Services Division, Chicago Area Transportation Study, Illinois, 1994.
9. Christopher, E. J., J. J. Nam, and M. J. Rogus. *CATS Traffic Analysis Zone Structure for the 1990 Census Transportation Planning Package.* Working Paper 94-01. Information Services Division, Chicago Area Transportation Study, Illinois, 1994.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Recursive Structure for Exact Line Probabilities and Expected Waiting Times in Multipath Transit Assignment

R. JAYAKRISHNAN, MICHAEL G. MCNALLY, AND ARUN G. MARAR

Exact analytical expressions for incorporation into transit network assignment frameworks are presented. These expressions apply to the case of random uniform passenger arrivals and fixed constant line headways. Previously, difficulty in specification has led to assumptions such as Poisson line arrivals; the reality, however, conforms more to fixed schedules than to Poisson line arrivals. The exact expressions that are derived define expected waiting times and line ridership probabilities. Recursive schemes are developed for computational implementation of these expressions by which to facilitate their application in practical transit assignment. The expressions were developed for multipath assignment schemes and can be used to enhance existing transit assignment algorithms in commonly used planning packages; applications can be either in the line enumeration phase or in the line ridership probability calculations. Numerical examples are provided to illustrate the application of the recursive schemes, and the predicted line probabilities are compared with simulated passenger and line arrivals.

The motivation for this paper is to improve on certain assumptions and methods that are used in existing transit assignment models and applied in practice for transit planning and operations. The focus is on developing exact expressions for finding line selection probabilities and expected waiting times without making traditional assumptions such as Poisson line arrival probability distributions. Recursive procedures are developed to facilitate the practical implementation of these expressions. The successful design and operation of an efficient transit system rely heavily on the successful implementation of the assignment models utilized in systems planning packages. The transit assignment algorithm developed by Spiess and Florian (*1*) is one of the most popular models and, hence, this algorithm is used as a benchmark to apply the exact expressions and calculation schemes developed. Although the proposed models are possibly most applicable to the Spiess-Florian algorithm, they can be used in other assignment schemes as well.

The paper presents a brief discussion on the history of transit assignment through current multipath assignment models used in practice. It also highlights the relative merits of existing transit assignment models and the implications of passenger arrival distributions and waiting strategies. The proposed models are applied to candidate transit networks, including a simplified real-world network.

OVERVIEW OF TRANSIT ASSIGNMENT

An important component of transit planning is transit path assignment, the prediction of how transit passengers choose paths. Path assignment involves determining the level of service on various paths serving an origin and destination (O-D) and assigning passenger demand to these paths. There are various ways by which traveler route choice may be formulated. One possible assumption would be that all used paths will have the same minimum expected travel time and any unused paths will have travel times that are at least as great as this minimum (a variation of Wardrop's first principle). This principle implies that transit passengers choose a path from a set of paths with the minimum expected travel time.

Transit path assignment, however, rarely considers other important aspects of passenger behavior. Some of the conventional transit assignment methods assume negative exponential headways for transit services. Although it does not necessarily reflect actual behavior, this assumption has been attractive to researchers in the past because of its simplicity. Most multipath assignment models have used approximate expressions for the expected waiting time at a stop and the resulting ridership probabilities. Multipath assignment models assign passengers to a set of paths on the basis of some optimal strategy that typically seeks to minimize the expected total travel time to the destination.

Transit assignment is different from conventional traffic assignment because of the waiting aspect and line transfer requirements. Associated techniques can be broadly classified on the basis of the nature of the assignment as deterministic and probabilistic. Deterministic transit assignment models find a single shortest path between an O-D pair by considering waiting time at a node and possible transfers to other lines. The earliest model developed is Dial's transit pathfinder algorithm (*2*), which is an extension of Moore's shortest-path algorithm accommodating the peculiarities of transit minimum paths. Le Clercq (*3*) suggested a different shortest-path algorithm known as the once-through algorithm for transit assignment as an improvement over the pathfinder algorithm.

Probabilistic models consider the possibility of choosing from a set of lines. Perhaps the earliest work that considers transit assignment on the basis of perceived travel time is by Chriqui and Robillard (*4*). The idea of choosing the first line among a set of lines has been adopted by recent transit researchers (*1,5–7*); the nature of the algorithms differs in the definition of the choice set of transit lines. Spiess and Florian (*1*) developed a mathematical model for enumerating an optimal strategy set of transit lines that aims to minimize the expected total travel time from an origin to a destination. Horowitz (*5*) modified Dial's multipath assignment model to accommodate various level-of-service parameters (such as travel time, waiting time, and capacity), which are used to find the set of reasonable paths between an origin and destination; a logit model was employed to estimate the probability of choosing a particular transit line.

Institute of Transportation Studies and Department of Civil and Environmental Engineering, University of California, Irvine, Calif. 92717.

Passenger Arrivals and Waiting Strategies

In most of the conventional transit assignment methods, the expected waiting time for each route is assumed to be one-half of the route headway (assuming that a passenger arrives randomly in a perfectly reliable system). The expected waiting time for any transfer is again equal to one-half the headway on the connecting route (regardless of transfer timing). Even recently proposed multipath assignment models still assume random passenger arrivals and approximate waiting times as some fraction of the headway. The assumption of random uniform passenger arrivals at a node might hold true in the case of frequent transit service but may not necessarily be warranted for less-frequent bus service (for example, headways greater than 10 min). For such cases, several researchers have suggested that alternate distributions of passenger arrivals be used (*8–10*).

Route Choice

The choice of a route depends on an optimal strategy (or a waiting strategy). A strategy is a set of rules that, when applied, allows a traveler to reach the desired destination. Early transit assignment models specify a single shortest path, whereas probabilistic assignment models specify a set of paths on the basis of an optimal strategy. Actual route choice is more complex because it is a function of a passenger's perception of different level-of-service parameters, such as waiting time, number of transfers, and line capacity.

The validity of Wardrop's principles in transit path choice is questionable because riders often may not select one initial complete path but may show adaptive behavior. In a comprehensive study of the transit route choice problem, Hall (*11*) suggested that passengers are able to improve their travel time over the Wardrop optimum by adaptively selecting routes. The research focus was to determine the importance of real-time information to passengers in making route choice decisions. Passengers can improve their path choice after arriving at a node by using additional information available at that node. The advantages of knowing how much time a passenger has waited at a node in determining route choice when there are overlapping bus routes also was investigated by Marguier and Ceder (*12*).

MULTIPATH ASSIGNMENT MODELS

The most common multipath transit assignment models used in practice are those developed by Spiess and Florian (*1*) and Horowitz (*5*). Spiess and Florian enumerate a set of paths on the basis of an optimal strategy that minimizes the total expected travel time; this model has been implemented in EMME/2. Horowitz applied Dial's stochastic multipath assignment to enumerate the set of reasonable paths from a node to a destination on the basis of the disutility of a transit trip; this transit assignment model is implemented in QRS II.

The Spiess-Florian transit assignment model is an algorithm for solving the transit assignment problem with a fixed set of transit lines. The traveler chooses the strategy that allows a desired destination to be reached at minimum expected cost. A strategy is a set of rules that, when applied, allows the traveler to reach that destination. For the special case in which the waiting time at a stop depends only on the combined line frequency, the problem may be formulated as a linear program of a size that increases linearly with network size. The problem of transit assignment is solved by a label-setting algorithm in polynomial time. In the algorithm's first pass (from the destination node to all origins), the optimal strategy is enumerated and the expected travel times from each node to the destination are computed. In the second pass (from all origins to the destination), the demand is assigned to the network according to the optimal strategy.

The multipath transit assignment model developed by Horowitz (*5*) is a modification of the general stochastic multipath assignment approach of Dial (*13*). Although Dial's model is considered to be efficient, there are instances where travel behavior is inaccurately represented. However, in the case of transit assignment, it can be shown that the potential inaccuracies of Dial's model are of little consequence relative to the anomalies of Dial's algorithm that arise in its performance in automobile networks. Any multipath assignment is based on hypothesized behavior of users; therefore, it is important to understand the passenger's perception of the (dis)utility of the individual routes. The disutility of a transit trip may be represented as a weighted function of the components of travel time, including access and egress walking time, waiting time, and transfer time.

DEVELOPMENT OF EXACT EXPRESSIONS

The major focus of this paper is the development of exact expressions for the line probabilities and expected waiting time for passengers randomly arriving (via a uniform distribution) at a node and choosing a transit line from a candidate set of lines. This candidate set could, for instance, be based on an optimal strategy that minimizes the expected total travel time (such as that proposed by Spiess and Florian). Of particular interest is the case of uniform random arrivals of passengers and constant interarrival times of vehicles, which have been considered difficult in the past and have led to assumptions of Poisson line arrivals. The Spiess-Florian algorithm, for instance, makes such an assumption for finding the line choice probabilities (based on line frequencies).

The expressions developed for the expected waiting time under fixed uniform line arrivals and random uniform passenger arrivals are applicable to assignment algorithms that use calculated expected waiting times during the candidate transit line enumeration phase. The second set of expressions developed for the line selection probabilities for a random passenger is of use in any assignment algorithm, which may be identifying candidate lines using different variables other than travel times and expected waiting times.

Link Probability Expressions

Link selection probability calculations are required in an assignment algorithm that develops the candidate line set by adding links to the set rather than paths. A link here refers to a transit line between two nodes in the network. The link probabilities can be derived from line frequencies, such as in the original Spiess-Florian algorithm:

$$P_a(A_i^+) = \frac{f_a}{\sum_{\forall a' \in A_i^+} fa'} \qquad a \in A_i^+ \tag{1}$$

The probability $P_a(A_i^+)$ of selecting link a at node i from a set of links A_i^+ is the ratio of the frequency of transit service f_a for link a

to the sum of the frequencies of the transit services for all the links. This expression holds true only for negative exponential vehicle headways. This expression is only an approximation in the case of constant headways. Because transit services tend to follow a fixed schedule (with some variance around the schedule), the Poisson arrival assumption does cause significant errors in some cases, and this is examined in a later section. The developments in this paper are based on the belief that a constant interarrival time is a much more logical assumption.

Consider a set of m transit lines at a node that the traveler may pick from to reach the desired destination. Let $H_1, H_2, \ldots, H_m$ be the headways for Lines 1, 2, . . . , m, respectively. Let H_k be the minimum of all these headways (that is, the kth line has the smallest headway). Ties are not restricted and either of the tying lines can be selected.

The assumption of constant headways and uniform passenger arrivals results in a waiting time probability density function for Line i given by $1/H_i$, implying that the randomly arriving passenger will find the arrival of line i to be after any waiting time of up to H_i with equal probability. The joint probability density function of all the m lines is simply the reciprocal of the product of the headways of all the transit lines.

The expression "a line arriving" is equivalent to a vehicle from a particular line arriving. The probability P_i of line i being selected from the set of m lines is the probability of that transit line arriving first among the set of lines. The bus headways $H_1, H_2, \ldots, H_m$ may be in any order. This probability is given by the following m-space integral:

$$P_i = \frac{1}{\prod_{j=1}^{m} H_j}\left[\int_0^{H_k} dh_i \int_{h_i}^{H_1} dh_1 \int_{h_i}^{H_2} dh_2 \ldots\ldots\ldots \int_{h_i}^{H_m} dh_m\right] \qquad (2)$$

The motivation behind this expression is simple: if Line i is to be the first line selected, then it must arrive within a time interval equal to the minimum headway among the set of lines under consideration, and all the remaining lines must arrive at a later time. This exact probability expression is a polynomial integration of degree $(m-1)$ and hence results in a polynomial of degree m. To find a general expression, consider the case for two and three lines. Without loss of generality, assume that the minimum headway among the set of lines is H_1. Expanding that expression yields the probability of selecting line i from the two or three line choice set (the number of lines are indicated parenthetically):

$$P_1(2) = \frac{1}{\prod_{j=1}^{2} H_j}\left[\frac{(-1)H_1^2}{2} + H_1 H_2\right] \qquad (3)$$

$$P_1(3) = \frac{1}{\prod_{j=1}^{3} H_j}\left[\frac{H_1^3}{3} + \frac{(-1)H_1^2}{2}(H_2+H_3) + H_1H_2H_3\right] \qquad (4)$$

Defining $S_j^m(i)$ as the sum of the products of headways of all possible combinations from the set of m lines, not including line i, leads to a general expression for the ridership probability.

$S_1^2(1) = H_2$

$S_1^3(1) = H_2 + H_3$

$S_2^3(1) = H_2 H_3$

From induction, the general result is as follows:

$$P_i = \frac{1}{\prod_{j=1}^{m} H_j}\left[\frac{(-1)^{m-1} H_k^m}{m} + \frac{(-1)^{m-2} H_k^{m-1}}{m-1} S_1^m(i) + \ldots + H_k S_{m-1}^m(i)\right] \qquad (5)$$

The probability expression derived has lost the simplicity of the proportionality expression originally advocated by Spiess and Florian; however, the expression manifests the theoretical predictions without the use of approximations. This is a more precise approach than that of Jansson and Ridderstolpe (*14*), where the probability of selecting a line is found by converting the already selected set to an "equivalent" route. Once a new route is added with share p_u each of the shares in the "equivalent" set is scaled by the factor $(1 - p_u)$.

Expected Waiting Time Formulation

Just as the probability of picking a link is a function of the line frequencies, waiting time is also a function of the headway distribution of the transit services. Again, an example is the expression for the expected waiting time at a node as specified by Spiess and Florian:

$$E[\text{wait}] = \frac{\alpha}{\sum_{a \in A_i^+} f_a} \qquad \alpha > 0 \qquad (6)$$

Spiess and Florian state that

> The case $\alpha = 1$ corresponds to an exponential distribution of interarrival times of the vehicles with mean $1/f_a$ and a uniform passenger arrival rate at the nodes. The case $\alpha = ½$ is an approximation of a constant interarrival time $1/f_a$ for the vehicles on link a. This measure of waiting time is the most widely used approach in practice, in spite of the fact that it is based on a rough approximation. (*1*, p.91)

Although this is a widely used expression, it significantly underestimates the expected waiting time at a node for the special case of constant interarrival times of the transit lines. The actual expected waiting time is the expected value of min $[t_1, t_2, \ldots, t_m]$ where t_i is the time a randomly arriving passenger has to wait until the arrival of a line i service. This is the expected waiting time at a node because the passenger waits for the first transit line that arrives from among the optimal strategy set. The theoretical expression for expected waiting time in the case of constant line headways and uniform random passenger arrivals is

$$E[\text{wait}] = \frac{1}{\prod_{i=1}^{m} H_i} \int_0^{H_k} \left[t \sum_{j=1}^{m} \prod_{\substack{i=1,2..m \\ i \neq j}} (h_i - t)\right] dt \qquad (7)$$

The probability density function (PDF) for Min $[t_i]$ (that is, the PDF for the line i to be picked first from the set of transit lines) is given by:

$$\text{PDF for } \min[t_i] = \frac{1}{\prod_{i=1}^{m} H_i}\left[\prod_{\substack{i=1,\ldots,m\\ i\neq j}} (h_i - t)\right] \tag{8}$$

The summation term over all the lines is to take into account the case for each line being the minimum one to be selected from the set of lines. The limits of integration are 0 and H_k because the maximum waiting time is equal to the minimum headway among the set of transit lines. This seemingly cumbersome equation takes a similar expression as that of the link probabilities, with the polynomial integration being of the order m resulting in a polynomial expression of degree $(m + 1)$. To find a general expression, consider, as before, the expected waiting times for the two and three-line cases (the number of lines are indicated parenthetically):

$$E[\text{wait}](2) = \frac{1}{\sum_{i=1}^{3} H_i}\left[\frac{(-1)2H_k^3}{3} + \frac{H_k^2}{2}(H_1 + H_2)\right] \tag{9}$$

$$E[\text{wait}](3) = \frac{1}{\prod_{i=1}^{2} H_i}\left[\frac{3H_k^4}{4} + \frac{(-1)2H_k^3}{3}(H_1 + H_2 + H_3) + \frac{H_k^2}{2}(H_1H_2 + H_1H_3 + H_2H_3)\right] \tag{10}$$

Denote as R_j^m the sum of the products of the headways of all possible combinations of j lines taken from the set of m lines. Note that the joint probability density function of headways (or waiting time) for the m lines can be denoted as $1/R_m^m$. The above cases result in the following:

$R_1^2 = H_1 + H_2$

$R_2^2 = H_1H_2$

$R_1^3 = H_1 + H_2 + H_3$

$R_2^3 = H_1H_2 + H_2H_3 + H_1H_3$

$R_3^3 = H_1H_2H_3$

From induction, it is possible to find the general form of the R_j^m expressions as

$$E[\text{wait}] = \frac{1}{\prod_{i=1}^{m} H_i}\left[\frac{(-1)^{m-1}mH_k^{m+1}}{m+1} + \frac{(-1)^{m-2}(m-1)H_k^m}{m} R_1^m + \ldots \frac{H_k^2}{2} R_{m-1}^m\right] \tag{11}$$

RECURSIVE SOLUTION FOR MODIFIED ALGORITHM

Theoretical expressions for the expected waiting time at a node and for the corresponding link probabilities have been derived from basic probability fundamentals. However, these expressions increase the computational complexity when implementing the modified model in practice. Computational time is of significant concern in assignment for large transit networks; thus, the expressions must be specified efficiently.

The method proposed here is suitable for algorithms based on shortest-path finding (such as Spiess-Florian), where the process of enumeration of the optimal strategy set of transit lines at a node is started by selecting the link with the least expected travel time. Another link is then added to the optimal set if it improves the expected total travel time. If it does not improve the expected travel time, the link is discarded and never again considered in the enumeration. Thus, links are added sequentially to define the optimal strategy.

A major portion of the computational time is spent in the calculation of the expressions for R and S at each stage of the algorithm. For example, when the fourth line is added, R_1^4, R_2^4, R_3^4, and R_4^4 must be calculated. It is possible, however, to compute these values at each stage of the algorithm in a more elegant way without a significant increase in computational burden. The following steps will illustrate the manner in which the algorithm can incorporate the expected waiting time and link probability expressions while maintaining control of the computational expense.

Consider an optimal strategy set consisting of m lines and assume that each line is added in the order 1, 2, . . . , m (i.e., Line 1 is selected first, then Line 2, and so on, until Line m is selected last). The algorithm has as many stages as there are transit lines in the optimal strategy. Stage 1 is executed when the first line is added to the optimal strategy set at a node. Stage 2 is executed when the second line is added to the optimal strategy set and so on until Stage m, which is executed when the mth line is added to the optimal strategy set. The R_j^m values at stage m can be obtained from the R_j^{m-1} values of the prior stage; the S-values for each stage are directly obtained from the R-values of that stage. The expected waiting times and ridership probabilities are calculated by using the following recursive structure:

- **Stage 1: Number of lines = 1; Line 1 enters.**
  $R_1^1 = H_1$
  $E[\text{wait}] = H_1/2$
  $P_1 = 1.00$
- **Stage 2: Number of lines = 2; Line 2 enters.**
  $R_1^2 = R_1^1 + H_2$
  $R_2^2 = R_1^1 H_2$
  $S_1^2(1) = R_1^2 - H_1$
  Find min $[H_1, H_2]$
  Find E [wait]
  Find P_i, $i = 1,2$
- **Stage 3: Number of lines = 3; Line 3 enters.**
  $R_1^3 = R_1^2 + H_3$
  $R_2^3 = R_2^2 + R_1{}^2 H_3$
  $R_3^3 = R_2^2 H_3$
  $S_1^3(i) = R_1^3 - H_i$ **for** $i = 1,2$
  $S_2^3(i) = R_2^3 - H_i[S_1^3(i)]$ **for** $i=1,2$
  Find min $[H_1, H_2, H_3]$
  Find E [wait]
  Find P_i, $i=1,3$
- **Stage m: Number of lines = m; Line m enters.**
  $R_1^m = R_1^{m-1} + H_m$
  $R_j^m = R_j^{m-1} + R_{j-1}^{m-1} H_m$ **for** $j=2, (m-1)$
  $R_m^m = R_{m-1}^{m-1} H_m$
  Do $i=1, (m-1)$
  $S_1^m(i) = R_1^m - H_i$
  $S_j^m(i) = R_j^m - H_i[S_{j-1}^m(i)]$ **for** $j=2, (m-1)$

Continue
Find min $[H_1, H_2, \ldots, H_m]$
Find E [wait]
Find P_i, $i=1,m$

The recursive steps developed facilitate the computation of the values of the R's and the S's at each stage on the basis of the values from the prior stage.

ILLUSTRATED EXAMPLE

A four-line case has been chosen to illustrate the recursive algorithm; let the headways of these four lines be 30, 20, 12, and 4 min. Assume that these lines form the optimal strategy set. Note that whether a link becomes a part of the optimal strategy is a function of link cost (which typically in the case of the transit assignment is the link travel time), whereas the expected total waiting time and link probabilities are functions of headways only. Therefore, it is not necessary to consider the link travel times for each transit line.

- **Stage 1: Line 1 enters the optimal strategy set.**
  $R_1^1 = H_1 = 30$
  E [wait] = 15 min
  $P_1 = 1.00$
- **Stage 2: Line 2 enters the optimal strategy set.**
  $R_1^2 = R_1^1 + H_2 = 50$
  $R_2^2 = R_1^1 H_2 = 600$
  $S_1^2(1) = R_1^2 - H_1 = 20$
  Min $[H_1, H_2]$ = 20 min
  E [wait] = 7.78 min
  $P_1 = 0.3333$
  $P_2 = 0.6667$
- **Stage 3: Line 3 enters the optimal strategy set.**
  $R_1^3 = R_1^2 + H_3 = 62$
  $R_2^3 = R_2^2 + R_1^2 H_3 = 1{,}200$
  $R_3^3 = R_2^2 H_3 = 7{,}200$
  $S_1^3(1) = R_1^3 - H_1 = 32$
  $S_1^3(2) = R_1^3 - H_2 = 42$
  $S_2^3(1) = R_2^3 - H_1 S_1^3(1) = 240$
  $S_2^3(2) = R_2^3 - H_2 S_1^3(2) = 360$
  Min $[H_1, H_2, H_3]$ = 12 min
  E [wait] = 4.24 min
  $P_1 = 0.16$
  $P_2 = 0.26$
  $P_3 = 0.58$
- **Stage 4: Line 4 enters the optimal strategy set.**
  $R_1^4 = R_1^3 + H_4 = 66$
  $R_2^4 = R_2^3 + R_1^3 + H_4 = 1{,}448$
  $R_3^4 = R_3^3 + R_2^3 H_4 = 12{,}000$
  $R_4^4 = R_3^3 h_4 = 28{,}800$
  $S_1^4(1) = R_1^4 - H_1 = 36$
  $S_1^4(2) = R_1^4 - H_2 = 46$
  $S_1^4(3) = R_1^4 - H_3 = 54$
  $S_2^4(1) = R_2^4 - H_1 S_1^4(1) = 368$
  $S_2^4(2) = R_2^4 - H_2 S_1^4(2) = 528$
  $S_2^4(3) = R_2^4 - H_3 S_1^4(3) = 800$
  $S_3^4(1) = R_3^4 - H_1 S_2^4(1) = 960$
  $S_3^4(2) = R_3^4 - H_2 S_2^4(2) = 1{,}440$
  $S_3^4(3) = R_3^4 - H_3 S_2^4(3) = 2{,}400$
  Min $[H_1, H_2, H_3, H_4]$ = 4 min
  E [wait] = 1.60 min
  $P_1 = 0.055$
  $P_2 = 0.085$
  $P_3 = 0.150$
  $P_4 = 0.710$

Note that the link probabilities of $(m - 1)$ lines are calculated using the expression that, for the m^{th} line, is computed as follows:

$$1 - \sum_{i=1}^{m-1} P_i \qquad (12)$$

DISCUSSION OF RESULTS

The comparison of the Spiess-Florian method with the exact method indicates a significant difference in the values of the expected waiting time and the probabilities predicted for uniform passenger arrivals at a node (Table 1). It is important to note that the actual probabilities depend on the initial line starting times (the relative headway gaps at the start of the time horizon). The link probabilities obtained from the exact method give the probability of taking a line, which is unconditional on the initial arrival times. If two routes have identical 10 min headways, with the first arrival on Line 1 on the hour and arrivals on Line 2 lagged by 2 min, then the first route receives 80 percent of the randomly arriving passengers and the other only 20 percent. This is not the equal split that would be expected using either the exact method or the Spiess-Florian algorithm. Particularly in the case of routes with identical headways, the variance of probabilities conditional on the starting time is high. With the aid of a simulation study, it was shown that, in most cases,

TABLE 1 Comparison of Spiess-Florian with Exact Method

Stage No:	Line No:	Headways (minutes)	Spiess-Florian Method		Exact method	
			E [Wait]	Probability	E [Wait]	Probability
1	1	30	15	1.000	15	1.000
2	1 2	30 20	6	0.400 0.600	7.78	0.333 0.667
3	1 2 3	30 20 12	3	0.200 0.300 0.500	4.24	0.160 0.260 0.580
4	1 2 3 4	30 20 12 4	1.2	0.080 0.120 0.200 0.600	1.60	0.055 0.085 0.150 0.710

the variance of the probabilities caused by different starting times was low enough to minimize any concerns (Table 2). Moreover, the initial starting times for transit services over a large network are variable and hence the predicted probabilities, which are unconditional on the starting times, are the best estimates on average, especially during the planning process when assignments are used. For combinations of headways that do not share many common divisible factors (for example, headways of 5 and 12 min), the actual link probabilities, irrespective of the starting time, closely approximate the theoretical unconditional values provided that the time horizon is sufficiently long. A typical peak-hour operation of 3 hr is sufficient to expect the system to converge to the theoretical values of the probabilities, as confirmed by the simulation study.

IMPLICATIONS OF USING EXACT FORMULATIONS

The exact formulations derived for the probabilities and waiting time for constant headways and uniform random passenger arrivals were implemented in an operational version of the Spiess-Florian algorithm. This task was somewhat formidable given the complexity of both the derived expressions and the resulting algorithm. The Spiess-Florian algorithm performs efficiently largely because of the nature of the expressions for line probabilities and waiting times. This algorithm adds lines in the order of increasing link costs. Consider a simple example where there are n lines operating between a single O-D pair. Let the link costs be $t_{1,2}, \ldots, t_{n-1}, t_n$ such that $t_n > t_{n-1} > \ldots > t_1$. Let $H_1, H_2, \ldots, H_n$ be the corresponding headways. Two rules are presented next. The first rule is implicit in the Spiess-Florian algorithm and is used as the criterion for choosing an optimal strategy set. Because the optimal strategy set is different according to the exact expressions, the second rule is provided as a supporting rule to enumerate the optimal strategy set in this case.

Rule 1: Conventional Rule for Enumeration of Optimal Strategy Set

The Spiess-Florian algorithm adds Line 1 into the optimal strategy set and then Line 2 if $t_2 < E$[travel time]. For the general case, assume that $(n - 1)$ lines are already a part of the optimal strategy set. The expected waiting time for the $(n - 1)$ lines is

$$E[\text{wait}] = \frac{1}{2\sum_{i=1}^{n-1} \frac{1}{H_i}} \tag{13}$$

The probability of picking a line i is given as

$$P_i = \frac{\frac{1}{H_i}}{\sum_{i=1}^{n-1} \frac{1}{H_i}} \tag{14}$$

Using these expressions, the expected total travel time is given by

$$E[\text{total travel time}] = E[\text{wait}] + \sum_{i=1}^{n-1} P_i t_i \tag{15}$$

Define a and b as

$$a = \frac{1}{2} + \sum_{i=1}^{n-1} \frac{t_i}{H_i}$$

$$b = \sum_{i=1}^{n-1} \frac{1}{H_i} \tag{16}$$

It follows from these expressions that the expected total travel time is given by (a/b). If the nth line is added to the optimal strategy, then the new a' is given by $[a + (t_n/H_n)]$ and the new b' is given by $[b + (1/H_n)]$. The new expected travel time is given by a'/b'. This line would be a part of the optimal strategy only if $(a'/b') < a/b$ {i.e., if $[a + (t_n/H_n)]/[b + (1/H_n)] < [a/b]$}.

Cross-multiplying yields

$$ab + b(t_n/H_n < ab + a(1/H_n) \tag{17}$$

This expression readily simplifies to

$$t_n < a/b \tag{18}$$

Because a/b is the expected travel time with $(n - 1)$ lines, the inclusion of a new line has to satisfy the preceding criterion as employed by the Spiess-Florian algorithm, that is, the travel time on the new link is less than the current expected travel time to the destination. Any line that satisfies this criterion also becomes a line in the opti-

TABLE 2 Summary of Simulation Results for Two-Line Case

	Same Headways (min)		Different Headways (min)	
	10	10	12	5
Time Horizon (hours)	3		3	
Probability (Simulation)	0.485	0.515	0.209	0.791
Probability (Theoretical)	0.500	0.500	0.210	0.790
Expected Wait Time (Simulation)	3.907		2.149	
Expected Wait Time (Theoretical)	3.333		2.160	
Probability Variance	0.086		0.001	
Expected Waiting Time Variance	8.075		3.856	

mal strategy set because it also improves the expected total travel time. It is a crucial difference in the context of implementing the exact expressions because all the lines that satisfy this condition may not be in the optimal strategy set. It is not necessary, however, that an entirely new criterion be developed. The Spiess-Florian criterion can be used to eliminate unfavorable lines; the strategy set needs to be further reduced to make it optimal because the addition of a line that satisfies the above criterion may not improve the expected total travel time according to the exact expressions.

Rule 2: Supporting Rule To Enumerate Optimal Strategy Set in Exact Formulation

A supporting rule must be developed for the previous criterion to find the optimal set of lines at a node that minimizes the expected total travel time. The following rule is proposed:

> One by one, add the lines from the ordered set that satisfy the criterion stated in Case 1 and determine the expected total travel time as each line is added. Store the expected total travel time values at each stage (typically the dual variables in the algorithm). Once the set of lines is exhausted, pick the minimum and eliminate all the lines that were included after the minimum line.

This rule ensures that some of the lines included in the strategy set according to Rule 1 are removed once all the lines are considered. Only the lines up to and including the line that caused the minimum expected travel time remain in the final set. Thus Rule 1 is used to make an initial reduction of the strategy set, and Rule 2 is used to find the optimal set. It can be seen that because the strategy set between any two nodes is optimal, all the arguments presented by Spiess and Florian for the optimality of the algorithm apply here too.

Special Example of Implementation of Exact Expressions

Consider a two-line example with headways for Lines 1 and 2 of 20 and 6 min, respectively. The lines connect Nodes A and B directly. Applying the Spiess-Florian algorithm, Line 1 is added first to the optimal strategy set because it has the lowest link cost (the expected total travel time via Lines 1 and 2 are 13 and 15 min).

Considering Line 2, the selection criterion $u_B + C_a$ is 12 min and would be considered favorable by the Spiess-Florian algorithm because it is less than u_A (currently equal to 13 min). This is true because the expected total travel time from A to B is improved to 12.24 min. However, the exact expressions yield an expected total travel time of 13.35 min, which is inferior to the current value of u_A. The inclusion of Line 2 has therefore increased the expected total travel time despite satisfying the selection criterion.

- Spiess-Florian algorithm:
  –Expected wait time = $1/2[1/(1/20+1/6)] = 2.31$
  –Probability of picking Line 1 = $1/20/[1/20+1/6] = 0.23$
  –Probability of picking Line 2 = $1/6[1/20 + 1/6] = 0.77$
  –Expected total travel time = $2.31 + 0.23 * 3 + 0.77 * 12 = 12.24$ min.
- Exact expressions:
  –Expected wait time = $1/120[(-1) * 2 * 6^3/3 + 1/2 * 6^2 * (20 + 6)] = 2.70$
  –Probability of picking Line 1 = $1/120[1/2 * 6 * 6] = 0.15$
  –Probability of picking Line 2 = $1/120[120 - 1/2 * 6 * 6] = 0.85$
  –Expected total travel time = $2.7 + 0.15 * 3 + 0.85 * 12 = 13.35$ min

Line 2 cannot be excluded from further consideration at this stage because the addition of another line could make it favorable. The main implication in the application of the exact expressions is that the expected total travel times calculated at each stage must be stored and compared for the minimum value only after all lines at a node have been evaluated.

APPLICATION TO REAL TRANSIT NETWORK

To illustrate the implementation of the new transit assignment model, a small network of four lines has been extracted from the in transit network Orange County, California. The optimal strategy to reach the University of California, Irvine (Node B) from the Santa Ana Transit Terminal (Node A) is to be identified. Line 65 connects the O-D pair directly, whereas Line 61 requires a transfer at Node C to either Line 382 or Line 74 to reach the destination. Figure 1 illustrates the subnetwork along with the headways and link travel time (both in minutes). Figures 2 and 3 give the ridership probabilities according to the Spiess-Florian algorithm and according to the proposed algorithm using the exact expressions for expected waiting time and ridership probabilities, respectively. The Spiess-Florian algorithm was independently coded; the EMME/2 program was not used.

The network was simulated with an O-D demand of 500 person trips per hour and for a time horizon of 3 hr. To ensure that the resulting line probabilities reflected unconditional values with respect to the initial starting times, 200 initial line starting times were simulated. Figure 4 gives the ridership probabilities. The expected waiting time at the origin node, expected waiting time at the transfer node, and the minimum expected total travel time to reach the destination for all three cases are presented in Table 3. The results obtained from the exact expressions are clearly comparable to those of the network simulation. Table 3 also indicates that the Spiess-Florian algorithm underestimates the expected waiting time at a node and consequently the minimum total travel time to the destination. The expected waiting time at the transfer node validates the assumption that transfer passengers behave similarly to random uniform passenger arrivals, particularly if the simulation is run for a large number of combinations of initial line starting times. The O-D demand rate of 500 persons per hour is considerably heavy; the time horizon of 3 hr is a representative value for a peak period. In the case of the exact expressions, the simulation results are only a validation of their correctness because the simulations were performed under the same assumptions as those behind the expressions.

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

The expressions used in transit assignment models for ridership probabilities and expected waiting time are often based on approximations and assumptions, such as Poisson line arrivals. However, real-world situations possibly conform better to assumptions of constant-line headways and random uniform passenger arrivals,

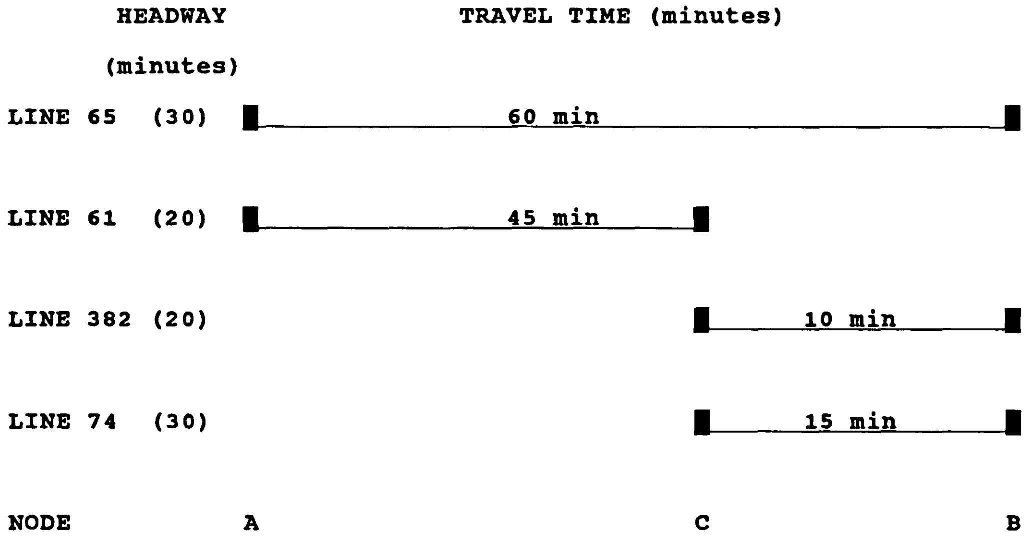


FIGURE 1 Selected lines from extracted network.

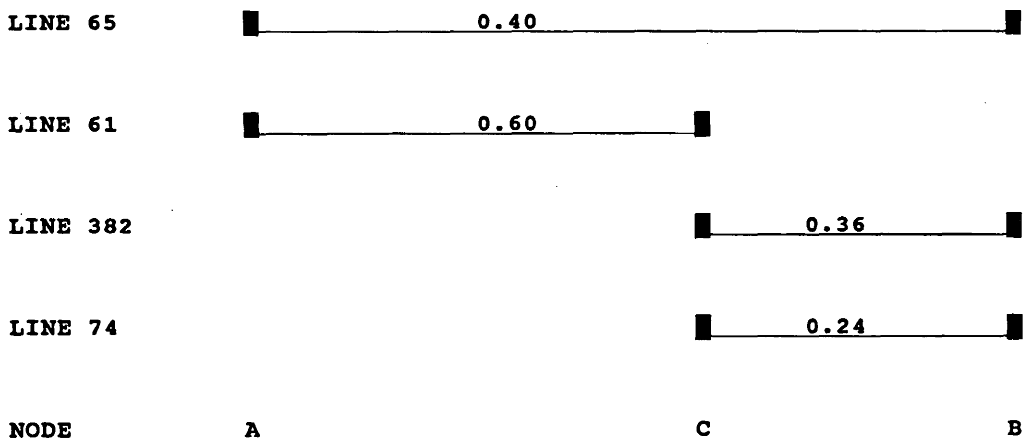


FIGURE 2 Ridership probabilities using Spiess-Florian algorithm.

which have been considered mathematically difficult. Exact expressions for these ridership probabilities and expected waiting times have been derived in this paper. The expressions are most applicable to assignment frameworks that enumerate the choice set of paths on the basis of travel times and expected waiting times, as well as to those that assign the ridership on the basis of line probabilities among a selected set of candidate lines (even if these are based on additional criteria besides travel and waiting times). The Spiess-Florian algorithm has both these characteristics, hence it is used as a benchmark to implement and compare the proposed expressions. The modified transit assignment model yields more robust values for line ridership than the original Spiess-Florian model in experiments based on a simulation of simple transit lines.

Research goals were not limited to developing the exact expressions in the algorithm; they also included the development of an efficient implementation procedure to minimize computational efforts. A recursive approach that avoids calculation of the expressions by brute force at each stage has been successfully implemented in the modified algorithm.

The assumption of uniform random passenger arrivals may not be justified in the presence of improved traveler information. A possible extension would be to develop a comparable model that considers the advantage of real-time information provided both at terminals and in-vehicle. This would result in a dynamic choice set of transit lines as a function of real-time information provided. Such models can then be used for the planning and evaluation of advanced public transit systems. This proposed extension to the model would essentially involve more complex strategies that change in real time, depending on the information provided to the user. The reliability of a transit service can be significantly

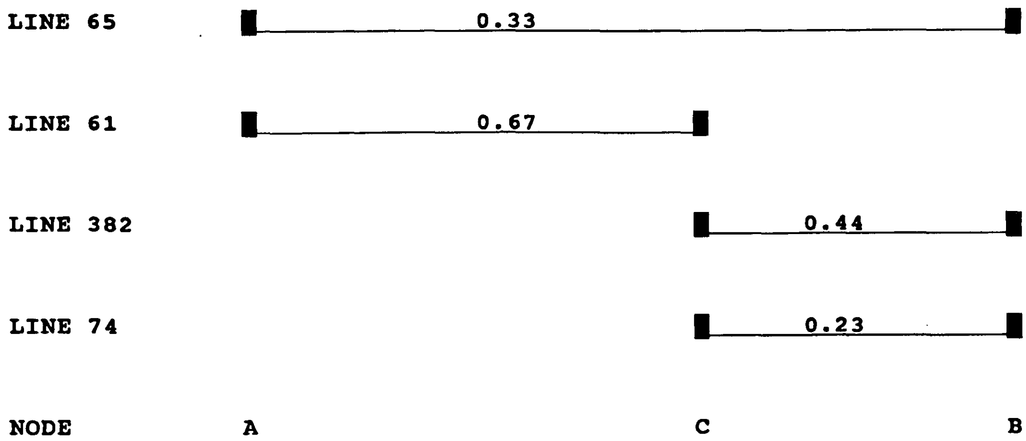


FIGURE 3 Ridership probabilities using exact expressions.

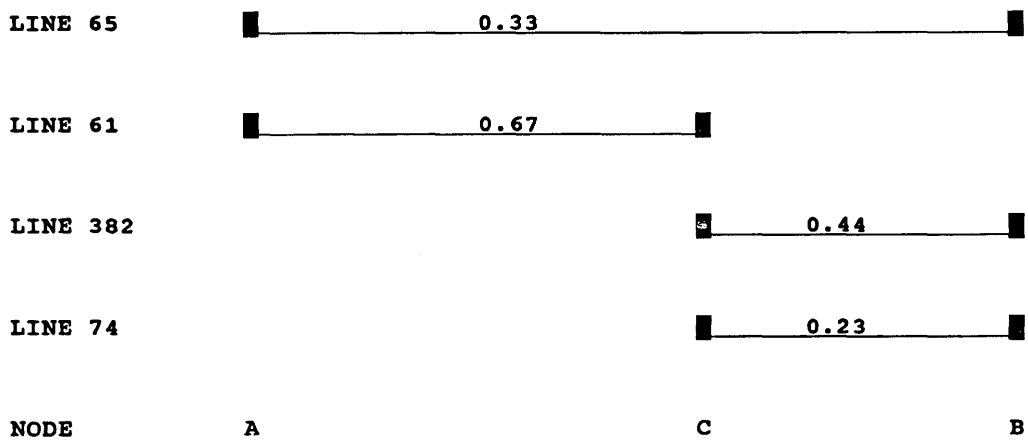


FIGURE 4 Ridership probabilities using network simulation.

improved if real-time information is provided; future research in this direction is needed.

The ability of the assumption that the passenger's choice set consists of multiple paths that give minimum expected travel times still needs to be examined via field validations. Data describing the nature of passenger arrival distributions are critical in evaluating the applicability of the proposed models to transit systems with real-time information availability.

ACKNOWLEDGMENTS

This paper was sponsored in part by the Transit Systems Research Program of the California Department of Transportation and by the University of California Transportation Center. Their support is gratefully acknowledged. The authors thank the referees for their suggestions, specifically with regard to some recent research literature on the topic.

TABLE 3 Comparison of Results

	Spiess-Florian Algorithm	Exact Expressions	Network Simulation
Minimum Expected Travel Time (min)	67.80	70.74	70.83
Expected Wait at Origin (min)	6.00	7.78	7.80
Expected Wait at Transfer (min)	6.00	7.78	7.83

REFERENCES

1. Spiess, H., and M. Florian. Optimal Strategies: A New Assignment Model for Transit Networks. *Transportation Research B,* Vol. 23B, No. 2, 1989, pp. 83–102.
2. Dial, R. Transit Pathfinder Algorithm. In *Highway Research Record 205,* HRB, National Research Council, Washington, D.C., 1967, pp. 67–85.
3. Le Clercq, F. A Public Transport Assignment Method. *Traffic Engineering and Control,* Vol. 14, No. 2, 1972, pp. 91–96.
4. Chriqui, C., and P. Robillard. Common Bus Lines. *Transportation Science,* Vol. 9, No. 1, 1975, pp. 115–121.
5. Horowitz, A. Extensions of Stochastic Multipath Trip Assignment to Transit Networks. In *Transportation Research Record 1108,* TRB, National Research Council, Washington, D.C., 1987, pp. 66–72.
6. De Cea, J., and J. Fernandez. Transit Assignment to Minimal Routes: An Efficient New Algorithm. *Traffic Engineering and Control,* Vol. 30, No. 10, 1989, pp. 491–494.
7. De Cea, J., and J. Fernandez. Transit Assignment for Congested Public Transport Systems: An Equilibrium Model. *Transportation Science,* Vol. 27, No. 2, 1993, pp. 133–147.
8. Jolliffe, J., and T. Hutchinson. A Behavioral Explanation of the Association Between Bus and Passenger Arrivals at a Bus Stop. *Transportation Science,* Vol. 9, No. 3, 1975, pp. 248–282.
9. Turnquist, M. A Model for Investigating the Effects of Service Frequency and Reliability on Bus Passenger Waiting Times. In *Transportation Research Record 663,* TRB, National Research Council, Washington, D.C., 1978, pp. 70–73.
10. Bowman, L., and M. Turnquist. Service Frequency, Schedule Reliability, and Passenger Wait Times at Transit Stops. *Transportation Research A,* Vol. 15A, No. 6, 1981, pp. 465–471.
11. Hall, R. Traveler Route Choice: Travel Time Implications of Improved Information and Adaptive Decisions. *Transportation Research A,* Vol. 17A, No. 3, 1983, pp. 201–214.
12. Marguier, P., and A. Ceder. Passenger Waiting Strategies for Overlapping Bus Routes. *Transportation Science,* Vol. 18, 1984, pp. 207–230.
13. Dial, R. A Probabilistic Multipath Transit Assignment Model Which Obviates Path Enumeration. *Transportation Research,* Vol. 5, 1971, pp. 83–111.
14. Jansson, K., and B. Ridderstolpe. A Method for the Route-Choice Problem in Public Transportation Systems. *Transportation Science,* Vol. 26, No. 3, 1992, pp. 246–251.

The contents reflect the views of the authors only. Publication of this paper sponsored by Committee on Transportation Supply Analysis.

Improving Efficiency of Commercial Vehicle Operations Using Real-Time Information: Potential Uses and Assignment Strategies

AMELIA C. REGAN, HANI S. MAHMASSANI, AND PATRICK JAILLET

Advances in communication, automatic vehicle location, and geographic information system technologies have made available several types of real-time information with benefits for commercial vehicle operations. Continuous updates on vehicle locations and demands create considerable potential for developing automated, real-time dispatching systems. The potential benefits of a diversion strategy in response to real-time information are explored under idealized conditions, and the technologies that are available for use in commercial vehicle operations and selected results derived from simulation are described. The results illustrate potential savings from simple diversion strategies under real-time information and highlight the need for methodological development to support improved truckload carrier operations decisions.

Telecommunications and information technologies provide unprecedented opportunities for using real-time information to enhance the productivity, performance, and energy efficiency of the commercial transportation sector. Achieving the benefits of real-time information requires development of fleet operating strategies, including vehicle assignment and dispatching rule with increased flexibility, along with suitable decision support methodologies. There appears to be virtually no methodology in the literature intended specifically for truckload or other surface carrier operations under the kind of real-time information possible with emerging technologies. The lack of methodological development applies to both the analysis of carrier operations to evaluate the effectiveness of real-time information and to actual tools that could be used by carriers to take advantage of such information. The area of vehicle routing and scheduling, including dynamic vehicle allocation and load assignment models, has evolved rapidly in the past few years, both in terms of underlying mathematical basis and actual commercial software tools (*1–3*). Although these approaches may well be adaptable to operations under real-time information availability, they are currently unable to take full advantage of such information because their underlying formulations do not recognize possible decisions that are meaningful only under real-time information. One such decision is the possibility to divert in response to customer demands.

After briefly describing some of the technologies available, this work identifies and explores potential uses of real-time information for the efficient management of truckload carrier operations. In particular an en route "diversion" strategy in response to unfolding customer demands is proposed and analyzed. A simulation model to explore the profitability of such diversion strategies under various operational conditions and demand arrival patterns is described and conditions under which such strategies might be profitable are derived. Findings suggest that meaningful potential exists for improving truckload carrier operations. These findings and related operational issues are discussed.

In truckload operations, carriers typically know only a portion of the loads that must be moved before the beginning of the day. Typically 60 percent of a given day's loads may be accepted on the same day that they are moved (*1*). The assignment of an available driver to a load therefore takes place almost in real time or at least shortly after the request is received. In addition, the load acceptance decision made by a carrier must be executed in real time and may have a significant impact on the carrier's ability to accept other loads later in the day or in the days that follow. This research explores ways to make "good" assignment decisions, and ultimately load acceptance decisions, that lead to overall cost-effective operations but rely on local (current) rather than on long-term or forecasted information. Although various forecasting methods may be used to estimate future demands, this information is not reliable in practice because of the large number of possible origin and destination combinations and the inherent randomness of the process (*1*).

INTRODUCTION TO TECHNOLOGIES

Automatic vehicle location (AVL) systems are finding increasing application in a variety of contexts, including truckload and less than truckload trucking companies; local delivery and courier services; fire, rescue, and police departments; utility companies; security companies; public transportation companies; high-value and hazardous materials shippers; and taxi and limousine services. Not all applications require the same degree of accuracy. The dispatcher in a long-distance trucking company will most likely derive the same benefit from knowing the locations of the company vehicles to within 50, 1000, or even 10 000 m, whereas a police dispatcher may need to determine, with certainty, on which streets tracked vehicles are located.

Although global positioning satellite (GPS) technology is often perceived as the leading AVL technology, the companies with a major market share in the long-haul trucking AVL market do not employ GPS technology in their standard products. One of these systems uses a group of nationwide specialized mobile radio towers with optional Loran-C location tracking, and the other uses a

A. C. Regan and H. S. Mahmassani, Department of Civil Engineering, University of Texas, Austin, Tex. 78712. P. Jaillet, Department of Management Science and Information Systems, University of Texas, Austin, Tex. 78712.

network of two geosynchronous satellites to perform tracking and communication, with Loran-C an optional addition. The 500- to 1000-m accuracy these systems provide is adequate for dispatchers to estimate which highways the trucks are on. Commercial applications that require street-level location information could benefit from increased accuracy. Applications that include navigation, either on board or at a central location, require more accurate position information than 500- to 1000-m estimates, as these do not ensure street-level accuracy.

In most AVL applications, the position location obtained must be transmitted to the dispatch center over an available communication link. Although position estimates and even point-to-point routing could, with an appropriate microcomputer, be calculated on board the vehicle, the vehicle's position must be transmitted back to the dispatch center for display. Communication links available for this purpose differ in cost and sophistication. VHF, cellular, or subtitle link may all be used, with digital cellular becoming more and more widely available. The link used typically depends on the frequency of communication and the distance between the dispatch center and the vehicles. The communication cost of such a system may be high, with messages costing as much as ten cents per brief packet for satellite communication systems, about five cents per packet for transmission over 800 or 900 MHz trunked radio lines, and more for standard cellular in which rates are determined by the minute rather than the data packet. If the vehicle locations are 'polled' often by the central dispatcher, these costs add up quickly (*4*).

This study is most interested in irregular route common carrier operations. Discussions with operators of trucking companies of various sizes have made clear that whatever the particular technologies chosen, AVL and two-way communication systems will be necessary for many trucking companies to compete in a market where the location and magnitude of demands for service are highly dynamic (*5*). A 1992 survey performed at the University of Texas of just under 300 carrier companies pointed out the fact that carriers agree that AVL and two-way communications technologies will lead to improvements in many aspects of their operations. Figures 1 and 2 share some of their responses about what they perceive as the potential benefits of these technologies. Although it is clear that these technologies are beginning to see widespread use, it is equally clear that the full potential of these technologies will not be realized until responsive real-time dispatching tools become available.

REAL-TIME ASSIGNMENT STRATEGY: DIVERSION

Because of the length of some empty moves made to pick up loads, it is possible that new information on demands to be serviced may arrive while a driver is en route to a pickup. Assuming that time windows for movements are flexible, this new demand information may be used to order demands in such a way as to reduce empty distances driven. Quasi-continuous dispatcher-to-driver communication makes it possible to divert a driver en route to a pickup location to an alternative load, thereby inducing a resequencing or reassignment of the original load. Such diversion strategies are not generally feasible under current operations because dispatcher-driver communication takes place at discrete instances only, typically at a load pickup or delivery point (*5*).

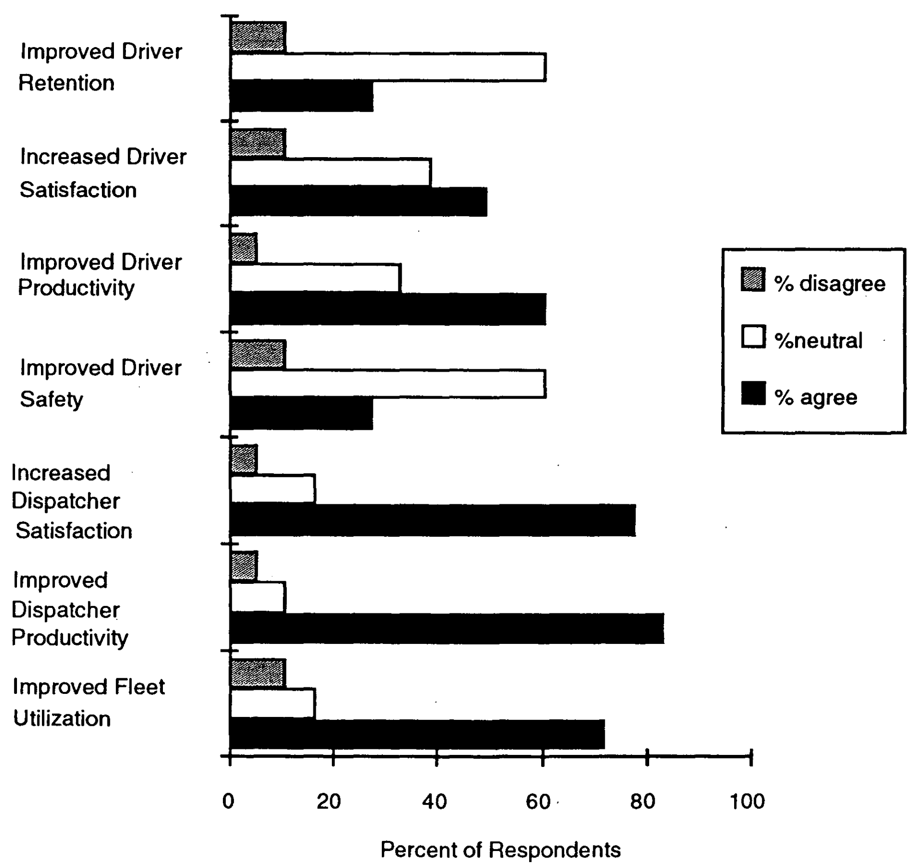


FIGURE 1 User assessment of two-way communication and AVL system benefits: Part 1.

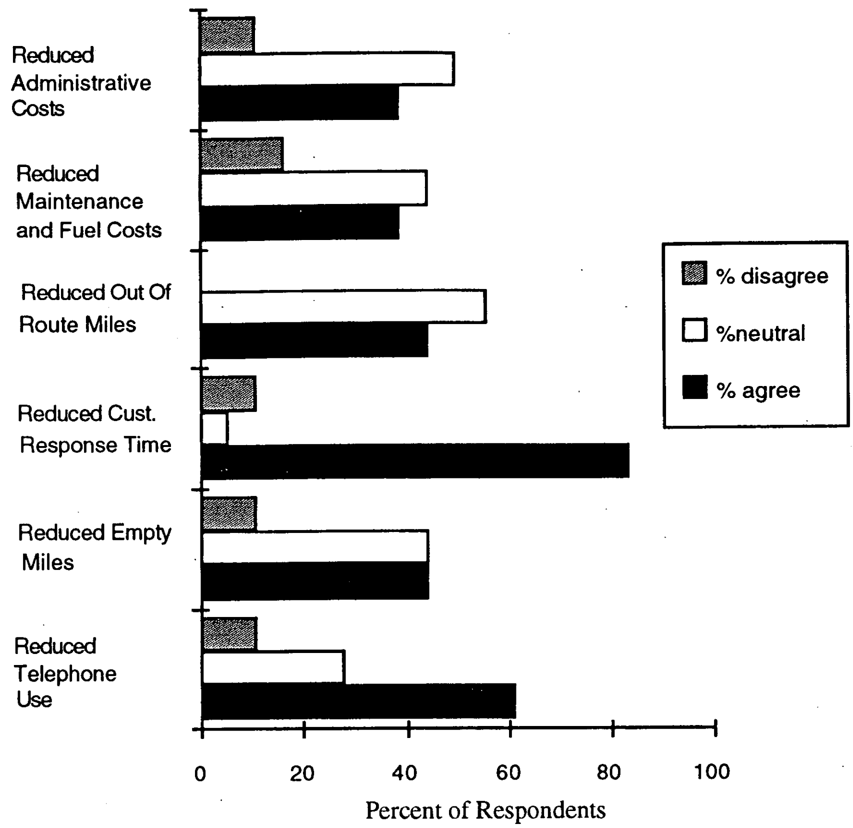


FIGURE 2 User assessment of two-way communication and AVL system benefits: Part 2.

The relative improvement possible under this strategy depends on the relative locations of the alternative pickup and delivery points. Under some distributional assumptions about the locations of these points, the interest is in the probability that diverting the driver to a new demand while en route to a previously assigned pickup will be beneficial. In even the simplest case, it is difficult to derive this probability analytically because the various cost components are not independent. For this reason, these probabilities and various other performance measures are evaluated through simulation of such diversion strategies over service horizons of varying lengths, under different arrival stream distributions, and under load acceptance rules that either require all loads to be serviced or allow less profitable loads to be rejected. The scenarios examined up to this point are not intended to exactly replicate actual operating conditions, but to provide a simplified representation that allows derivation of basic insights into the potential benefits of real-time information and the factors that affect these benefits, as well as the identification and design of strategies that merit examination under more realistic operating conditions.

Diversion Probabilities Under Simple Assumptions

To begin with the most basic case, while a driver is en route to a load origin, information about another load (and in this initial case, only one other load) to be moved becomes available. Answers to the following questions are desired: What is the probability, given various diversion decision rules, that the driver will be diverted to serve the new load first? What is the probability that following such diversion decision rules will result in a reduction of overall distance traveled? And, what is the associated expected reduction in travel?

To clarify, consider in Figure 3, a vehicle that begins at the center, *c*, of a circle and moves toward the origin of a loaded movement between Points X_1 and X_2, where these points are uniformly and randomly generated over the area of the circle. Given a diversion point (the point at which another load to be moved becomes available) some fraction of the distance from the center of the circle and origin X_1, the probability is derived that the distance between the diversion point to a new origin X_3 will be less than the distance from the diversion point to origin X_1. Let α, $0 \leq \alpha \leq 1$ denote the fraction of the distance from the center to X_1 traveled to reach the diversion point. The probability that the distance from the

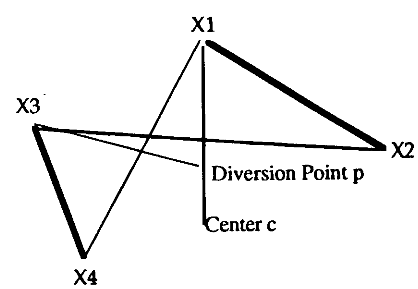


FIGURE 3 Diversion example.

diversion point to the new origin is less than that to the old origin is given by $(1 - \alpha)^2/2)$, as shown hereafter.

Let $B(c, r)$ denote the circle of center c and radius r, and $d(x, y)$ the Euclidean distance between points x and y. Consider two random points in $B(c, r)$, say, X_1 and X_3. For $0 \leq \alpha \leq 1$, let $W_1(\alpha)$ be the point on the segment (c, X_1) such that $d\,[c, W_1\,(\alpha)] = \alpha\,[d\,(c, X_1)]$. Define the following two random variables $Y_1 = d\,[W_1(\alpha), X_1]$ and $Y_2 = d\,[W_1\,(\alpha), X_3]$, where Y_1 and Y_2 represent the distances from the potential diversion point to the current and potential load origins.

Let Z be the radial distance of $W_1(\alpha)$ so $Z = d\,[c, W_1\,(\alpha)]$, and $f_z\,(\cdot)$ be its probability density function:

$$P(Y_2 > Y_1) = \int^{P} (Y_2 < Y_1 \mid Z = z) f_z\,(z)\,dz$$

$$= \int^{P} \{X_3 \in B\,[W_1\,(\alpha), z/\alpha - z]\} f_z\,(z)\,dz \qquad (1)$$

Because $W_1(\alpha)$ is a random point in $B(c, \alpha)$,

$$P(Y_2 < Y_1) = \int_0^{\alpha} [(1 - \alpha)\,z/\alpha]^2\,(2z/\alpha^2)\,dz = (1 - \alpha)^2/2 \qquad (2)$$

If a myopic strategy of diverting to the new demand origin, X_3, is followed if it is closer to the diversion point than origin X_1, then $(1 - \alpha)^2/2$ represents the fraction of loads for which one actually diverts. This probability $[P(Y_2 < Y_1)]$ is shown graphically as a function of the diversion point location parameter, α, in Figure 4. However, under this strategy, even if the diversion decision at point $\alpha = 0$ is evaluated, the resulting average savings in terms of reduced distance traveled while serving the two loads is less than 1 percent, and diverting at points further downstream actually results in a slight increase in traveled distance, on average.

A more plausible diversion strategy would also consider the relative distances between the destination point of the first movement and the origin point of the next load. In Figure 3 these are given by $d(X_2, X_3)$ and $d(X_4, X_1)$. In this case diversion is chosen if

$$d(p, X_3 + d(X_4, X_1) < d(p, X_1) + d(X_2, X_3) \qquad (3)$$

Analytic derivation of the corresponding diversion probability under this strategy is no longer straightforward because the respective distances are not independent. The diversion likelihood and associated expected benefit are evaluated using a simulation program under the following underlying assumptions:

- A circular work area with a radius of 1 unit of travel as in Figure 3,
- Uniformly and independently generated demand locations,
- Euclidean (straight-line) travel distances, and,
- Diversion results compared with a 'base' case where demands are serviced in order of their arrival.

Simulation was used to evaluate the case shown in Figure 3, in which the vehicle begins at the center of the circle, c, and only two demands are served. A total of 75,000 independent trials were executed for each of 10 values of α, the diversion point fraction, that varied between 0 and 1. When $\alpha = 0$, that is, resequencing occurs before departing for the first demand, as should be expected, resequencing occurs in half the cases. The fact that resequencing occurs in more than 10 percent of the cases when the diversion decision is evaluated at the origin point of the first load, that is, $\alpha = 1$, is somewhat counterintuitive and results from the cases in which the loaded movement of the candidate load takes the vehicle close to the origin of the original load. Average savings resulting from such a diversion strategy, in which the demand horizon (the number of demands served in a single simulation instance) is only two loads, vary from 7 percent of the total distance traveled if demands are taken in order with $\alpha = 0$, down to 1 percent with $\alpha = 1$. These results are shown in Figure 4.

This analysis is extended beyond the first diversion decision. After serving the load selected, the vehicle begins to move toward the unsatisfied demand. Again, a new demand arises along the way, creating a new diversion opportunity. With a demand horizon of 100 loaded movements, evaluated sequentially on a pairwise basis, simulation results indicate overall benefits (of the diversion strategy) in the range of 1.5 to 12.5 percent of overall distance traveled, depending on the diversion point fraction, relative to the base case of servicing demands in the order in which they arrive. Figure 5

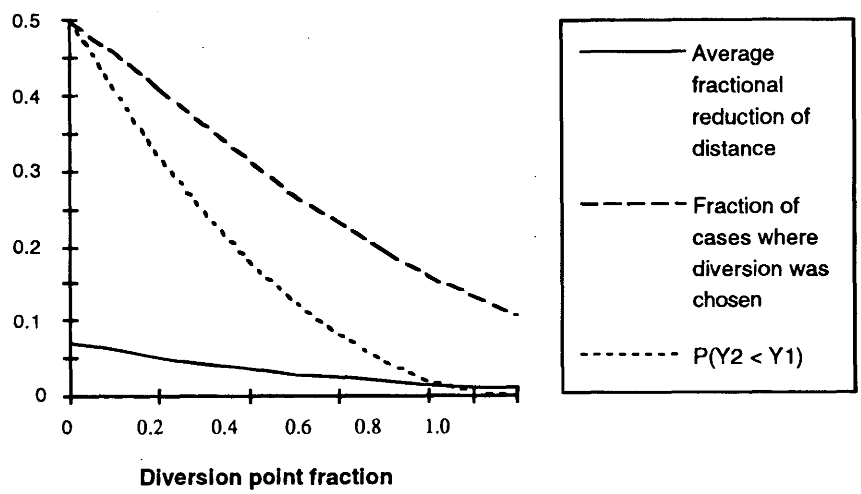


FIGURE 4 Probability of diversion when distances to load origins and between two demand points are considered.

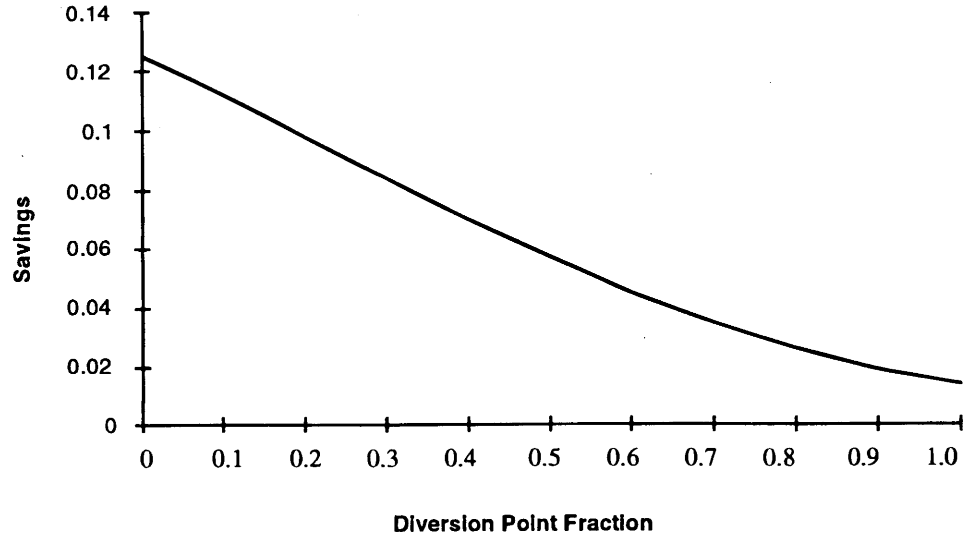


FIGURE 5 Reduction of travel with diversion strategy as fraction of overall travel with 100 demands served.

shows the overall reduction in travel cost as the diversion point fraction is varied from 0 to 1. A regression model that assumes an exponential functional form approximates this curve as

$$\text{Reduction} = (0.151)0.111^{\alpha} \tag{4}$$

Note that in the simulations the new load is assumed to be known by the time the vehicle reaches the diversion decision point. If α, the diversion point fraction, is a uniform random variable taken between 0 and 1, which would correspond to a scenario in which the new load may become known at any point along the route between the vehicle's last load destination and the next load origin, then the average reduction in travel is more than 6 percent of the total distance traveled.

In addition to these average numbers, it is important to gain insight into the worst-case performance of this strategy. If the service horizon is short, say, less than 10 demands served, it is possible to make one or more diversion decisions that result in an overall cost (distance) increase over the demand horizon. However, over a longer service horizon, diversion outperforms the base case more than 99 percent of the cases. Figure 6 gives the expected probability of overall gain and loss, respectively, along with the associated magnitudes of the gains and losses over different demand horizons. Each set of numbers is based on 10,000 simulated realizations of the corresponding sequence of random demand locations. Of course, overall gain here corresponds to the sequence, not to individual diversion decisions. The expected gains (or losses) are given in terms of fractions of the overall distance traveled under the base case (no diversions). The reported decreases (and increases) are conditional values given that the particular sequences experienced a decrease (or increase) under the diversion rule. The expected overall gain (or loss) over a sequence of calls is given by

$$\begin{aligned} \text{E[gain]} = {} & \text{E[gain} \mid \text{gain} > 0]\, p(\text{gain} < 0) \\ & - E[\text{gain} \mid \text{gain} < 0]\, p(\text{gain} < 0) \end{aligned} \tag{5}$$

The results in Figure 6 indicate that even with a horizon with as few as 10 diversion points the probability of overall loss is only 11.6 percent, with a corresponding expected loss of 2.3 percent of the overall distance under the base case. On the other hand, the 84.5 percent likelihood of gain is accompanied by an expected gain greater than three times the expected loss (7.1 percent reduction in overall cost). Note that for the 10-demand case the likelihood of 0 gain (no diversions chosen at all) is about 3.9 percent. The probability of loss rapidly decreases with the service horizon considered, to less than 1 percent with 50 demands (accompanied by an insignificant loss of under 1 percent, whereas the corresponding gain is over 6 percent with over 99 percent probability of gain. The fact that the probability of loss and the expected conditional loss are extremely small makes the diversion operating strategy appear to be somewhat of a win-win strategy. The simple diversion criterion of comparing the relative distances to serve a pair of loads sequentially appears relatively robust. If it suggests that diversion is profitable diversion is done, with a very high probability of realizing some meaningful benefit, and if it does not, the original plan is followed.

To further reduce the likelihood of loss over a finite number of decisions one can introduce a threshold in the diversion rule, whereby the local gain is required to exceed some minimum level to trigger a diversion, as follows:

If

$$\begin{aligned} & d(p, X_3) + d(X_4, X_1) < d(p, X_1) + d(X_2, X_3) \\ & - T[d(p, X_1) + d(X_2, X_3)] \end{aligned} \tag{6}$$

then

divert and serve load X_3 to X_4 first,

where

p = current diversion point,
X_1, X_2 = origin and destination locations of current load,
X_3, X_4 = origin and destination of newly arrived load, and
T = threshold multiplier corresponding to the minimum relative improvement associated with a given diversion.

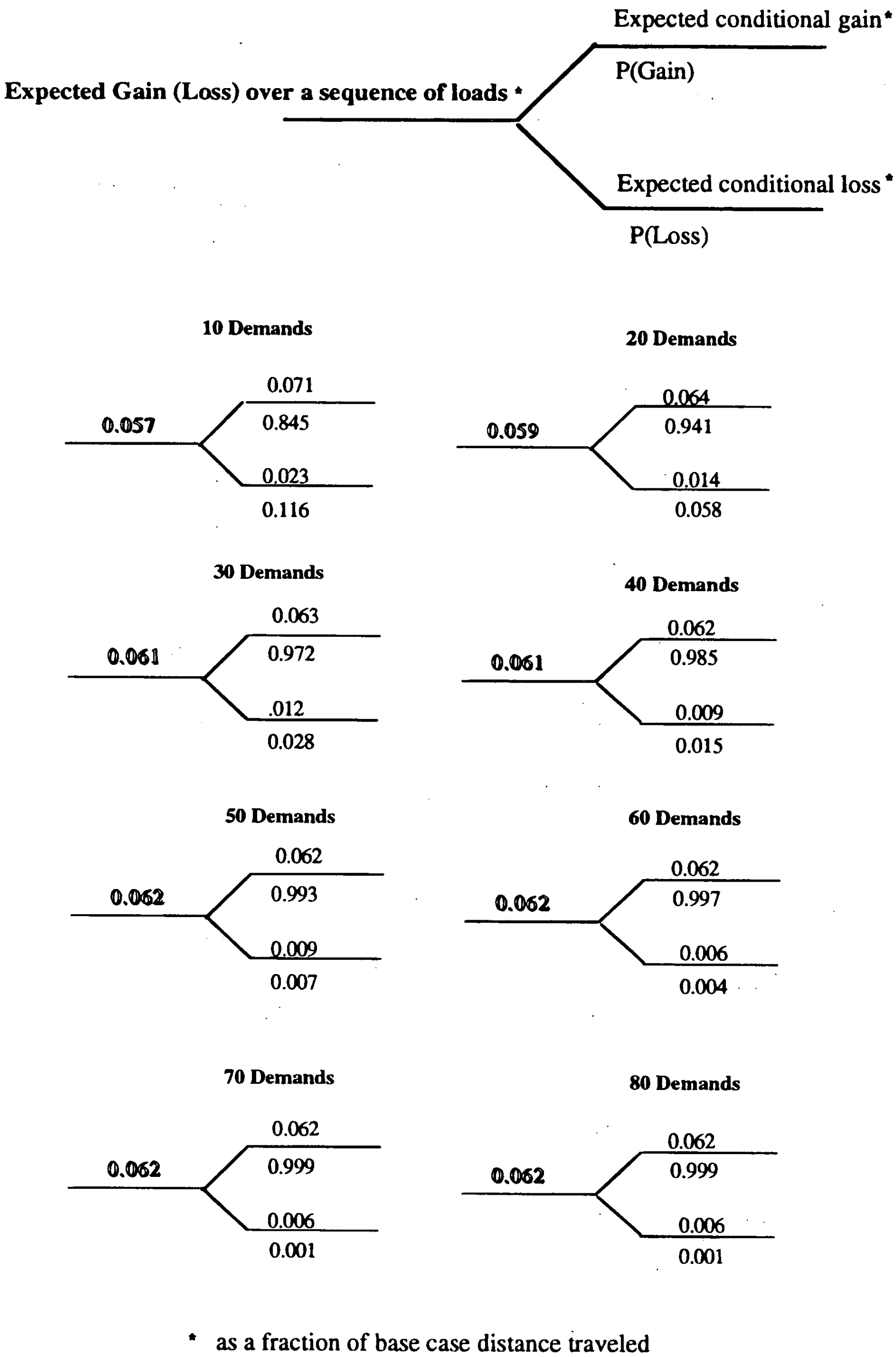


FIGURE 6 Benefits of diversion (distances to serve both loads considered).

This multiplier was varied from 0 to 0.5 (50 percent in the present analysis. Results suggest that a threshold value of about 10 percent of the base case cost yields the best performance. However, although the addition of a threshold for diversions reduces the risk of bad diversions, any threshold that precludes many positive diversions results in a reduction of expected benefits. The addition of the threshold rule for diversion increased the overall benefits by an amount between 1 and 0.3 percent, depending on the demand threshold. However, more significantly, the thresholds cut nearly in half the already low probability of loss in each case.

In these simulations, the diversion point fraction, α, varied uniformly between 0 and 1. The diversion decision was based on an entire sequence of moves associated with the current and new demand points. Alternatively, Figure 7 shows the results of a diversion strategy under a strictly myopic or greedy strategy of diverting to the closest origin point, that is, if $d(p, X_3) < d(p, X_1)$ then divert to load X_3.

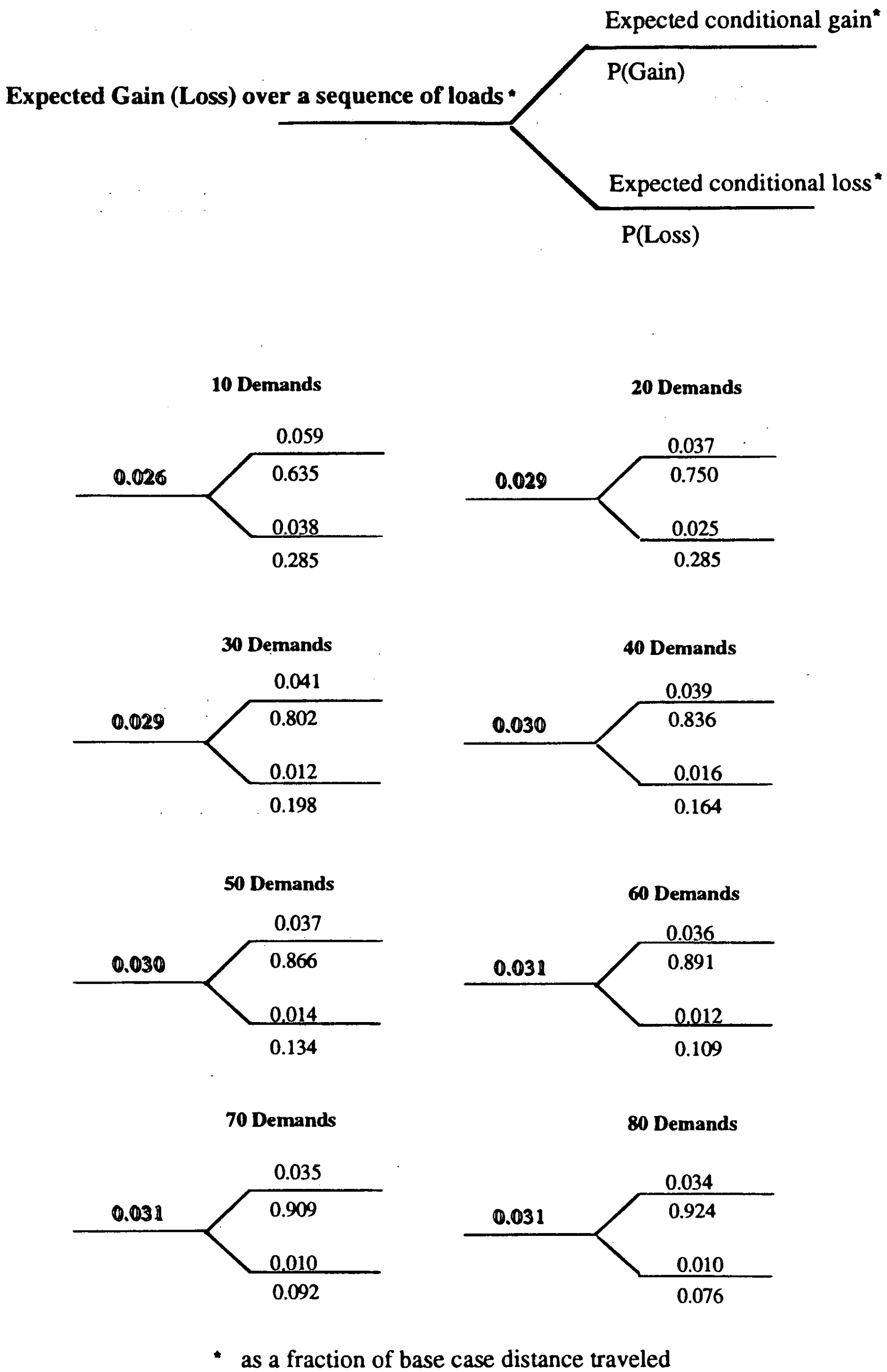


FIGURE 7 Benefits of purely greedy diversion strategy.

Two interesting points can be noted about these two sets of results in Figures 6 and 7. The first is that despite the limited information employed, that of which origin is closer, the greedy diversion strategy consistently leads to a reduction in overall expected travel over the demand horizons considered. The second is that by slightly increasing the amount of information considered, by also considering the distances that must be traveled empty between the two loads, the overall benefits double from 3.1 percent of the distance traveled to 6.2 percent over a demand horizon of 60 points. More importantly, the risk of overall loss is reduced significantly. For example, the probability that the diversion strategy will result in an overall cost increase with 60 demands is reduced from 0.109 (in the greedy case) to 0.004 when the distances to the load destinations are also considered. This is nearly 2 orders of magnitude less. Overall, these results demonstrate the potential power of reacting to even small amounts of real-time information on the state of the system.

EXTENSIONS OF SINGLE-VEHICLE, TWO-DEMAND CASE

This exploration under highly idealized conditions suggests that even a simple local diversion strategy is highly likely to result in a reduction of overall distance traveled. After considering only two demands at a time, the analysis is extended to consider several demands in a particular decision to divert and to look at demands that are uniformly generated in space but arrive according to a Poisson arrival stream as well as from a uniform distribution. In addition, operational constraints in which every demand must be served, and those in which one has the freedom to accept or reject demands according to the cost of serving them have been explored (*6*).

Naturally the performance of a given diversion strategy can be compared relative to several possible benchmarks or base cases with differing results. In addition to the base case of serving demands in the order in which they arrive, an "intelligent base" case in this analysis, is also considered next.

Poisson Arrival Stream, Optimal Resequencing, All Demands Accepted

This scenario has the following assumptions:

- Demands are generated from a Poisson arrival stream over time,
- The rate of arrival is rapid enough that more than one new demand may arrive while the vehicle is en route to a pick-up, and
- Demands diverted away from or not chosen for diversion are added to a queue and resequenced optimally with respect to overall distance traveled before being served.

This diversion strategy is compared with two different base cases. The first assumes service in the exact order of arrival, as considered previously, whereas the second, an intelligent base case, assumes that any demands waiting for service are resequenced optimally at the completion of each loaded movement. This intelligent base strategy is applied with optimal resequencing of up to five demands at a time. This itself results in solutions that are only 1 to 2 percent higher than those attained under the comparable diversion strategy in terms of overall travel distance. Under the assumption that all demands must be served, with demands generated from a Poisson arrival stream and with an arrival rate rapid enough to produce diversion opportunities, this intelligent base scenario leads to savings of more than 12 percent of the base case travel distance, and the en route diversion strategy tends to improve on the intelligent base by about 1 to 2 percent.

The dashed lines in Figure 8 show the various distances (costs) compared when choosing to divert, resequence, or serve the demands as they arrived in the case where two demands are in the queue and while the driver is en route to the current demand, origin X_1. The strategy chooses the minimum cost case of the $n!$ alternative orderings, where n is the number of demands in queue. It is assumed that n is a small number, say, less than 6, since to enumerate all alternatives for even slightly larger queues would take a prohibitive amount of time. This assumption of short queues makes sense in the trucking application where drivers typically have one or two jobs queued at most. Using the notation in Figure 8, when a new demand arises the minimum of the following six quantities corresponding to all possible service sequences is evaluated:

$$a\colon d(p, X_1) + d(X_2, X_3) + d(X_4, X_5)$$

$$b\colon d(p, X_1) + d(X_2, X_5) + d(X_6, X_3)$$

$$c\colon d(p, X_3) + d(X_4, X_5) + d(X_6, X_1)$$

$$d\colon d(p, X_3) + d(X_4, X_1) + d(X_2, X_5)$$

$$e\colon d(p, X_5) + d(X_6, X_1) + d(X_2, X_3)$$

$$f\colon d(p, X_5) + d(X_6, X_3) + d(X_4, X_1) \qquad (7)$$

Cases e and f represent a diversion to the new load; Cases c and d represent diversion to a load already in the queue, one that was previously found unprofitable to divert to or to place first in the queue but was resequenced because of the information on the new demand. Cases a and b represent no-diversion cases, that is, the vehicle proceeds as before, but the second and third loads may be resequenced if that is beneficial.

Various extensions of these rules were explored, and it seems that under the assumption that eventually all demands must be serviced, a strategy that allows diversion but limits the number of times that one diverts before some demand is serviced is better than one that allows diversion whenever it is locally better. If diversion is allowed whenever it appears (locally) beneficial, under these assumptions costs may be low early in the service horizon and then considerably higher at the end.

Investigation of Poisson Arrival Stream, Optimal Resequencing, Loads Accepted or Rejected on Basis of Cost to Existing Route

This investigation has a different assumption from the preceding case with respect to load acceptance, namely, many demands are generated over time and loads may be accepted or rejected. This case assumes a rapid arrival rate for new demands. Whenever a new demand becomes known and space is available in the queue adding it to the current queue is considered. Rather than inserting the new demand into the existing route, resequencing the route in light of the new demand is considered. Because to optimally resequence the whole route would be computationally expensive (and possibly infeasible) the marginal cost of adding a demand to one of the first five slots in the queue is determined. If the additional empty distance needed to service the first four demands along with the new demand exceeds a given threshold value the new load is rejected. Otherwise the load takes an empty slot in the queue. These (up to) five demands are then resequenced optimally. Under the diversion strategy the load acceptance or rejection decision is made as soon as the demand becomes known if the vehicle is moving empty and as soon as it becomes empty if it is moving loaded. In the intelligent base case load acceptance decisions are made for all loads that have become known during the last period of service immediately after service is complete. These loads are evaluated for acceptance or rejection using the same logic as that in the diversion case (marginal cost to add to the first five slots in the queue) in the order in which they arrived. It was an a priori thought that the diversion strategy would perform well under these circumstances. However, it appears that excessive diversion creates a sort of "zig-zag" effect where a

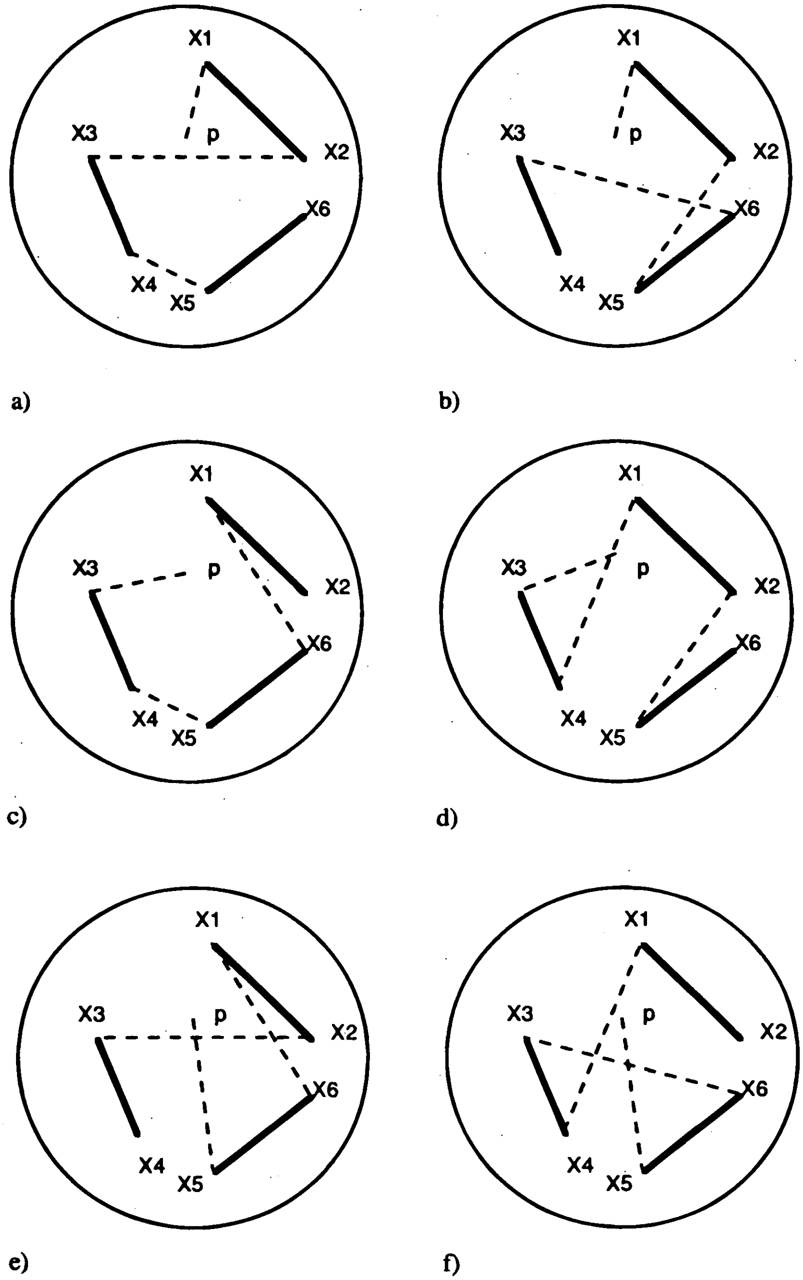


FIGURE 8 Alternatives with two queued demands and a third arrival: *a*, no change in plan; *b*, no diversion, resequence; *c*, divert to previously considered demand; *d*, divert to previously considered demand; *e*, divert to new demand; *f*, divert to new demand. (Dashed lines represent empty movements, and solid lines, loaded movements.)

vehicle is en route and then diverts and then diverts again. It appears that without additional constraints to restrict the amount of diversion a comparable intelligent base case in which the first few queued demands are optimally resequenced performs better. Table 1 provides a summary of the results presented in the last few sections.

CONCLUSIONS

Trucking operations consume a vast quantity of economic and environmental resources. A reduction in overall travel of even a few percentage points would represent a significant savings to both suppliers and consumers of trucking services. The U.S. Department of Transportation estimated that in 1991 motor vehicle fuel purchases accounted for 7.9 percent of common carrier costs or about $8.7 billion nationally (*7*). If 10 percent of these vehicles had a 5 percent reduction in fuel consumption, $43.5 million would be saved each year. In this work we identify and explore potential uses of real-time information for the efficient management of truckload carrier operations. Findings suggest that the diversion strategy examined may result in reduced travel distances and hence improved efficiency under certain conditions. Such strategies could become one part of an overall assignment and load acceptance strategy for truckload operations. The exploration of idealized scenarios suggests that a reduction of overall travel distance of between 5 and 10 percent would not be unreasonable. Although they are not intended to exactly replicate actual operating conditions, these scenarios do provide a simplified representation that allows the extraction of basic insights into the potential benefits of real-time information, the factors that affect their benefits, and the identification and design of strategies that merit examination under more realistic operating conditions.

Continuing developments include extending this analysis to a more 'realistic' scenario with respect to geographic region studied and customer demand stream. In addition, a fleet of vehicles rather than a single truck is examined. As the availability of automatic vehicle location and two-way communication technologies improves, and as the cost of equipping vehicles with these technologies decreases, more and more fleets will incorporate these technologies into their daily operations. Recent interviews with carrier company executives and fleet managers suggest that they are eager to incorporate communications technologies and optimization tools into their operations but that much work remains to be done in terms of developing such tools in a manner that is responsive to actual operating realities. An additional benefit of such tools is that they would enable companies to find good solutions to a complicated multiobjective problem. An addition to the goal of reducing overall distance driven is that of matching a driver with the load that best meets his or her needs. Discussions with industry executives have pointed out that irregular route truck drivers may stay on the road for more than 3 weeks at a time before finding an opportunity to pull a load in the direction of their home base. Flexible assignment strategies would improve the chances of finding a load that meets the preferences of an individual driver. A data base management system that would likely be an adjunct to any computer-based dispatching system could make the preferences of individual drivers easily accessible. It is clear that there are many potential uses of new technologies in commercial vehicle operations in general, and freight carrier operations in particular. Although technologies are beginning to see widespread use, it is equally clear that their full potential will not be realized until responsive real-time dispatching tools become available.

TABLE 1 Summary of Key Results

Scenario	Results
Demands generated a fraction of the way towards the first demand, 2 demands evaluated in each simulation.	Savings of 1-7 percent of the base travel distance depending upon the diversion point fraction.
Demands generated a fraction of the way towards the current demand, 100 demands evaluated in each simulation.	Savings of 1.5-12.5 percent of the base travel distance depending upon the diversion point fraction.
Demands generated from a Poisson arrival stream, all demands served, optimal resequencing of up to first five demands, diversion to new or queued demand.	Savings of 13-14 percent base case travel distance when compared to the base base, 1-2 percent of the intelligent base case travel distance when compared to the intelligent base case.
Demands generated from a Poisson arrival stream, accepted or rejected for service based upon space in queue and the cost of providing service given the current queued demands, optimal resequencing of up to first five demands, diversion to new or queued demand.	No comparison to the base case, little or no savings when compared to intelligent base case because of zig-zag effect.

Notes: The base case refers to serving the demands in the order in which they arrive. The 'intelligent base' refers to the case in which en-route diversion is not allowed, but a short queue of demands is resequenced optimally prior to the start of new service.

REFERENCES

1. Powell, W. B., A Comparative Review of Alternative Algorithms for the Dynamic Vehicle Allocation Problem. In *Vehicle Routing: Methods and Studies* (B. L. Golden and A. A. Assad, eds.). Elsevier Science Publishers, Amsterdam, The Netherlands, 1988, pp. 249–291.
2. Golden, B. L., and A. Assad. Perspectives in Vehicle Routing: Exciting New Developments. *Operations Research,* Vol. 34, No. 5, Nov.–Dec. 1986, pp. 803–810.
3. Dejax, P. J., and T. G. Crainic. A Review of Empty Flows and Fleet Management Models in Freight Transportation. *Transportation Science,* Vol. 21, No. 4, Nov. 1987, pp. 227–249.
4. French, R. L. Introduction to Intelligent Vehicle Highway Systems (IVHS). Presented at VNIS '91 International Conference on Vehicle Navigation and Information Systems, Dearborn, Mich., Oct. 1991.
5. Regan, A. C. Emerging Technologies in Automatic Vehicle Location and Two-Way Communication Systems for Common Carrier Operations. Presented at 1993 International Motor Carrier Symposium, Williamsburg, Va., May 1993.
6. Regan, A. C., and H. S. Mahmassani. *Evaluating Real-Time Dispatching Strategies for Truckload Trucking Operations: Simulation Model Results.* Working paper. Department of Civil Engineering, University of Texas, Austin, July 1994.
7. Schmitt, R. R., and T. D. Feinberg (eds.). *Transportation Statistics Annual Report 1994.* U.S. Department of Transportation, 1994.

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

Quasi-Continuous Dynamic Traffic Assignment Model

BRUCE N. JANSON AND JUAN ROBLES

Several variants of combined dynamic travel models in discrete time with dynamic user equilibrium or system optimality as the assignment objective have been presented recently. This modeling approach is converted into quasi-continuous time, which enables two key model improvements: (*a*) traffic volumes are spread over time intervals in continuous time, allowing trips to be split among successive time intervals, and (*b*) the first-in first-out ordering of trips between all zone pairs is more precisely maintained. The means by which capacity losses are approximated on upstream links caused by spillback queueing from oversaturated links and accidents are also described. Trips are assumed to have scheduled departure times and variable arrival times, but notational variations allowing other model forms are briefly mentioned. Application of this model to a Denver-area network with comparison of results to observed speeds and volumes is described elsewhere.

Among several key issues addressed by researchers developing dynamic traffic models over the past 20 years, three critical issues are capabilities of the model to (*a*) fractionally split trip flows among multiple paths at any road juncture, (*b*) validly maintain the first-in first-out (FIFO) ordering of trips between all zones pairs, and (*c*) account for spillback queueing effects caused by incidents and oversaturated links. This paper formalizes the inclusion of these three concerns in a dynamic user-equilibrium (DUE) formulation. Robles and Janson (*1*) apply this model to a Denver-area network covering about 100 mi^2(260 km^2) with comparisons of results with observed speeds and volumes.

DUE formulations presented by Janson (*2,3*) use 0-1 variables called node time intervals to track trips across the network in both time and space (i.e., to identify whether trips departing zone *r* in time interval *d* cross node *i* in interval *t*). However, trips departing within a given time interval from any node or zone do not all depart at a single point in time but instead depart as a uniform trip rate over that time interval. This paper improves these earlier DUE formulations by representing time more continuously. Herein, the integer node time intervals are used to compute "trip flow fractions" (i.e., the fraction of trips departing zone *r* in time interval *d* to cross node *i* in time interval *t*). All trip flows are tracked through the network in continuous rather than discrete time. This modeling revision improves its (*a*) spillback queueing effects and dynamic link capacity adjustments, (*b*) FIFO trip ordering between all origin-destination (O-D) pairs, and (*c*) link volume and speed transitions between time intervals.

In DUE, the full analysis period of several hours is sliced into shorter intervals. Each trip has a known departure or arrival time (but not both) and its corresponding trip-end zones. Because travel times are variable, times of departure and arrival cannot both be fixed for any trip. DUE1 assumes known departure times for trips from each zone but only total arrivals to each zone over the full analysis period. DUE2 assumes known arrival times of trips to each zone, but only total departures from each zone over the full analysis period. Janson (*4*) formulates DUE2 similarly to DUE1, except for numerous notational changes to make node time intervals and node-to-destination travel times on the basis of destination zones and arrival times. Janson (*5*) describes a general model including both trip types, and Janson and Robles (*6*) combined route choice with departure or arrival time choice in DUEA.

DUE1 as formulated in this paper with scheduled departures is defined as follows:

> Given a set of zone-to-zone trip tables containing the number of vehicle trips departing from each origin zone in successive time intervals of 1 to 10 min each, and also the destination zone but not the arrival time of each trip, determine the volume of vehicles on each link in each time interval such that, for each O-D pair of zones, no path has a lower travel time than any used path for trips departing within a given time interval.

This DUE condition for fixed departure times, as derived by Janson (*3*), is a temporal generalization of Wardrop's (*7*) condition for static user equilibrium (SUE). Here, the term departing can be replaced by arriving for cases with variable departure times and scheduled arrival times. As an outcome of equal travel time paths, trips between the same O-D pair with the same departure time also have equal arrival times. At equilibrium, each trip arrives at its destination on an equal travel time path that departs from its origin within the same time interval.

DUE1 presented here is quasi-continuous in that link volumes are split fractionally between discrete time intervals from which speeds are calculated. In comparison, Friesz et al. (*8*), Wie (*9*), and Wie et al. (*10*) present optimal control theory formulations of dynamic traffic assignment in continuous time for which the equilibrium condition is that no used path between any two nodes has a higher travel time than any other path at any instant. Path choice according to time-of-departure conditions or en route revisions, or both, according to updated conditions in each time interval lead to models in which complete O-D paths used by trips can have unequal travel times for any given departure time.

Wie (*11*) and Ran et al. (*12,13*) refine and extend optimal control models to include elastic demand and departure time choice in user equilibrium or system optimal forms. Friesz et al. (*14*) formulate the simultaneous route choice and departure time problem in continuous time as a variational inequality. Although proposed solution procedures are prohibitive, these papers address important considerations about the behavioral assumptions of alternative model forms.

Department of Civil Engineering, University of Colorado, Campus Box 113, P.O. Box 173364, Denver, Colo. 80217.

DYNAMIC USER EQUILIBRIUM WITH SCHEDULED DEPARTURES

Dynamic user equilibrium with scheduled departures (DUE1) can be stated equivalently in terms of path flows, but the link flow form shown here does not implicitly assume complete enumeration of all paths between zone pairs. Turn movements at each intersection are represented by separate links at each node. The exact form of each link's impedance function can be specific to the intersection or link type. The O-D trip matrix can be developed from traffic counts or from survey data and trip distribution models. In DUE1 stated by Equations 1 through 4, link lengths are computed on the basis of monotonically nondecreasing impedance functions dependent on each link's volume in each time interval.

- Upper Problem (UP)

Minimize

$$\sum_{ij \in K} \sum_{t \in T} \int_0^{x_{ij}^t} f_{ij}^t(w)\, dw \qquad (1)$$

subject to

$$x_{ij}^t = \sum_{r \in Z} \sum_{d \leq t} v_{rij}^d \, \phi_{ri}^{dt} \qquad \text{for all } ij \in K, t \in T \qquad (2)$$

$$q_{rn}^d = \sum_{t \geq d} \left[\sum_{in \in K} v_{rin}^d \phi_{ri}^{dt} - \sum_{nj \in K} v_{rnj}^d \phi_{rn}^{dt} \right] \qquad \text{for all } n \in N, r \in Z, d \in T \qquad (3)$$

$$v_{rij}^d \, \phi_{ri}^{dt} \geq 0 \qquad \text{for all } r \in Z, ij \in K, d \in T, t \in T \qquad (4)$$

$$\Delta b_{ri}^d = b_{ri}^d - b_{ri}^{d-1} \qquad \text{for all } r \in Z, i \in N, d \in T, \text{ and } b_{ri}^0 = b_{ri}^1 - \delta t \qquad (5)$$

and, for all $r \in Z$, $i \in N$, $d \in T$, $t \in T$ in Equations 6a through 6c

$$\phi_{ri}^{dt-k} = \left(\min \{1, [b_{ri}^d - (t-1)\Delta t]/\Delta b_{ri}^d\}\right) \alpha_{ri}^{dt} \qquad \text{for k} = 0 \qquad (6a)$$

$$\phi_{ri}^{dt-k} = [\Delta t/\Delta b_{ri}^d]\alpha_{ri}^{dt} \qquad \text{for all } k > 0 \text{ for which } b_{ri}^{d-1} - (t-1-k)\Delta t \leq 0 \qquad (6b)$$

$$\phi_{ri}^{dt-k} = \left(\max \{0, [\Delta t(t-k) - b_{ri}^{d-1}]/\Delta b_{ri}^d\}\right) \alpha_{ri}^{dt} \qquad \text{for min } k \text{ for which } b_{ri}^{d-1} - (t-1-k)\Delta t > 0 \qquad (6c)$$

where all $\{\alpha_{ri}^{dt}\}$ and $\{b_{ri}^d\}$ are optimal for:

- (Lower problem):

Maximize

$$\sum_{r \in Z} \sum_{i \in N} \sum_{d \in T} b_{ri}^d \qquad (7)$$

subject to

$$\alpha_{ri}^{dt} = (0,1) \qquad \text{for all } r \in Z, i \in N, d \in T, t \in T \qquad (8a)$$

$$\sum_{t \in T} \alpha_{ri}^{dt} = 1 \qquad \text{for all } r \in Z, i \in N, d \in T, t \in T \qquad (8b)$$

$$[b_{ri}^d - t\Delta t]\, \alpha_{ri}^{dt} \leq 0 \qquad \text{for all } r \in Z, i \in N, d \in T, t \in T \qquad (9a)$$

$$[b_{ri}^d - (t-1)\Delta t]\, \alpha_{ri}^{dt} \geq 0 \qquad \text{for all } r \in Z, i \in N, d \in T, t \in T \qquad (9b)$$

$$b_{rr}^d = d\Delta t \qquad \text{for all } r \in Z, d \in T \qquad (10)$$

$$b_{ri}^d = \max [e_{ri}^d, b_{ri}^{d-1} + h\Delta t] \qquad \text{for all } r \in Z, i \in N, d \in T, \text{ and } b_{ri}^0 = b_{ri}^1 - \Delta t \qquad (11)$$

$$\theta_{ri}^{dt} = [(b_{ri}^d - (t-1)\Delta t)/\Delta t]\, \alpha_{ri}^{dt} \qquad \text{for all } r \in Z, i \in N, d \in T, t \in T \qquad (12)$$

$$g_{rij}^{dt} = [\theta_{ri}^{dt} f_{ij}^t(x_{ij}^t) + (1 - \theta_{ri}^{dt}) f_{ij}^p(x_{ij}^p)]\, \alpha_{ri}^{dt} \qquad \text{for all } r \in Z, ij \in K, d \in T, t \in T, p = t-1 \qquad (13)$$

$$\left(e_{rj}^d - \max \{b_{ri}^d, (t-1)\Delta t + \Delta f_{ij}^{tp}\}\right)\alpha_{ri}^{dt} \leq g_{rij}^{dt}\, \alpha_{ri}^{dt} \qquad \text{for all } r \in Z, ij \in K, d \in T, t \in T, p = t-1, \Delta f_{ij}^{tp} = f_{ij}^p(x_{ij}^p) - g_{rij}^{dt} \qquad (14)$$

where

N = set of all nodes;

Z = set of all zones (i.e., trip-end nodes);

K = set of all links (directed arcs);

Δt = duration of each time interval (same for all t);

T = set of all time intervals in the full analysis period (e.g., 18 intervals of 10 min each for 3-hr peak-period assignment);

x_{ij}^t = number of vehicle trips between all zone pairs assigned to link ij in time interval t (variable);

v_{rij}^d = number of vehicle trips departing zone r in time interval d assigned to link ij at some time (variable);

$f_{ij}^t(x_{ij}^t)$ = average travel impedance on link ij in time interval t (variable);

q_{rn}^d = number of vehicle trips from zone r to node n departing in time interval d via any path; 0 for any node $n \notin Z$ (variable);

e_{ri}^d = time (including $d\Delta t$) at which last trip departing zone r in time interval d crosses node i via its shortest path less FIFO delay time at node i (variable);

b_{ri}^d = time(including $d\Delta t$) at which last trip departing zone r in time interval d crosses node i via its shortest path (variable);

α_{ri}^{dt} = 0–1 variable indicating whether last trip departing zone r in time interval d crosses node i in time interval t (henceforth called a "node time interval") (0 = no; 1 = yes) (variable);

ϕ_{ri}^{dt} = fraction of all trips departing zone r in time interval d to cross node i in time interval t (henceforth called a trip flow fraction) (variable);

θ_{ri}^{dt} = fraction of time interval Δt into time interval t that last trip departing zone r in time interval d crosses node i (variable);

g_{rij}^{dt} = "average" travel time on link (i,j) of last trip departing zone r in time interval d adjusted for time into interval t versus $t - 1$ that this trip enters link (variable); and

h = minimum fraction of time interval that trips departing zone r in time interval d must follow trips departing in time interval $d - 1$.

Equation 2 defines total flow on link ij in time interval t to be the sum of flows departing any zone r in any time interval $d \leq t$ using link ij in time interval t to formulate the objective function as given by Equation 1. Conservation of flow Equation 3 constrains inflow minus outflow at each node and zone in each time interval to sum to the proper trip departure totals in each time interval between each O-D pair, and Equation 4 requires all link volumes to be nonnegative. DUE1 requires nonlinear mixed-integer constraints with "node time intervals" and "trip flow fractions," indicating the time intervals in which trips from each origin cross each node so as to ensure temporally continuous trip paths and to spread the trips more continuously in time over these intervals.

A node time interval α_{ri}^{dt} differs from a trip flow fraction ϕ_{ri}^{dt} as follows. A node time interval is a $0 - 1$ variable that indicates the time interval in which trips departing zone r in time interval d "last cross" node i. Each node time interval acts as an "if-then" operator to activate or deactivate certain constraints as needed. A node time interval applies to the last vehicle of each departure time interval, with the time of the last vehicle departing zone r in time interval d to cross node i, via its shortest path given by b_{ri}^{d}. The difference between node i crossing times of "last" vehicles departing in successive time intervals, defined as Δb_{ri}^{d} in Equation 5, is used in Equation 6*a* through *c* to determine the temporal spread of dispersion of trips crossing node i from the same origin.

In Figure 1 explained later, trips v_{rij}^{d} departing zone r in time interval d using link (i,j) are uniformly spread temporally over the vertical difference between the times when "last" vehicles in these two streams pass a given node. In contrast to previous papers, the trip variable v_{rij}^{d} does not have a time superscript t because it is always joined by the variable ϕ_{ri}^{dt}. Arrays $\{\alpha_{ri}^{dt}\}$ and $\{b_{ri}^{d}\}$ used in Equations 5 and 6 are passed in from the lower problem. These equations are not solved simultaneously with the upper problem, because they provide only exogenous inputs to the upper problem and have no variables that vary while solving the upper problem.

One may ask whether shorter time intervals could be used instead of trip flow fractions. Shorter time intervals add to the computational burden of the problem for several reasons. First, more time intervals to span the same analysis period require many more calculations and storage of floating point values than using trip flow fractions. Second, FIFO trip ordering in this formulation is best maintained if link lengths remain well below the time interval length as explained shortly. Using shorter time intervals may require dividing links into shorter links, thus increasing the computational burden of the problem in all three dimensions (nodes, links, and time intervals).

Equations 8 and 9 compute node time intervals that define temporally continuous trip paths with the zone-to-node travel times. Equation 8a defines each α_{ri}^{dt} term to be 0 or 1, which defines the last time interval t in which arcs incident from node i are used by trips departing zone r in time interval d. For trips departing zone r in time interval d, any link ij incident from node i can only be "last" used (if at all) in time interval t, and first used in the last time interval for trips departing in the previous interval d-1. Equation 8*a* allows only one interval t in which trips departing zone r in interval d can last cross node i.

According to Equations 8 and 9, links are traversed within time intervals that trip paths cross their tail nodes. For 5-min intervals, Interval 1 begins at 0, Interval 2 begins at 5 min, Interval 3 begins at 10 min, and so on. If the travel time from zone r to node i is within $t\Delta t$, then α_{ri}^{dt} is 1 because these trips must cross node i in time interval t. Because of tracking last vehicles, if any path crosses a node at the exact start of a time interval (to the degree of floating point precision being used), then the solution algorithm sets $\alpha_{ri}^{dt} = 1$, but all trips from that origin for that departure interval will be assigned to the link over previous time intervals.

In Equation 10, b_{rr}^{d} (equal to the start time of the "last" vehicle departing zone r in time interval d) is set to $d\Delta t$ to correctly set the clock to the end of each time interval and also prevents the (LP) maximization from having an infinite solution. A second subtle change from previous papers is that b_{ri}^{d} includes $d\Delta t$ so that it rep-

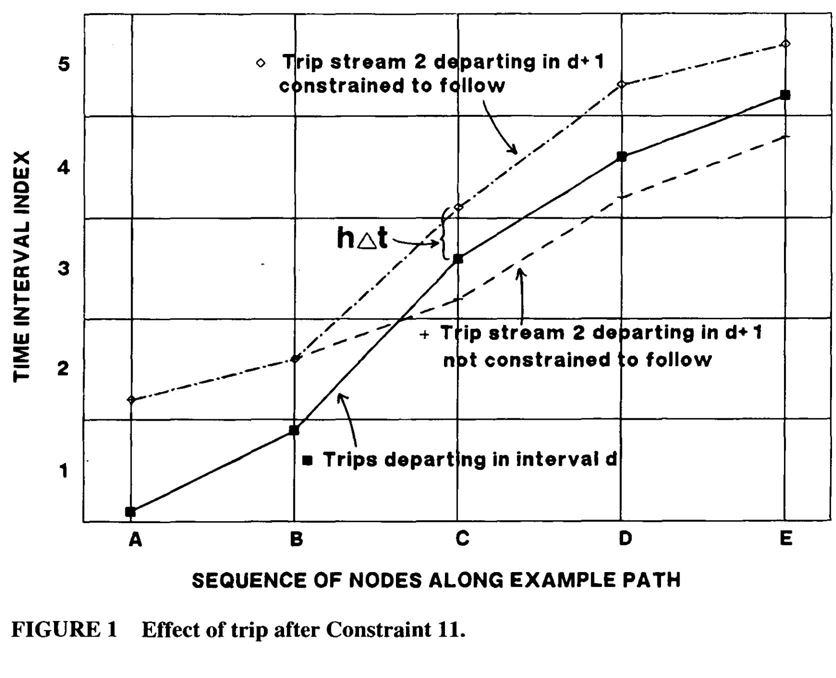


FIGURE 1 Effect of trip after Constraint 11.

resents "clock time" from the start of the entire analysis period rather than travel time from the time of departure. Using "clock time" enables a clearer accounting of time in this quasi-continuous model. Although α_{ri}^{dt} can never equal 1 when $t=1$, trips departing in interval 1 are uniformly distributed over the previous time span $b_{ri}^{1} - \Delta t$ at each node of the network such that some trip volumes are still assigned in interval 1.

Equations 11 through 14 impose FIFO trip ordering between all O-D pairs according to their travel times in successive time intervals as explained next. Vehicles are assumed to make only one-for-one (or zero-sum) exchanges of traffic positions along any link, which is acceptable and expected in aggregate traffic models.

Equation 11 is a vehicle following constraint that prevents later trips from "getting too close" to trips departing earlier from the same zone in successive time intervals so as to prevent these trips from bunching. The value h is the fraction of a time interval that the last trip departing from zone r in time interval d must follow the last trip departing from zone r in interval $d - 1$. Note that when solving for b_{ri}^{d} on the left side of Equation 11, b_{ri}^{d-1} on the right side is held fixed.

Figure 1 illustrates the effect of Constraint 11 on two trip streams (1 and 2) traversing the same series of nodes and departing from the same zone in intervals d and $d + 1$, respectively. The node sequence (A,B,C,D,E) denotes a series of links along the trip path. A trip stream consists of trips with the same origin and departure time interval. At Node A, Trip Stream 2 is $0.8\Delta t$ behind Trip Stream 1. The two trip streams traverse Node A in different time intervals and thus have different travel times for Link A,B. At Node B, the separation between the trip streams has decreased to $0.6\Delta t$. The two trip streams traverse Node B in different time intervals and have different travel times for Link B,C. Without Constraint 11, Stream 2 would pass Stream 1 and traverse Node C earlier. With Constraint 11, Stream 2 is forced to traverse Node C at least $h\Delta t$ behind Stream 1, where $h = 0.5$ in this figure.

Allowing that many other paths include the node sequence (A,B,C,D,E), it can be deduced from Figure 1 that Constraint 11 is less likely to be binding in networks with shorter arc lengths relative to Δt. Hence, this representation works best for networks in which most arc lengths have free-flow travel times less than 20 percent of the interval duration and in which loadings on the network do not cause arc lengths to exceed $h\Delta t$. Example runs revealed the solution algorithm explained later to converge more easily if time-varying travel demands do not cause arc travel times to exceed these bounds.

Although Constraint 11 prevents successive trip streams from passing each other, the exact specification of h can be improved in further research. Reasonable values of h lie between $0.3\Delta t$ and $0.7\Delta t$, but the exact value of h depends on traffic densities of arcs incident to a node in each time interval. However, the value of h must lie between 0 and 1. If $h = 0$, a trailing trip stream can completely overlay (but not overtake) a leading trip stream so that the two streams become coincident, which is not realistic. If $h = 1$, then trailing trips can never partly "gain ground on" leading trips, and later departing trips can never have lower travel times than earlier departing trips. Because Constraint 11 applies to trips departing in intervals $d + 1$ and $d + 2$, trips departing from zone r in interval $d + 2$ must follow trips departing in interval d by at least $2h\Delta t$ at any node. Constraint 11 also applies to all nodes in the network regardless of whether any trips departing from zone r in intervals d and $d + 1$ actually cross node i.

Because Equation 11 does not insure FIFO trip ordering between all O-D pairs, Equations 12 through 14 are also required. Equations 12 and 13 compute an average travel time on link (i,j) of the "last" trip departing zone r in time interval d adjusted for the time into interval t versus $t - 1$ that this trip enters the link. Equations 12 and 13 "smooth out" speed transitions between time intervals in a "quasi-continuous" manner so that vehicle speeds do not abruptly change if they enter links just split seconds before or after a time interval change. When finding shortest paths in the upper problem, equations 12 and 13 are also used to calculate link travel times based on when trips enter links so as to be consistent with how paths are found in the lower problem.

Even without Equation 14, this improvement to the model eliminates FIFO violations for trips between all O-D pairs unless a link's travel time exceeds a full time interval (and even for most of those cases) as required by the following inequality.

$$\theta_{ri}^{dt} f_{ij}^{t} (x_{ij}^{t} + (1 - \theta_{ri}^{dt}) f_{ij}^{p} (x_{ij}^{p}) + \theta_{ri}^{dt} \Delta t \geq f_{ij}^{p} (x_{ij}^{p})$$
$$\text{for all } r \in Z,\ ij \in K, d \in T,\ t \in T,\ p = t - 1 \qquad (15)$$

which simplifies to

$$f_{ij}^{t} (x_{ij}^{t}) + \Delta t \geq f_{ij}^{p} (x_{ij}^{p}) \qquad \text{(q.e.d.)} \qquad (16)$$

If no link travel times exceed Δt, then the left side of Equation 14 could be written more simply as $(e_{ij} - b_{ri}^{d})\, \alpha_{ri}^{dt}$, but Equation 14 is written as shown in DUE1 to prevent FIFO violations in cases where link travel times exceed Δt if needed.

Figure 2 illustrates the effect of Constraint 14 on two trip streams (1 and 2) between any two zones traversing the same series of nodes departing in two successive time intervals. Without Constraint 14, Trip Stream 2 could depart from Node C in Time Interval 3 at a faster speed than Trip Stream 1 and pass stream 1, which departs from node C in Interval 2 at a slower speed. With Constraint 14, Trip Stream 2 will depart from Node C such that it reaches Node D no earlier than Trip Stream 1. Kaufman and Smith (*15*) show that FIFO adjustments such as Constraint 14 are easily added to shortest-path label-correcting algorithms (but not label-setting algorithms) so long as labels are properly updated when it occurs.

Constraint 14 does not entirely replace the need for Constraint 11. Constraint 14 allows trips between different O-D pairs to become concurrent while sharing the same path, whereas Constraint 11 ensures a minimum separation of "last" vehicles departing from the same zone in successive time intervals. If Constraint 11 is removed from the problem, then trips from the same zone can "bunch" together in overly dense flows. As mentioned earlier, additional research will lead to better treatment of this bunching problem.

Albeit counter-intuitive, the maximization of zone-to-node travel times in subproblem (LP) is the correct determination of node time intervals and shortest-path travel times subject to arc lengths plus FIFO delay in Equations 10 through 14. Using the mechanical analogy of Minty (*16*, p. 724), Bertsekas (*17*) defines this formulation as the dual shortest-path problem according to the min-path/max-tension theorem defining this primal-dual relationship.

OPTIMALITY CONDITIONS OF DUE1

Although DUE is nonconvex over the domain of feasible node time intervals for all trip departures to all destinations, DUE1 is convex with a unique global optimum for any given set of fixed node time intervals. The optimality conditions of DUE1 stated in the paper's

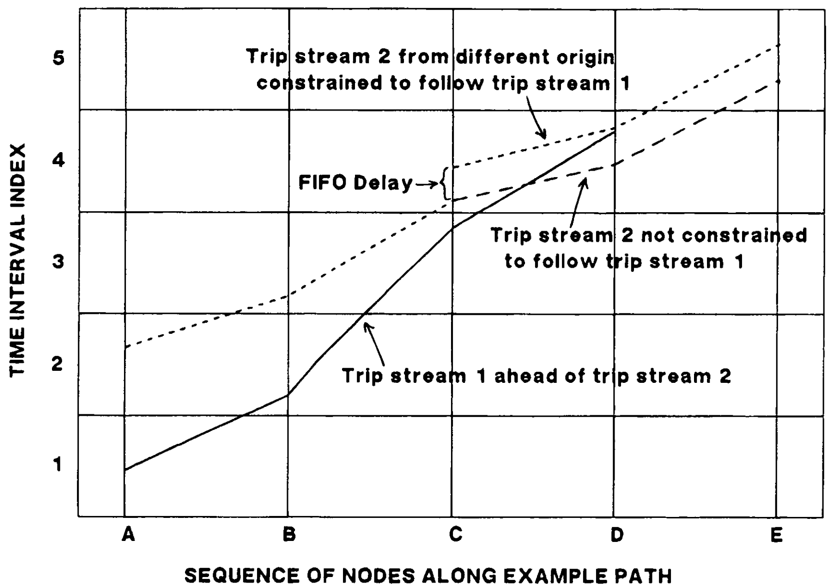


FIGURE 2 Effect of Constraint 14 on FIFO order of all O-D trips.

first section can be derived from (UP) for a given set of node time intervals as given by an optimal solution to (LP). Subproblem (LP) is a shortest-path linear program for which there exists an optimal solution for any given solution to (UP). Any set of node time intervals resulting from (LP) defines a directed network for which (UP) is a convex nonlinear program with a global optimum solution. Because node time intervals resulting from (LP) are uniquely determined by a given set of link volumes resulting from (UP), they can be assumed to be known in the derivation of optimality conditions for DUE1. The optimality conditions of DUE1 are next derived from (UP) for a given set of node time intervals to which all temporally continuous trip paths in the optimal solution must conform.

Equation 17 is the Lagrangian of (UP) with fixed-integer node time intervals. Over the domain of variable integer values, Equation 17 is nonconvex, and there are many local optima that are inferior to the global optimum. For a given set of fixed node time intervals, the bordered Hessian matrix of Equation 17 is positive definite, which means that there is a unique global optimum with no local optima (*18*). The bordered Hessian matrix of Equation 17 is only positive definite so long as each impedance function is a monotonically nondecreasing function of flow on link ij in time interval t alone, as was stipulated earlier.

$$
\begin{aligned}
L(X, V, \lambda, \mu, \tau) = {} & \sum_{ij\epsilon K} \sum_{t\epsilon T} \int_0^{x^t_{ij}} f^t_{ij}(w)\, dw \\
& - \sum_{ij\epsilon K} \sum_{t\epsilon T} \gamma^t_{ij} \left[x^t_{ij} - \sum_{r\epsilon Z} \sum_{d \leq t} v^d_{rij}\ \phi^{dt}_{ri} \right] \\
& + \sum_{r\epsilon Z} \sum_{n\epsilon N} \sum_{d\epsilon T} \mu^d_{rn} \\
& \times \left(q^d_{rn} - \sum_{t \geq d} \left[\sum_{in\epsilon K} v^d_{rin}\ \phi^{dt}_{ri} - \sum_{nj\epsilon K} v^d_{rnj}\ \phi^{dt}_{rn} \right] \right) \\
& + \sum_{r\epsilon Z} \sum_{ij\epsilon K} \sum_{d\epsilon T} \sum_{t\epsilon T} \tau^{dt}_{rij} \left(-v^d_{rij} \phi^{dt}_{ri} \right)
\end{aligned}
\qquad (17)
$$

The optimality conditions are given by Equation 18 through 20.

$$\partial L/\partial x^t_{ij} - < f^t_{ij}(x^t_{ij} = \lambda^t_{ij} \qquad \text{for all } ij \in K,\ t \in T \qquad (18)$$

$$\partial L/\partial v^d_{rij} - < \mu^d_{rj} - \mu^d_{ri})\ \phi^{dt}_{ri} = (\lambda^t_{ij} - \tau^{dt}_{rij})\ \o^{dt}_{ri} \qquad \text{for all r} \in Z,\ ij \in K,\ d \in T,\ t \in T \qquad (19)$$

$$\partial L/\partial\ \tau^{dt}_{rij} - > \tau^d_{rij}\ v^d_{rij}\ \phi^{dt}_{ri} = 0, (\tau^{dt}_{rij} \geq 0) \qquad \text{for all } r \in Z,\ ij \in K,\ d \in T,\ t \in T \qquad (20)$$

where τ^{dt}_{rij} equals 0 if $v^d_{rij}\ \phi^{dt}_{ri} > 0$, nonnegative otherwise; equals impedance difference from node i to *n*ode j via a used path versus by link ij (used or unused) in time interval t for trips departing from zone r in time interval d.

The last part of Equation 17 ensures nonnegative link flows and results in a third optimality condition given by Equation 20, which requires τ^{dt}_{rij} to be 0 if any trips departing zone r in time interval d are assigned to link ij in time interval t and nonnegative otherwise. According to Equation 18, the optimal solution has a unique equilibrium impedance for each link in each time interval. According to Equations 19 through, 20, for any given pair of nodes, all used paths for a given origin and departure time must have equal impedances, and any unused path between these nodes cannot have a lower impedance.

Optimality conditions for dynamic user equilibrium can be stated similarly to Wardrop's (*7*) statement of necessary conditions for static user equilibrium. Let v^d_{rij} be the equilibrium flow on link ij in interval t of trips departing zone r in interval d, and λ^t_{ij} be the equilibrium impedance of link ij in interval t. Also, for trips departing zone r in interval d, let μ^d_{rj} be the equilibrium impedance of all used paths to node i. At equilibrium, according to Equations 21 and 22, all paths from zone r to node j used by trips departing in interval d have impedance μ^d_{rj} and no unused path for this same (r,j,d) combination has a lower impedance.

$$\mu^d_{ri} + \lambda^t_{ij} = \mu^d_{rj} \qquad \text{if } v^d_{rij}\ \o^{dt}_{ri} > 0$$

and

$$\tau^{dt}_{rij} = 0 \qquad \text{for all } r \in Z,\ ij \in K,\ d \in T,\ t \in T \tag{21}$$

$$\mu^d_{ri} + \lambda^t_{ij} = \mu^d_{rj} \qquad \text{if } v^d_{rij}\ \ \phi^{dt}_{ri} = 0$$

and

$$\tau^{dt}_{rij} \geq 0 \qquad \text{for all } r \in Z,\ ij \in K,\ d \in T,\ t \in T \tag{22}$$

CONVERGENT DYNAMIC TRAFFIC ASSIGNMENT ALGORITHM

Whereas SUE can be solved efficiently by linear combination methods for nonlinear programs with all linear constraints [e.g., Frank-Wolfe (F-W) and PARTAN], these methods can easily create temporally discontinuous flows if applied directly to DUE1. Instead, the two subproblems of DUE1 are solved successively by a convergent dynamic algorithm (CDA). As indicated in Figure 3, CDA first solves (UP) with fixed node time intervals using the F-W method of linear combinations (or a similar technique), and then solves (LP) (which is a linear program) to update all node time intervals for the next F-W solution of (UP). Adjustments to link capacities are made between the upper and lower problems to account for spillback queueing effects or signal timing changes as explained in the next section.

The CDA algorithm terminates when fewer than an acceptable number of node time intervals change from one (LP) solution to the next. With fixed node time intervals, subproblem (UP) is solved without fixing which links are used. Only time intervals in which links are used by trips are fixed, depending on their trip origins and departure times. To clarify the CDA algorithm, the following steps are performed successively to solve subproblems (UP) and LP to near convergence.

1. Input all network data, temporal trip departure matrixes, and initial link flows. Initial link flows are optional, and can be set to 0, but SUE link flows reduced to the chosen time interval duration may be good starting values. Calculate initial node time intervals by solving (LP) with initial link flows. Set iteration counter $n = 0$.
2. Increment iteration counter $n = n + 1$.
3. (UP) Minimize Equation 1 subject to Equations 2 through 4, where all x^t_{ij} are variable and all $\{\phi^{dt}_{ri}\}$ are fixed as computed with Equations 5 and 6 according to the optimal values of $\{b^d_{ri}\}$ and $\{\alpha^{dt}_{ri}\}$ from solving (LP).
4. Adjust link capacities, which alters the impedance function $f^t_{ij}(x^t_{ij})$, for links and time intervals affected by spillback queues or other events.
5. (LP) Maximize Equation 7 subject to equations 6 through 14, where all $\{\alpha^{dt}_{ri}\}$ and $\{b^d_{ri}\}$ are variable and all x^t_{ij} are fixed to their optimal values from (UP).
6. Sum NDIFFS = total number of node time interval differences between iterations $n - 1$ and n. Compare each $\{\alpha^{dt}_{ri}\}^{n-1}$. If NDIFFS $\leq$ small percent of all node time intervals [$Z(N - 1)T$], then stop. Otherwise, return to Step 2.

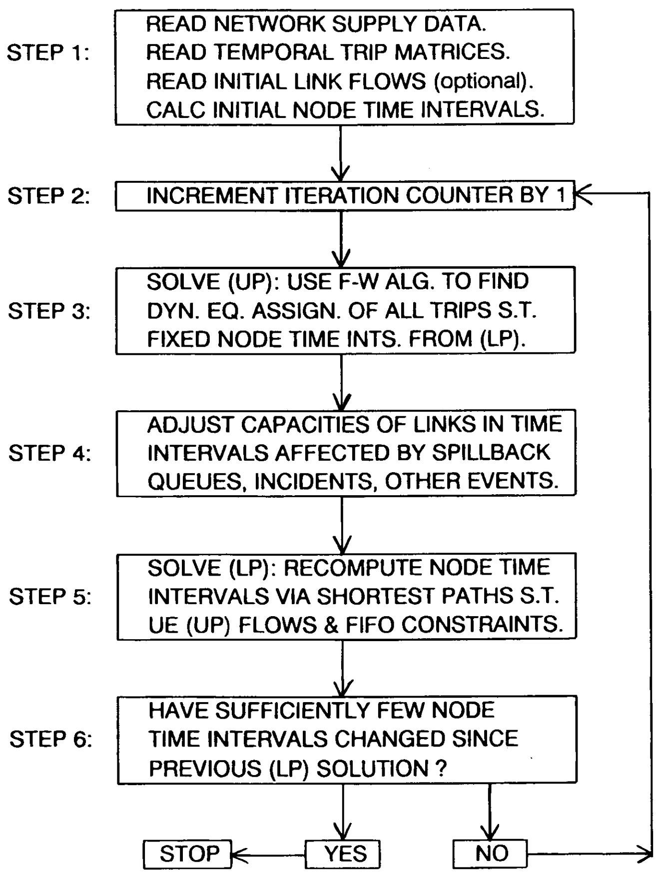


FIGURE 3 Steps of CDA solution algorithm.

CDA converges toward a dynamic user-equilibrium solution for the following reasons. First, if node time intervals corresponding to the true equilibrium are known, then solving (UP) will reproduce the equilibrium link volumes from which these node time intervals can be calculated. That convergence proof follows from the fact that any set of node time intervals resulting from (LP) defines a directed network for which (UP) is a convex nonlinear program for which a global optimum exists. Second, given node time intervals that do not correspond to a true dynamic equilibrium, then solving (UP) with the F-W algorithm will produce link volumes that shift the node time intervals toward their correct values. For example, if a node time interval is too early, then solving (UP) will assign more traffic to paths leading up to that node such that the node time interval is shifted later when recalculated in (LP). Oppositely, if a node time interval is too late, then solving (UP) will assign less traffic to paths leading up to that node such that the node time interval is shifted earlier when recalculated in (LP). Thus, CDA converges to a set of node time intervals that, when used to assign trips to the network in solving (UP), result in temporal link volumes that give rise to the same node time intervals when recalculated in (LP).

DYNAMIC LINK CAPACITY ADJUSTMENTS FOR SPILLBACK QUEUES

A key feature of this modeling approach is the adjustments of link capacities input exogenously or generated endogenously. Exogenous changes (e.g., accidents, weather effects, signal timing changes, or time-of-day road restrictions for special events or construction) can be input to the program for specific links and times of day when they occur (expected or unexpected). Endogenous link capacity changes occur when spillback queues reduce the capacities of upstream links, or when integrated algorithms for ramp metering and signal timing make adjustments. When accidents or recurrent congestion create oversaturated conditions, then upstream link capacities are reduced in time intervals affected by spillback queues using wave propagation speeds to determine the timing of upstream effects. These capacity losses create further upstream effects to the extent and duration of the oversaturated condition. For example, if an accident occurs at 7:30 a.m. and blocks the right lane of a three-lane freeway, then the capacity of that link is reduced by roughly 50 percent from 7:30 a.m. until the estimated clearance time of the accident.

A subroutine called QUECAP is executed between the upper and lower problems of DUE1 to recognize exogenous link capacity changes, and toe compute endogenous link capacity changes. Three essential steps of QUECAP are to (*a*) track the locations of multiple spillback queues in a network, (*b*) weight the effects of multiple spillback queues when jointly affecting inflows to any node, and (*c*) adjust the capacities of inflow links to each node in proportion to the fractions of flows affected by each queue. Endogenous link capacity changes stabilize with successive solutions to the two subproblems of DUE1 and are made only between subproblem solutions so that the upper problem remains convex for each given set of supply specifications.

The queue propagation and capacity adjustment procedure is explained next. First note that all nodes are configured such that every node has only one outflow link (a merge node) or one inflow link (a diverge node). No node has multiple inflow and outflow links (i.e., no merge/diverge nodes). Intersections always have turn movement links, and weave sections always have connecting links between the merge and diverge nodes. Capacity adjustment steps starting from original unadjusted capacities are as follows:

1. Increment the pass number from 1 until convergence is achieved,
2. Increment the time interval from 1 to T,
3. Increment the node number from 1 to N,
4. For each outflow link from a node, compute the cumulative queue equal to all excess flow above capacity through the current time interval. Compute the link length fraction occupied by queueing during the current interval on the basis of the cumulative queue and wave speed. Exact times and locations where accidents initiate queueing can be specified to the program. Spillback from an oversaturated bottleneck link not caused by an accident is assumed to begin at the start of the time interval and at the entry to the link. The link length fraction occupied by a queue is computed by comparing the cumulative queue on the link to what the link can absorb as the vehicle stream compresses to higher density. Wave speed (with upstream being the negative direction) is computed as

$$\text{wave speed} = (\text{flow}^2 - \text{flow}^1) / (\text{density}^2 - \text{density}^1)$$

where

flow^1 = flow before queue (below original but above adjusted capacity),
flow^2 = flow inside queue (assumed to equal the adjusted capacity),
speed^1 = link length / travel time for flow^1 with original capacity,
speed^2 = link length / travel time for flow^2 with adjusted capacity,
density^1 = flow / speed^1 = density before queue (less dense), and
density^2 = flow^2 / speed^2 = density inside queue (more dense).

5. The link length fraction occupied by a queue equals the portion of link length (perhaps all) covered by the queue over successive time intervals. Inflow links are unaffected until a queue extends beyond the tail node of an outflow link, and only a fraction of the time interval will be affected when this first occurs. The affected time interval fraction equals (the interval start time)—(time that queue reaches link's tail) divided by the time interval duration Δt.
6. Compute the weighted volume-to-capacity (V/C) ratio of the outflow links from this node, which is weighted by each outflow link's volume and fraction of the time interval affected. A detailed explanation of this computation is not possible within the brevity of this paper.
7. Adjust the capacity of each inflow link so that its V/C ratio equals the weighted V/C ratio of the outflow links just found. Return to Step 3 until all nodes are processed; then return to Step 2 until all time intervals are processed. Then, if the capacity of any link in any time interval changes by more than an acceptable percent, return to Step 1 for another pass or, else, stop.

To capture the effects of multiple queues spilling back from several places in the network, and queues spilling back farther than one link in any time interval, multiple passes of the above steps are performed until all capacities in all time intervals do not change significantly. Bounding rules are added to prevent endogenously adjusted capacities from becoming too small (which generates a warning), and to prevent queues from dissipating too quickly. The rule here is that a link's capacity during dissipation cannot exceed the average of the above calculation and its reduced capacity in the prior interval. Adjusted capacities are then returned to the upper problem. Hence, both flows and capacities are time dependent in the travel time functions of the DUE1 objective function. This method of adjusting capacities has performed well in these applications but can be improved with further refinements.

SUMMARY AND CONCLUSIONS

This paper converts the DUE modeling approach presented in previous papers by the authors into quasi-continuous time, which enabled three key model improvements: (*a*) inclusion of spillback queueing effects and dynamic link capacity adjustments, (*b*) more accurate FIFO trip ordering between all O-D pairs, and (*c*) better link volume and speed transitions between time intervals. This paper also describes how capacity losses are approximated on upstream links because of spillback queueing from oversaturated links.

This dynamic traffic modeling approach has been applied to several networks of realistic size, detail, and complexity. In addition to the paper by Robles and Janson (*1*) mentioned at the start, Darjardi and Janson (*19*) apply it to the Colorado ski region of rural arterials and I-70 linking Denver and surrounding communities to ski resorts. The CDA found very good solutions to each of these applications without convergence difficulties. These diverse applications show the model to be a flexible analysis tool and CDA to be a robustly convergent solution algorithm for these types of dynamic traffic assignment problems.

REFERENCES

1. Robles, J., and B. N. Janson. Dynamic Traffic Modelling of the I-25/ HOV Corridor Southeast of Denver. In *Transportation Research Record,* in press.
2. Janson, B. N. Dynamic Traffic Assignment for Urban Road Networks. *Transportation Research,* Vol. 25B, Nos. 2/3, 1991, pp. 143–161.
3. Janson, B. N. A Convergent Algorithm for Dynamic Traffic Assignment. In *Transportation Research Record 1328,* TRB, National Research Council, Washington, D.C., 1991, pp. 69–80.
4. Janson, B. N. Network Design Effects of Dynamic Traffic Assignment. *Journal of Transportation Engineering,* ASCE, Vol. 121, No. 1, 1995, pp. 1–13.
5. Janson, B. N. A General Dynamic Assignment Model and Solution Algorithm. Presented at 39th North American Meeting of the International Regional Science Association, Chicago, Ill., Nov. 1992.
6. Janson, B. N., and J. Robles. Dynamic Traffic Assignment with Arrival Time Costs. In *Transportation and Traffic Theory* (C.F. Daganzo, ed.), Elsevier Science Publishers, Amsterdam, The Netherlands, 1993, pp. 127–146.
7. Wardrop, J. G. Some Theoretical Aspects of Road Traffic Research. *Proc., Institution of Civil Engineers,* Vol. 1, Part 2, 1952, pp. 325–378.
8. Friesz, T. L., J. Luque, R. L. Tobin, and B. Wie, Dynamic Network Traffic Assignment Considered As a Continuous Time Optimal Control Problem. *Operations Research,* Vol. 37, No. 6, 1989, pp. 893–901.
9. Wie, B. An Application of Optimal Control Theory to Dynamic User-Equilibrium Traffic Assignment. In *Transportation Research Record 1251,* TRB, National Research Council, Washington, D.C., 1989, pp. 66–73.
10. Wie, B., T. L. Friesz, and R. L. Tobin. Dynamic User-Optimal Traffic Assignment on Congested Multi-Destination Networks. *Transportation Research,* Vol. 24B, 1990, pp. 431–442.
11. Wie, B. Dynamic Analysis of User Optimized Network Flows with Elastic Demand. In *Transportation Research Record 1328,* TRB, National Research Council, Washington, D.C., 1991, pp. 81–87.
12. Ran, B., D. E. Boyce, and L. J. LeBlanc. A New Class of Instantaneous Dynamic User-Optimal Traffic Assignment Models. *Operations Research,* Vol. 41, No. 1, 1993, pp. 192–202.
13. Ran, B., and D. E. Boyce. Dynamic Urban Transportation Network Models: Theory and Implications for IVHS. In *Lecture Notes in Economics and Mathematical Systems.* Springer-Verlag, Berlin, (in press).
14. Friesz, T. L., D. Bernstein. T. E. Smith, R. L. Tobin, and B. Wie. A Variational Inequality Formulation of the Dynamic Network User-Equilibrium Problem. *Operations Research,* Vol. 41, 1993, pp. 179–191.
15. Kaufman, D. E., and R. L. Smith. Fastest Paths in Time-Dependent Networks for IVHS Applications. *IVHS Journal,* Vol. 1, No. 1, 1993, pp. 1–12.
16. Minty, G. J. A Comment on the Shortest Path Problem. *Operations Research,* Vol. 5, 1957.
17. Bertsekas, D. P. *Linear Network Optimization: Algorithms and Codes.* MIT Press, Cambridge, Mass., 1991.
18. Sheffi, Y. *Urban Transportation Networks.* Prentice-Hall, Englewood Cliffs, N.J., 1985.
19. Darjadi, P., and B. N. Janson. *Dynamic Modelling of Winter Weekend Ski Traffic in Colorado.* Colorado Department of Transportation, Denver, 1994.

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

Contributions to Logit Assignment Model

FABIEN M. LEURENT

In the past, research in traffic assignment modeling has been directed primarily toward the deterministic model. Alternative, more behavioral principles were thought to be too demanding computationally. Two mathematical contributions that enable one to solve a logit assignment model with flow-dependent travel times at a reduced cost are presented. First, a convergence test for Fisk's minimization program is introduced on the basis of a duality gap principle. Second, a new definition of Dial's STOCH fixed-time logit assignment procedure is given, in which the set of available paths is defined only once and the computations are reinterpreted. A numerical experiment indicates that these tools make the logit assignment model competitive compared with the procedures conventionally used to solve the deterministic model.

Traffic assignment is the fourth and final step in the conventional travel demand forecasting scheme; by partitioning the origin-destination (O-D) trip rates between several paths, the assignment program attempts to duplicate the vehicular flows on the network. Most assignment models assume that travelers behave rationally. The most well-known assignment principle is that of Wardrop (*1*); that travelers strive to maximize the utility derived from their transportation choices—in other words they try to minimize their generalized travel time. Thus, a user-optimal equilibrium is achieved when no traveler may decrease travel time by unilaterally switching paths.

To account for errors in trip-makers' perception of travel time, Daganzo and Sheffi (*2*) defined the stochastic user principle, according to which all trip makers strive to minimize their stochastic generalized travel time. This rule allows for partitioning O-D trip rates between several alternative paths, even if their true travel times differ from each other.

Two stochastic models are of particular interest: the logit model (*3*) and the probit model (*2,4,5*). The latter, although behaviorally more appealing, is impractical because only Monte-Carlo procedures are available, unless all paths can be identified. The logit model however, is endowed with both an extremely efficient fixed-time assignment procedure (Dial's STOCH2) and a convex minimization formulation with a closed-form objective function (*6*). Nevertheless, computational difficulties have prevented the logit model from enjoying more widespread use. Among other drawbacks, Fisk's (*6*) objective function was thought difficult to evaluate. Only recently have heuristic methods been developed (*7,8*).

In this paper two developments that make computation of a logit user equilibrium competitive with its deterministic counterpart are presented. First, a theoretically sound convergence test for an equilibrium algorithm such as the method of successive averages (MSA) is designed; then it is possible to check whether an equilibrium has been reached. Second, the definition of the set of available paths in Dial's STOCH2 procedure is modified; the procedure is problematic if it is implemented within an equilibration scheme because the path set is likely to change from one iteration to the next. Some changes that remedy this flaw are suggested.

Département Economie et Sociologie des Transports, Institut National de Recherche sur les Transports et leur Sécurité, Avenue du Général Malleret-Joinville, 2, 94114 Arcueil, France.

The organization of the paper is as follows. First the problem is stated formally. Second the convergence test for Fisk's model is introduced. Third a definition of efficient paths that does not depend on congestion phenomena is derived; it is inspired from Dial's STOCH2, and a related path loading procedure is provided, wherein it is easy to compute all the terms in Fisk's objective function. Fourth a numerical experiment is carried out to demonstrate that the MSA, combined with the proposed tools, is indeed an efficient algorithm when applied to the logit model. All proofs of the assertions presented here can be found in work by Leurent (*9*), in which elastic demand and capacity constraints are also considered and a dual solution scheme is proposed.

PROBLEM FORMULATION AND MODELING NEEDS

Logit Equilibrium Model

Let $r-s$ be an O-D pair with traffic flow q_{rs}, θ a nonnegative parameter, k a path from r to s with deterministic travel time T^k_{rs} and flow f^k_{rs}. In the logit assignment model (*3*), it is assumed that the path flow f^k_{rs} is proportional to a negative exponential function of the travel time T^k_{rs}:

$$f^k_{rs} = q_{rs} \frac{\exp(-\theta T^k_{rs})}{\Sigma_k \exp(-\theta T^k_{rs})} \tag{1}$$

Then it is automatically ensured that

$$q_{rs} = \Sigma_k f^k_{rs} \tag{2}$$

The travel time of path k is related to the travel times T_a of the links a that belong to it via

$$T^k_{rs} = \Sigma_{a \epsilon k} T_a = \Sigma_a \delta^{ak}_{rs} T_a \tag{3}$$

where $\delta^{ak}_{rs} = 1$ if $a \epsilon k$ or 0 if not.

Let x_a be the traffic flow on link a:

$$x_a = \Sigma_{rsk} \delta^{ak}_{rs} f^k_{rs} \tag{4}$$

Finally let t_a be the travel time function of link a (assumed to be continuous and nondecreasing):

$$T_a = t_a(x_a) \tag{5}$$

Then Equations 1 through 5 define a logit-based equilibrium. Figure 1 illustrates a logit split between two paths.

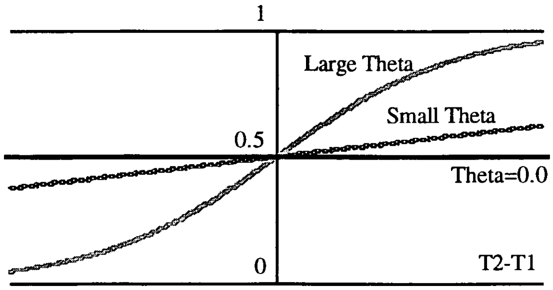


FIGURE 1 Proportion of travelers choosing Path 1 as function of Θ and time difference T_2-T_1 (binary case).

Fisk's Minimization Program

Fisk (*6*) characterized the logit equilibrium with variable travel times as the unique solution to the following convex minimization equation:

$$\min_f J_L(\mathbf{f}) = \Sigma_a \int_0^{x_a} t_a(x)\,dx + \frac{1}{\theta}\,\Sigma_{rsk} f_{rs}^k \log\left(\frac{f_{rs}^k}{q_{rs}}\right) \qquad (6)$$

subject to Equations 2 and 4 and of course to $f_{rs}^k \geq 0$. In Equation 6 Fisk's $\Sigma_{rsk}\ f_{rs}^k\ \log\ (f_{rs}^k)$ was replaced with $\Sigma_{rsk}\ f_{rs}^k\ \log\ (f_{rs}^k/q_{rs})$ to facilitate the understanding of the relationship between Equation 6 and the computations in the STOCH algorithm. This does not alter the existence and uniqueness results obtained by Fisk.

Fisk did not address a crucial question: How should the available paths be defined? In the deterministic model of Beckmann et al. (*10*), all existing acyclic paths may be considered; however, in a logit model a specific definition is required because the conventional shortest-path routines do not automatically find suboptimal paths.

In Dial's paper (*3*), two alternative definitions of efficient paths are provided, namely STOCH and STOCH2. But these definitions are consistent only with respect to fixed travel times (i.e., with constant functions t_a in Equation 6 and cannot be used in a variable-time program. A definition that is consistent with flow-dependent travel times will be provided later (STOCH3 procedure). First, equilibration issues are addressed.

Method of Successive Averages

Powell and Sheffi (*11*) proved the convergence of the MSA applied to minimization programs as Fisk's (provided that the definition of available paths cannot vary).

Fixed-time assignment (FTA) is defined as a path-loading procedure that partitions the O-D flow according to the logit rule, based on a given set of available paths. An FTA yields a solution to Equation 6 with constant travel time functions and a given set of utilized paths. The MSA equilibration algorithm is composed of the following steps:

- Step 0: Initialization.
  –Set iteration counter $n = 0$.
  –Choose a sequence $\alpha^{(k)}$ of real numbers such that $[0 \leq \alpha^{(k)} \leq 1]$, $[\Sigma\alpha^{(k)} = +\infty]$ and $[\Sigma\alpha^{(k)2} < +\infty]$.
  –Find an initial feasible flow pattern $x_a^{(0)} = x_a[\mathbf{f}^{(0)}]$. It may be obtained through an FTA on the basis of link times $t_a^{(-1)} = t_a(0)$.
- Step 1: Link travel time update.
  –Set $t_a^{(n)} = t_a(x_a^{(n)})$.
- Step 2: Direction finding.
  –Compute an FTA of traffic of all O-D pairs on the basis of link travel times $t_a^{(n)}$: this yields a path flow solution $\mathbf{g}^{(n)}$ and also an auxiliary are flow pattern

  $$y_a^{(n)} = x_a[\mathbf{g}^{(n)}]$$

- Step 3: Link flow update.
  –Set $x_a^{(n)+1)} = x_a(\mathbf{f}^{(n+1)}) = x_a^{(n)} + \alpha^{(n)}[y_a^{(n)} - x_a^{(n)}]$
- Step 4: Convergence test.
  –Apply a convergence test: either a maximum number of iterations or a test on the maximum value (over the arcs a of the network) of the change in $\Sigma_{k=1}^{n} \alpha^{(k)} x^{(k)}{}_a / \Sigma_{k=1}^{n} \alpha^{(k)}$ from the previous iteration $n - 1$ to the current one n. If the test is satisfied, then terminate; if not, increment the iteration counter $n = n + 1$ and go to Step 1.

The MSA has been widely applied to solve Fisk's program. However, the definition of efficient paths has not been adequately addressed. Thomas (*12*) wrote that "it seems likely that methods which incorporate definitions of acceptable paths similar to those of Dial and Gunnarsson are intrinsically non-convergent, though in practice users often claim them to be satisfactory in that respect." In the following section, a theoretically sound convergence test is provided for the equilibration algorithm that will be useful together with a formal definition of the efficient paths, as will be given later.

CONVERGENCE TEST FOR LOGIT MODELS

First the issue of designing a theoretically sound convergence test for an application of the MSA to Fisk's program is considered. It is based on a duality gap principle inspired from the deterministic model.

Duality Gap Principle in Deterministic Model

In the deterministic case, where only those paths whose travel times are minimal are used, the objective function reduces to $J_D(\mathbf{f}) = \Sigma_\alpha \int_0^{x_a} (\mathbf{f})\ t_\alpha(x)dx$. The usual convergence test is to evaluate a duality gap between the objective function $J_D[\mathbf{f}^{(n+1)}]$ and a lower bound estimate:

$$J_D[\mathbf{f}^{(n)}] + \nabla J_D[\mathbf{f}^{(n)}] \cdot [\mathbf{g}^{(n)} - \mathbf{f}^{(n)}]$$

where $\mathbf{g}^{(n)}$ is obtained in Step 2 of the MSA (or equivalently of the Frank-Wolfe method). Thus, the duality gap is given by

$$\begin{aligned} DG_D^{(n)} &= \Sigma_a t_\alpha^{(n)}\{x_\alpha[\mathbf{f}^{(n+1)}] - x_\alpha[g^{(n)}] \\ &= \Sigma_{rsk} f_{rs}^{k(n+1)}[T_{rs}^{k(n)} - \min_k T_{rs}^{k(n)}] \end{aligned}$$

The duality gap $DG^{(n)}{}_D$ is always positive, except at equilibrium, at which point it is 0. Hence, a convergence test involves checking whether DG is close to 0.

Convergence Test for Fisk's Model

Application of the duality gap principle to the logit model is suggested. Denoting the entropic part of the logit objective function indicates the following:

$$J_E(\mathbf{f}) = J_L(\mathbf{f}) - J_D(\mathbf{f}) = \frac{1}{\theta}\,\Sigma_{rsk}\, f^k_{rs} \log\left(\frac{f^k_{rs}}{q_{rs}}\right) \qquad (7)$$

Then the flow vector $\mathbf{g}^{(n)}$ considered in Step 2 of the MSA is the unique solution to the following auxiliary program:

$$\min_g J_f^{(n)}(\mathbf{g}) = J_D[\mathbf{f}^{(n)}] + \nabla J_D[\mathbf{f}^{(n)}]\cdot[\mathbf{g} - \mathbf{f}^{(n)}] + J_E(\mathbf{g}) \qquad (8)$$

The duality gap associated with the logit objective function is $DG^{(n)}_L = J_L[\mathbf{f}^{(n+1)}] - \mathrm{LBE}^{(n)}$, where the lower bound estimate $\mathrm{LBE}^{(n)}$ is defined as

$$J_D[\mathbf{f}^{(n)}] + \Delta J_D[\mathbf{f}^{(n)}]\cdot[\mathbf{g}^{(n)} - \mathbf{f}^{(n)}] + J_E[\mathbf{g}^{(n)}]$$

When applying the MSA algorithm to the logit model, generally it is not possible to compute $J_E(\mathbf{f})$, unless all paths are identified. However, for some models such as the one that will be described later, it is easy to compute $J_E(\mathbf{g})$. The trick is to evaluate the duality gap with respect to $\mathbf{g}^{(n)}$ and not with respect to $\mathbf{f}^{(n+1)}$. The following convergence test is also suggested on the basis of functions related to $\mathbf{g}^{(n)}$ rather than to $\mathbf{f}^{(n+1)}$:

if

$$J_L[\mathbf{g}^{(n)}) - \mathrm{LBE}^{(n)} \le \epsilon\,\{|J_L[\mathbf{g}^{(n)}]| + |LBE^{(n)}|\}$$

then

terminate and let $\mathbf{g}^{(n)}$ be the solution to the minimization equation (Equation 6) or else return to Step 1.

If true, the test gives a vector that solves the minimization equation on the basis of the convexity of J_L. Conversely, if the path flow vector $\mathbf{f}^*$ solves the program, then auxiliary vector $\mathbf{g}^*$ that corresponds to $\mathbf{f}^*$ is in fact equal to it and thus the convergence test is satisfied (*9*). If only a relative measure J_L − LBE of the duality gap is needed, then it is not necessary to compute J_E: the test can reduce to check if $J_L[\mathbf{g}^{(n)}] - \mathrm{LBE}^{(n)} \le \epsilon$; in other words to check if $J_D[\mathbf{g}^{(n)}] - J_D[\mathbf{f}^{(n)}] - \Delta J_D[\mathbf{f}^{(n)}]\cdot[\mathbf{g}^{(n)} - \mathbf{f}^{(n)}] \le \epsilon$.

DEVELOPMENT OF STOCH3 PROCEDURE

The results obtained so far apply to any set of utilized paths under the sole constraint that no path may include a given node more than once. Now a set of efficient paths that enable one to benefit from the efficiency of Dial's STOCH2 is defined.

Most previous logit assignment models have used Dial's second definition of efficient paths, according to which "a path is efficient (reasonable) if every link in it has its initial node closer to the origin than is its final node." The word "closer" refers to the travel time measured from the origin with respect to a current travel time vector that may change from one iteration to the next. Therefore, there was no use trying to compute an objective function for the logit assignment model.

Three problems had to be tackled:

1. Restrict Dial's set of efficient paths so as to limit its size and for each reasonable path not to be much longer than the shortest one,
2. Stabilize the definition of efficient paths so that it depends neither on congestion nor on the iteration number, and
3. Find a way to compute the entropic part of the objective function, so as to measure the convergence rate.

The first two problems are discussed first on the basis of previous work by Tobin (*13*). Then the STOCH3 procedure, which offers a practical way to perform a fixed-time logit assignment on the efficient paths defined formerly, will be introduced. Finally, a way to evaluate $J_E(\mathbf{g})$ in the STOCH3 model will be described.

Definition of Stable Set of Efficient, Not-Too-Long Paths

A path is considered "STOCH3 efficient" (or reasonable, or available) if it does not include the same node more than once, if every link has its initial node closer to the origin than its final node, if every link is "reasonable enough" compared with a reference shortest path.

More precisely, let

T^0_a = a reference generalized travel cost for Link a;
$C^0_r(n)$ = a reference shortest generalized travel cost from origin r to node n, based on link costs T^0_a;
h^a_r = a maximum "elongation ratio" for link a origin r;
B_a, E_a = the beginning and end nodes respectively, of Link a.

Definition 1; a path k from origin r to destination s is STOCH3 efficient if (*a*) it does not comprise a given node more than once, (*b*) $C^0_r(E_a) > C^0_r(B_a)\ \forall\, a \in k$; and (*c*) $(1 + h^a_r)\,[C^0_r(E_a) - C^0_r(B_a)] \ge T^0_a$, with $h^a_r \ge 0,\ \forall\, a \in k$.

A link a that satisfies the last two conditions is called STOCH3-reasonable wrt. origin r.

The last condition in Definition 1 limits the number of efficient paths by limiting their total reference generalized travel cost: defining $H_r = \max_a h^a_r$, summing over all links a that are incident to an efficient path k yields that

$$\begin{aligned}\text{Length}(k) &= \sum_{a\in k} T^0_a \le (1 + H_r)(C^0_r(s) - C^0_r(r))\\ &= (1 + H_r)\min_{k'} \text{length}(k')\end{aligned}$$

Conversely, if k satisfies length $(k) \le (1 + H_r) \min_{k'}$ length (k'), it may not be efficient because the first two conditions in Definition 1 must hold as well.

Definition 1 is inspired from Dial's specification STOCH2 (*3*), with respect to the second condition, and from Tobin (*13*) with respect to the third. The contribution of the author is to impose fixed reference travel costs, thus ensuring a stable definition of the efficient paths, whatever the congestion phenomena.

STOCH3 Procedure

Recall that in the STOCH3 procedure it is necessary to consider, on the one hand, the reference generalized travel costs to enumerate the available paths and, on the other hand, the "actual" travel times according to which the O-D flows are partitioned between the paths.

Equation Variables

The following variables apply.

n = node with reference travel cost $C^0_r(n)$ from origin r,
$O_r(i)$ = the ith node in order of increasing access cost $C^0_r(n)$ from r,
Ω^a_r = an indicator variable of 1 if link a is reasonable from r and 0 otherwise,
T_a = current travel time on link a,
$A(a)$ = impedance of link a,
$W_A(a)$ = link weight that accounts for importance of a in contributing to a reasonable path,
$W_N(n)$ = node weight,
$X_A(a)$ = flow on link a from current origin r,
$X_N(n)$ = flow passing through node n from current origin r, and
$F(a)$ = total current flow on link a (over all origins).

Index r can be omitted when writing variables A, W_A, W_N, X_A, and X_N because these variables do not need to be stored after dealing with origin r.

Algorithm STOCH3

- Step 0: Overall preliminaries: calculation of reasonable path.
  –From every origin node r, compute the shortest paths to all nodes n, on the basis of the reference link travel costs T^0_a, yielding the reference access costs $C^0_r(n)$ and a labeling $O_r(i)$ of the nodes n in the order of increasing access cost from r.
  –For each link a, set $\Omega^a_r := 1$ if $(1 + h^a_r)\,[C^0_r(E_a) - C^0_r(B_a)] \geq T^0_a > 0$, $\Omega^a_r := 0$ otherwise.

- Step 1: Preliminaries for a standard iteration.
  –Initialize the total link flow variables $F(a)$ to 0.
  –Set the link impedances $A(a) = \exp(-\theta T_a)$.

Steps 2, 3 and 4 are to be run for each origin node r.

- Step 2: Forward pass.
  –Set all $W_A(a)$ and $W_N(n)$ to 0. Set $W_N(r) = 1$.
  –For each node n taken in the order of increasing reference cost $C^0_r(n)$ (the ith node to be considered is indicated by $O_r(i)$), for each link a with beginning node $B_a = n$; if $\Omega^a_r = 1$, then compute $W_A(a) = A(a)W_N(n)$ and add $W_A(a)$ to $W_N(E_a)$ or else do nothing.

- Step 3: Backward pass.
  –For each node n, set $X_N(n) = q_{rn}$ if n is a destination node, 0 otherwise.
  –For each node n taken in the order of decreasing reference cost $C^0_r(n)$ (use the labeling $O_r(i)$ in decreasing order), for each link a with end node $E_\alpha = n$, if $\Omega^a_r = 1$ then compute $X_A(a) = X_N(n)W_A(a)/W_NE_a)$ and add $X_A(a)$ to $X_N(B_a)$ or else set $X_A(a) = 0$.

- Step 4: Contribution to total link flows.
  –$\forall$, $F(a) = F(a) + X_A(a)$

At the end of the procedure, the vector **F** gives the fixed-time logit assignment on the basis of link travel times T_a.

Computation of Entropic Part of Objective Function in STOCH3 Model

It is shown by Leurent (*9*) that, at the end of the forward pass from origin r, it holds that

$$W_N(s) = \Sigma_k \exp(-\theta T^k_{rs}) \tag{10}$$

$$g^k_{rs} = q_{rs} \exp(-\theta T^k_{rs}) / \Sigma_k \exp(-\theta T^k_{rs})$$

therefore

$$\frac{1}{\theta} \Sigma_k g^k_{rs} \log\left(\frac{g^k_{rs}}{q_{rs}}\right) = -\Sigma_k g^k_{rs} T^k_{rs} - \frac{q_{rs}}{\theta} \log\left[\Sigma_k \exp(-\theta T^k_{rs})\right]$$

By summing over all O-D pairs $r - s$,

$$J_E(\mathbf{g}) = \frac{1}{\theta} \Sigma_{rsk} g^k_{rs} \log\left(\frac{g^k_{rs}}{q_{rs}}\right) = -\Sigma_a x_a(\mathbf{g}) \cdot T_a - \frac{1}{\theta} \Sigma_{rs} q_{rs} \log\left[\Sigma_k \exp(-\theta T^k_{rs})\right]$$

Then, the convergence test designed earlier can be applied to the STOCH3 set of available paths.

COMPUTATIONAL EVIDENCE

In this section, a numerical example to compare the performance of the STOCH3 logit model using the MSA is compared with that of the deterministic model using both the Frank-Wolfe algorithm and the MSA.

Case Study

The application is related to the western part of the Paris metropolitan area during the evening peak period, with a typical trip travel time of 1 hr. The test network is composed of 2,000 directed links. There are 141 O-D zones.

The dispersion parameter θ is set to 0.233 mn^{-1} so that when two routes compete with each other, the first one with a travel time 5 min shorter than the second one, approximately three out of four drivers choose the first road. Because only the rate of convergence is of interest here, the elongation ratios $h^a{}_r$ are set to $+\infty$ [as noted from previous surveys they may be set to $h^a{}_r = 1.6$ for interurban studies (*14*) or $h^a_r \in [1.3; 1.5]$ for urban studies (*15*)].

Results

Figure 2 depicts the performance of the three algorithms, showing the evolution of

$$\log\left|\frac{X^{(n)}}{J^*} - 1\right|$$

where (*a*) in the logit model, J^* is the optimal value of the objective function in Equation 6, and $X^{(n)}$ is the value of $J_L\,[\mathbf{g}^{(n)}]$. In the MSA, the step size $\alpha^{(n)}$ is set to $1/(4 + n/10)$; (*b*) for the deterministic model, J^* is the optimal value of the deterministic objective func-

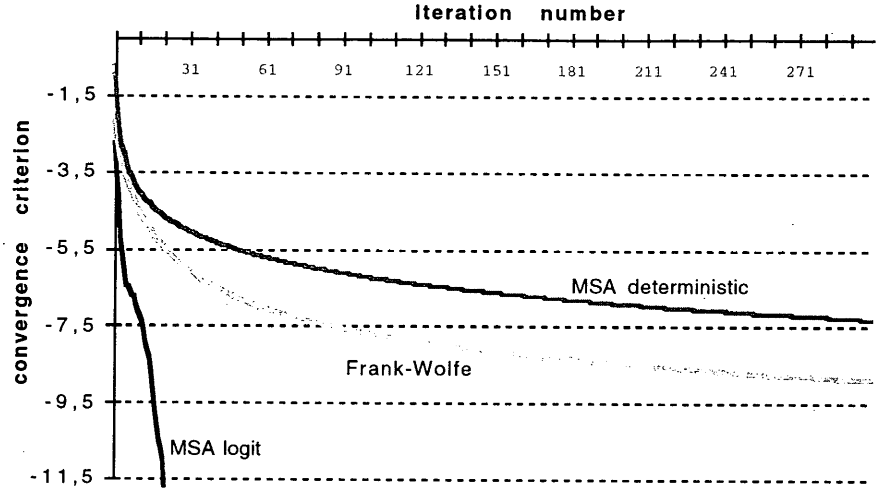


FIGURE 2 Convergence rates of three algorithms.

tion, and $X^{(n)}$ is the value of $J_D\,[\mathbf{f}^{(n+1)}]$. In the MSA, the step size $\alpha^{(n)}$ is set to $1/(1+n)$. The convergence rate is much better in the case of the logit model, notably because the descent direction includes information about all of the available paths—not only about the shortest path in each iteration.

COMMENTS AND CONCLUSION

Intelligent Vehicle-Highway System Implications

In an intelligent vehicle-highway system context, the logit model may be of particular interest for assessing the level of information provided to motorists by a route guidance system (*16*). One way to evaluate the effects of such a system is to model two classes of motorists, the first equipped with a route guidance device and characterized by a large dispersion parameter θ, and the other class of nonequipped drivers characterized by a small θ.

Model Extensions

The case of elastic demand and capacity constraints is addressed by Leurent (*9*). A dual solution scheme is also introduced, but for large-scale applications it is not efficient. The computational efficiency of the MSA applied to the logit assignment model facilitates the following possible extensions of the model:

- Diagonalization schemes, for example with travel time functions that depend on flows of several links (it is easy to derive a variational inequality formulation of Equation 6), and
- Simultaneous models that capture more than one step in the conventional transportation planning process.

Path Identification

- *It is useful to identify paths.* The STOCH procedure is a way to consider all available paths at a reduced cost. The numerical experiment demonstrates here, above all, that path-based equilibration algorithms are much more efficient than link-based algorithms. This conclusion is also supported by recent work (*17,18*).
- *Algorithms that identify paths should better address more behavioral models.* In a fixed-time path-loading procedure like STOCH, the O-D flow is partitioned between the paths according to a behavioral rule. Other available behavioral rules are the probit model [(*4,5*); see the work of Daganzo and Sheffi (*2*) and Powell and Sheffi (*11*) for a mathematical foundation], and the bicriterion, cost-versus-time model [(*19*); Leurent (*20*) gives a mathematical foundation]. By applying a behavioral rule, the need to search for an effective step size in the descent is bypassed. It is thus remarkable that, by the identification of paths, the computational process is greatly facilitated, especially in the case of behavioral models.

ACKNOWLEDGMENTS

Editorial suggestions from Laura Marmorstein-Wynter are greatly appreciated, as are comments from anonymous referees.

REFERENCES

1. Wardrop, J. G. Some Theoretical Aspects of Road Traffic Research. *Proc., Institute of Civil Engineers,* Part 1952, pp. 325–378.
2. Daganzo, C. F., and Y. Sheffi. On Stochastic Models of Traffic Assignment. *Transportation Science,* Vol. 11, No. 3, 1977, pp. 253–274.
3. Dial, R. B. A Probabilistic Multipath Traffic Assignment Model Which Obviates Path Enumeration. *Transportation Research,* Vol. 5, 1971, pp. 83–111.
4. Abraham, C. La Répartition du Traffic entre Itinéraires Concurrents. *Revue Générale des Routes et Aérodromes,* Paris, France, June 1961.
5. Burrell, J. E. *Multiple Route Assignment and Its Application to Capacity Restraint. Proc. 4th International Symposium Theory of Traffic Flow,* Karlsruhe, Germany, 1968.
6. Fisk, C. Some Developments in Equilibrium Traffic Assignment Methodology. *Transportation Research B,* Vol. 14, 1980, pp. 243–256.
7. Chen, M., and A.S. Alfa. Algorithms for Solving Fisk's Stochastic Traffic Assignment Model. *Transportation Research B,* Vol. 25, No. 6, 1991, pp. 405–412.
8. Damberg, O., J. T. Lundgren, and M. Patriksson. *An Algorithm for the Stochastic User Equilibrium Problem.* Working Paper. Department of Mathematics, Linköping University, Sweden, 1992.

9. Leurent, F. *Logit-Based Equilibrium Assignment Model with Elastic Demand Capacity Constraints and a Dual Solution Algorithm.* Actes INRETS, Vol. 38, 1994, pp. 125–162.
10. Beckmann, M., C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation.* Yale University Press, New Haven, Conn. 1956.
11. Powell, W. B., and Y. Sheffi. The Convergence of Equilibrium Algorithms with Predetermined Step Sizes. *Transportation Science,* Vol. 16 No. 1, 1982, pp. 45–55.
12. Thomas, R. *Traffic Assignment Techniques.* Avebury Technical, Aldershot, England, 1991.
13. Tobin, R. L. An Extension of Dial's Algorithm Utilizing a Model of Tripmakers' Perceptions. *Transportation Research,* Vol. 11, No. 5, 1977, pp. 337–342.
14. *Comportement des Automobilistes en matière de Choix d'Itinéraires.* Union des Sociétés d'Autoroutes à Péage, Paris, France, June 1992.
15. Tagliacozzo, F., and F. Pirzio. Assignment Models and Urban Path Selection Criteria: Results of a Survey of the Behaviour of Road Users. *Transportation Research,* Vol. 7, 1973, pp. 313–329.
16. Van Vliet, D. *Modelling Route Guidance Systems Using SATURN.* DRIVE Project V1054 (ASTERIX) Deliverable 4, 1990, pp. B-83–B-105.
17. Schittenhelm, H. *On the Integration of an Effective Assignment Algorithm with Path and Path-Flow Management in a Combined Trip Distribution and Traffic Assignment Algorithm. Proc., 18th Summer Annual Meeting of PTRC,* 1990.
18. Larsson, T., and M. Patriksson. Simplicial Decomposition with Disaggregate Representation for the Traffic Assignment Problem. *Transportation Science,* Vol. 26, 1992 pp. 4–17.
19. Marche, R. *Le modèle TRIP.* SETEC, Paris, France, 1973.
20. Leurent, F. Cost Versus Time Equilibrium over a Network. *European Journal of Operations Research,* Vol. 71, No. 2, 1993, pp. 205–221.

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