TRANSPORTATION RESEARCH
RECORD

No. 1328

Planning and Administration

Travel Demand Forecasting: New Methodologies and Travel Behavior Research 1991

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1991

Transportation Research Record 1328
Price: $22.00

Subscriber Category
IA planning and administration

TRB Publications Staff
Director of Publications: Nancy A. Ackerman
Senior Editor: Naomi C. Kassabian
Associate Editor: Alison G. Tobias
Assistant Editors: Luanne Crayton, Norman Solomon
Graphics Coordinator: Diane L. Snell
Production Coordinator: Karen S. Waugh
Office Manager: Phyllis D. Barber
Production Assistant: Betty L. Hawkins

Printed in the United States of America

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

Travel demand forecasting : new methodologies and travel behavior research 1991.
p. cm.—(Transportation research record, ISSN 0361-1981; no. 1328)
"A peer-reviewed publication of the Transportation Research Board."
ISBN 0-309-05161-4
1. Choice of transportation—Forecasting—Mathematical models. I. National Research Council (U.S.). Transportation Research Board. II. Series: Transportation research record ; 1328.
TE7.H5 no. 1328
[HE336.C5]
388 s—dc20
[388'.049] 92-3195
CIP

Sponsorship of Transportation Research Record 1328

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

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

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

Committee on Traveler Behavior and Values
Chairman: Ryuichi Kitamura, University of California-Davis
Elizabeth Deakin, Bernard Gerardin, Thomas F. Golob, David T. Hartgen, Joel L. Horowitz, Lidia P. Kostyniuk, Martin E. H. Lee-Gosselin, Hani S. Mahmassani, Patricia L. Mokhtarian, Frank Southworth, Peter R. Stopher, Janusz Supernak, Antti Talvitie, Mary Lynn Tischer, Carina Van Knippenberg, Martin Wachs, William Young

Committee on Transportation Supply Analysis
Chairman: Hani S. Mahmassani, University of Texas at Austin
David E. Boyce, Yupo Chan, Carlos F. Daganzo, Mark S. Daskin, Randolf W. Hall, William C. Jordan, Eric J. Miller, Anna Nagurney, Jossef Perl, Earl R. Ruiter, K. Nabil A. Safwat, Yosef Sheffi, Mark A. Turnquist

James A. Scott, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1990.

Transportation Research Record 1328

Contents

Foreword

Said et al. indicate that current trip generation techniques do not explicitly allow for treatment of the presence of structurally different socioeconomic groups in urban areas. The paper describes a statistical approach based on the general linear model that can be used in such cases. A case study based on data from the Kuwait metropolitan area is presented.

Kurth and Chang discuss a travel model that has been calibrated to project ridership on various circulator/distributor alternatives in the central area of Chicago. The model is based on the Downtown People Mover System.

Bajpai examines the sensitivity of the urban transportation planning process (UTPP) to differences (or errors) in socioeconomic input using the Dallas–Forth Worth area as a case study. The sensitivity analysis indicated that the final output of the UTPP, link volumes, is sensitive to errors in the district-level forecasting of population and employment. The author suggests that planners undertaking corridor-level studies should be most concerned about the reliability and accuracy of district-level socioeconomic forecasts, particularly for districts directly served by the corridor.

Morikawa et al. present a methodology for incorporating revealed preference and stated preference data in discrete choice models, apply the methodology to intercity travel mode choice analysis, and predict new mode shares for each origin-destination (OD) pair resulting from changes in service levels.

Davis and Nihan indicate that a number of researchers have sought to develop methods for estimating the OD matrix from observations of traffic volumes on the region's road network. The authors believe that the link counts on a traffic network are generated by a stochastic process that is parameterized by the means and variances of the separate OD flows. According to the authors, by using a tractable approximation to this traffic generating process, it is possible to develop maximum likelihood and method of moments estimators of these OD parameters, and these estimators have desirable consistency and asymptotic normality properties.

Chan presents a methodology to provide demand elasticities that are practical for patronage analyses in an operating agency. The results presented in this paper include guidelines for selecting an appropriate value of elasticity among the broad range of values and a method for converting the most commonly available elasticities to a more useful form, such as route-specific and time-of-day elasticities.

Prevedouros and Schofer present the results of an investigation of the variations in travel behavior across social groups and between location on the basis of a mid-1989 mail-back survey of individuals residing in Chicago suburbs. A primary objective of the research was to gain more knowledge about the causes of variations in traffic congestion. According to the authors, four prominent factors associated with traffic congestion are residence location, population aging, working women, and fixed work hours.

Talvitie and Koskenoja explore the socioeconomic determinants of automobile travel cost choices. The reported work is part of a study into the effects of quality of work on work-related travel conducted in the Road and Traffic Laboratory, Technical Research Center of Finland. The study started from the need to assess driving cost estimates for the mode choice model and was approached through separate functions for fixed, variable, and total driving costs of the respondent. Within-household effects were examined through driving cost regressions of the household members of the respondent. According to the authors, the driving cost models indicate that travel choices are a part of complex behavioral interplay within a household in which the costs of transport have a major role.

Janson presents a link flow formulation and a convergent solution algorithm for the dynamic user equilibrium (DUE) traffic assignment problem for road networks with multiple trip origins and destinations. DUE is a temporal generalization of the static user equilibrium assignment problem with additional constraints to ensure temporally continuous paths of

flow. As the author states, dynamic traffic assignment procedures are needed to evaluate the impacts of alternative travel demand management strategies during peak periods in urban areas and will play a key role in the development of real-time traffic management and in-vehicle route guidance systems.

Wie discusses an equivalent continuous time optimal control problem for dynamic user-optimized traffic assignment with elastic travel demand. Using the Pontryagin minimum principle, the author derives optimality conditions and provides economic interpretations that correspond to a dynamic generalization of Wardrop's first principle. Under steady-state assumptions, the model is shown to be a proper dynamic extension of Beckmann's equivalent optimization problem for static user-optimized traffic assignment with elastic demand.

Haghani presents an alternative approach to simultaneously consider several objectives in public- and private-sector location decisions. The approach is based on the analytical hierarchy process, which is a useful tool in multicriteria decision making. The approach is demonstrated by a numerical example.

Gendreau et al. describe a study to explore issues related to modeling and estimation of short-term demand for empty containers for subsequent movements on international shipping lines. The study is part of a larger research effort directed toward the development of models, methods, and integrated planning tools to address typical problems related to the management of the land distribution and transportation of containers.

Trip Generation Procedure for Areas with Structurally Different Socioeconomic Groups

GALAL M. SAID, DAVID H. YOUNG, AND HASSAN K. IBRAHIM

Many urban areas contain a mix of ethnic groups or households with structural differences in income and social status. Current trip generation techniques do not explicitly allow for the treatment of these structurally different socioeconomic groups. A statistical approach that can be used in such cases is described. The proposed framework allows for the treatment of different ethnic household groups; each group can have different subgroups based on one or more qualitative features, such as dwelling unit type or income. A case study based on data from the Kuwait metropolitan area is presented. Households are classified according to three nationality groups and four house types. The house type is found to be significant for only one of the groups. Differences in important quantitative household characteristics, such as car ownership and number of adults and children, appear both between and within subgroups.

Current household-level trip generation procedures assume a degree of homogeneity in the mix of households in the urban areas under investigation. Households are allowed to differ in such features as size, car ownership, and income, but normally there is no allowance for the presence of structurally different household groups. Many large cities in different parts of the world are growing increasingly into multiethnic urban areas, and the differences between groups may merit special treatment.

Structural differences in household groups in urban areas require that the adopted household-level trip generation procedure establish the extent to which the differences merit consideration and how the features can be allowed for without excessive sample sizes for calibration. Furthermore, there is a need for adopting mean trip rate estimate procedures that overcome the expected problem of poor reliability when trip rates are estimated for groups containing few households.

The purpose of this paper is to present the framework and an application of a procedure based on the concept of generalized linear models that can be used in estimating mean trip rates of households in urban areas with a distinct mix of household groups. The proposed procedure allows for

1. The presence of structurally different household groups;
2. The presence within each household group of subgroups, which differ in other qualitative features such as house type and income category; and
3. Households within each of the preceding classifications varying in other characteristics that can be described using quantitative variables, such as size, car ownership, and number of adults.

Kuwait University, P.O. Box 5969, Kuwait.

Identification of the significance of any of these classifications is done by using statistical testing procedures. Furthermore, the effects of each variable can be assessed. The proposed approach overcomes several problems associated with the widely used cross-classification analysis procedure, which has several shortcomings when used for urban areas with diverse characteristics.

These difficulties have been reported in the literature (*1–3*). Stopher and McDonald (*2*) presented a multiple classification analysis procedure based on an extension of analysis of variance (ANOVA) to respond to some problems of cross-classification analysis, especially those related to poor reliability of trip rate estimates. Dobson (*4*) discussed the possibility of using the general linear model analysis of variance in conjunction with cross-classification analysis. Rickard (*5*) described an application of generalized linear models to railway trips. Said and Young (*3*) proposed a general linear model (GLM) framework for modeling trips of one of the households groups in Kuwait as a function of quantitative household variables. Said et al. (*6*) extended this analysis to include qualitative variables and addressed the use of GLM with cross-classified household data using its regression and ANOVA specifications.

TRIP RATE DATA OF HOUSEHOLD GROUPS IN KUWAIT

The characteristics of the different population groups in Kuwait have been described previously (*1,7*). The Kuwait population reached 1.697 million in 1985; 40.1 percent were Kuwaitis, 37.9 percent were non-Kuwaiti Arabs, and 21.0 percent were non-Kuwaitis of Asian origin. The age-sex distributions of these groups are markedly different. The Kuwaiti population is dominated by younger age groups. Arab and Asian population groups are dominated by individuals in the working ages, and the number of males is almost double the number of females.

Labor force participation rates of the nationality groups are also different. Kuwaiti females have noticeably low participation rates compared with other groups. Participation rates of Arab males and Asian males and females are extremely high, a result of the labor laws that govern foreigners in Kuwait.

Households of different nationalities vary in size. The average sizes of Kuwaiti, Arab, and Asian households are 9.2, 5.4, and 3.9, respectively. Kuwaiti households are large, with the majority of household members in the school-age group. Non-Kuwaiti Arab households are mostly of the family type, of medium to large size. Asian households are small to medium, with few members at the pre-work ages. The average numbers of working persons in these households are 1.48, 2.00, and 3.30, respectively. The number of workers per household reflects labor force participation rate differences and the age-sex composition. There are also significant variations in the occupation status of workers in the household groups.

The nationality of households in Kuwait is an important qualitative factor that has implications for social status, income, household structure and composition, and possible occupational status of working individuals of these households. Variations in trip rates indicate an important nationality effect, so households of different nationalities need to be treated separately (*1*). Three household groups are identified on the basis of the nationality types listed earlier.

For households within each nationality, house type provides another important qualitative factor. Four house types are commonly available: private villas, government housing (mostly villas), apartments, and "others." The latter includes Arabian houses (old-style villas) and annexes. Only very small Kuwaiti households live in apartments; 70 percent are in private and public villas. Only 2 to 3 percent of non-Kuwaiti households live in villas, whereas more than 75 percent live in apartments, and the rest live in other housing types.

Table 1 classifies households included in the 1988 home interview into three nationalities and four house types. Because government housing is available only for Kuwaitis, the two cells for Arab and Asian households in government housing have zero frequencies.

The raw work trip data used in this study are restricted to trips made in the morning peak period between 6:30 and 8:00 a.m. This restriction has resulted in the recording of fewer trips than expected. If work trip data for a longer morning duration were available, greater differences in the mean trip rates of household groups would have been detected. This is evident from Figure 1, in which the recorded mean trips of households of different sizes are plotted against the number of full-time workers of these households. Figure 1 indicates that the mean number of full-time workers is generally about 50 percent more than the recorded mean trip rates.

Said (*1*) described variations in trip rates of households of the three nationality groups for different levels of household

TABLE 1 HOUSEHOLDS IN 1988 HOME INTERVIEW SURVEY OF KUWAIT CLASSIFIED BY NATIONALITY AND HOUSE TYPE

House Type	Nationality			
	Kuwaitis	Arabs	Asians	Total
Villas	1107	155	12	1274
Public Housing (MHA)	762	0	0	762
Apartments	140	2640	296	3076
Others	183	688	68	939
Total	2192	3483	376	6051

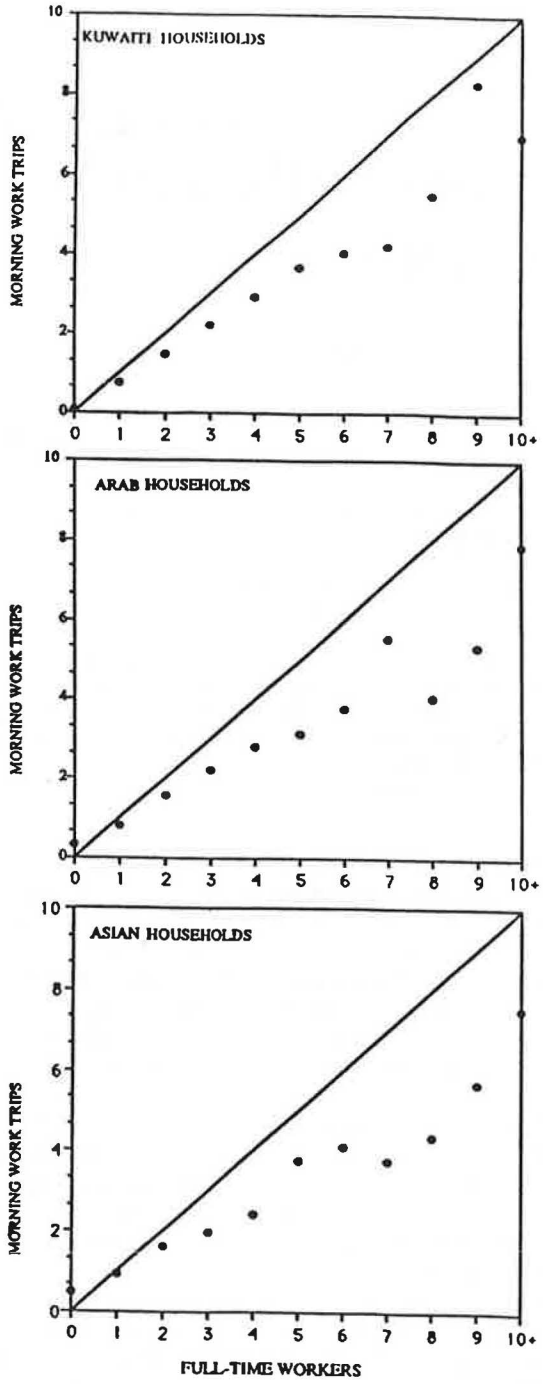

FIGURE 1 Mean morning work trip rates and number of full-time workers for households of various nationality groups.

size, car ownership, income, and numbers of adults and children in the household. Through the use of simple regression, Said and Young (*3*) showed that household size, car ownership, and number of adults in the household, when used as independent variables, are significant in explaining much of the variation in trips made by households. Said et al. (*6*) showed that, for Kuwaiti households, use of the number of children variable instead of household size is more appropriate because it complements the number of adults variable.

The regression analysis using ungrouped data for Kuwaiti households in villas (*3*) uses all combinations of values of the explanatory variables within each house type for which at least one trip rate observation was available. As a consequence, in many cases the mean trip rates are based on small cell household frequencies. Because forecasts of future numbers of households for combinations of specified values of the explanatory variables would be needed, it is advantageous to have a fairly broad grouping of the values of the variables.

The grouping varies for the three nationality groups. For Kuwaiti households a total of 80 cross-classification cells are used, based on five levels of the number of children variable X_1, four levels of the car ownership variable X_2, and four levels of the number of adults variable X_3. The house type effect is ignored for Kuwaiti and Asian households, and the reason will become apparent later in the paper. For Arab households a total of 48 cross-classification cells are used, based on four, four, and three levels for the variables X_1, X_2, and X_3, respectively. The grouping is performed for each of the three house types. A total of 27 cross-classification cells are used for Asian households. The variations in the grouping respond to observed differences between the three nationality groups. For example, whereas the range of the number of children variable extends from 0 to 20 for Kuwaiti households, few Asian households have more than 4 children.

Table 2 presents the cross-classification table for Kuwaiti households as well as the observed trip rates for each household cross-classification cell. Table 3 gives the household frequencies and trip rates for Arab households. The table is organized in three parts corresponding to three housing types (villas, apartments, and others). The most frequent households are those with one to two and three to five adults, one to three and four to eight children, and one or two cars per household. Table 4 presents household frequencies for Asian households. There are small discrepancies between the totals of Tables 2 through 4 and the relevant numbers in Table 1. These occur because a few outlying observations were eliminated.

STATISTICAL MODELS

In the previous discussion, the presence of large structural differences among household groups reflecting nationality was stressed. House type is another qualitative factor that must be evaluated. In addition, previous analyses made using data for Kuwaiti households alone indicate that the quantitative variables X_1 (number of children in the household), X_2 (number of cars owned per household), and X_3 (number of adults in the household) should be included as possible variables for explaining variations in trip rate (*6*).

The statistical models use the following notation, where the superscript i indicates nationality, with 1, 2, and 3 denoting Kuwaitis, other Arabs, and Asians, respectively. Define Cell $(j, k, l, m)^{(i)}$ as the cell corresponding to the group of households with Nationality i, House Type j, kth observed value of X_1, lth observed value of X_2, and mth observed value of X_3. Note that the range of values for j (house type) depends

TABLE 2 TRIP RATES AND HOUSEHOLD FREQUENCIES FOR KUWAITI HOUSEHOLDS

Adults in the Household	Car Ownership	Number of Children Under 18 Years: 0	1-3	4-7	8-11	12-15
1 - 2	0 - 1	0.59(17)	0.62(68)	0.62(120)	0.43(35)	0.00(1)
	2 - 3	1.18(17)	1.08(148)	1.07(203)	0.55(38)	0.00(2)
	4 - 6	1.00(1)	0.86(7)	1.09(11)	0.50(2)	--
	7 - 9	--	--	--	--	--
3 - 5	0 - 1	0.86(14)	0.73(30)	0.54(57)	0.53(40)	0.45(11)
	2 - 3	1.27(33)	1.16(149)	0.93(207)	0.70(81)	0.48(19)
	4 - 6	1.91(11)	1.67(86)	1.65(83)	1.25(24)	2.11(9)
	7 - 9	3.00(1)	2.50(2)	2.17(6)	2.50(2)	--
6 - 8	0 - 1	2.00(2)	0.80(5)	1.88(8)	0.88(8)	0.33(6)
	2 - 3	2.33(3)	1.72(47)	1.61(62)	1.34(41)	1.20(20)
	4 - 6	3.00(12)	2.45(98)	2.20(107)	1.78(37)	1.80(10)
	7 - 9	--	3.00(15)	3.11(18)	4.33(6)	3.50(2)
9 - 12	0 - 1	3.00(1)	3.50(2)	1.00(2)	2.00(2)	0.00(1)
	2 - 3	2.00(1)	--	1.55(11)	1.25(4)	1.00(6)
	4 - 6	--	3.00(11)	3.58(26)	3.10(21)	2.70(10)
	7 - 9	5.00(1)	4.39(18)	4.35(17)	3.42(12)	4.00(3)

-- indicates 0 household frequency

TABLE 3 OBSERVED TRIP RATES AND HOUSEHOLD FREQUENCIES FOR ARAB HOUSEHOLDS

Adults in the Household	Car Ownership	Number of Children: 0	1-3	4-8	9+
		(Villas)			
1 - 2	0	0.00 (1)	0.50 (2)	1.00 (1)	--
	1	0.00 (2)	1.10 (10)	0.61 (18)	1.00 (1)
	2	1.00 (1)	1.00 (14)	0.92 (12)	--
	3+	--	2.50 (2)	1.00 (1)	0.00 (1)
3 - 5	0	3.00 (1)	--	0.75 (4)	1.00 (2)
	1	0.00 (1)	0.86 (7)	1.13 (8)	1.25 (4)
	2	--	1.46 (11)	1.40 (5)	1.00 (1)
	3+	2.00 (1)	1.00 (2)	2.00 (4)	2.00 (1)
6+	0	--	--	1.00 (1)	--
	1	--	--	0.75 (4)	--
	2	--	3.50 (4)	1.75 (4)	0.00 (1)
	3+	3.00 (2)	2.33 (3)	2.00 (2)	1.33 (3)
		(Apartments)			
1 - 2	0	0.96 (24)	0.87 (67)	0.59 (68)	0.50 (2)
	1	0.97 (68)	1.05 (395)	0.83 (43)	0.86 (14)
	2	1.38 (21)	1.29 (161)	1.10 (109)	0.67 (3)
	3+	--	1.33 (3)	1.33 (6)	--
3 - 5	0	2.36 (47)	0.96 (44)	1.00 (55)	0.33 (3)
	1	1.53 (51)	1.22 (232)	1.03 (266)	1.28 (7)
	2	2.14 (42)	1.45 (129)	1.37 (76)	1.50 (4)
	3+	2.00 (18)	2.13 (47)	1.80 (5)	1.00 (1)
6+	0	3.87 (15)	1.00 (4)	2.57 (7)	--
	1	2.53 (15)	2.25 (56)	1.54 (26)	2.00 (3)
	2	3.00 (5)	2.20 (35)	2.58 (19)	3.00 (3)
	3	4.33 (12)	2.23 (13)	2.56 (9)	2.00 (1)
		('Other' Housing Types)			
1 - 2	0	1.00 (3)	0.67 (12)	0.70 (27)	0.43 (7)
	1	1.50 (2)	0.95 (20)	0.77 (104)	0.88 (25)
	2	1.00 (1)	1.36 (11)	1.00 (14)	0.83 (6)
	3+	--	--	2.00 (1)	1.00 (1)
3 - 5	0	2.48 (29)	1.00 (2)	0.82 (22)	1.18 (11)
	1	3.67 (3)	1.21 (34)	0.81 (69)	0.97 (30)
	2	1.55 (11)	1.63 (16)	1.24 (33)	1.52 (23)
	3+	2.50 (4)	1.86 (7)	1.43 (7)	1.25 (8)
6+	0	5.31 (13)	1.00 (2)	0.71 (7)	1.25 (4)
	1	3.75 (4)	1.77 (13)	1.20 (10)	1.00 (3)
	2	6.00 (2)	1.67 (6)	1.45 (11)	2.00 (11)
	3+	3.00 (2)	2.63 (8)	2.15 (13)	2.08 (12)

- indicates 0 household frequency

TABLE 4 TRIP RATES AND HOUSEHOLD FREQUENCIES FOR ASIAN HOUSEHOLDS

Adults in the Household	Car Ownership	Number of Children: 0	1-3	4+
1 - 2	0	0.77 (13)	1.29 (28)	1.00 (7)
	1	1.46 (11)	1.12 (87)	0.79 (19)
	2+	1.00 (1)	1.23 (13)	0.00 (1)
3 - 5	0	2.47 (19)	2.05 (21)	1.33 (6)
	1	2.00 (15)	1.41 (39)	1.12 (17)
	2+	2.00 (9)	2.20 (20)	2.00 (3)
6+	0	2.73 (11)	1.00 (1)	0.50 (2)
	1	1.75 (4)	2.86 (7)	2.00 (3)
	2+	6.00 (6)	3.50 (2)	1.00 (1)

on the i (nationality group) that is being considered and that the ranges of values for k, l, and m depend on which nationality/house type grouping (i, j) is being considered.

For Cell $(j, k, l, m)^{(i)}$ in the sample, let

$N^{(i)}_{jklm}$ = number of households;
$Y^{(i)}_{jklmr}$ = number of work trips observed for the rth household, $r = 1, \ldots, N^{(i)}_{jklm}$; and
$Y^{(i)}_{jklm}$ = total number of observed work trips.

In the population, let

$\mu^{(i)}_{jklm}$ = mean number of household work trips and
$\sigma^{(i)}_{jklm}$ = standard deviation of number of household work trips.

Examination of the observed within-cell variations of trip rates about the cell means indicated that the variance/mean ratios are reasonably stable with values close to 1; this extends a similar finding for Kuwaiti households in villas (*3*). Such a relation between mean and variance suggests that the trip rates for Cell $(j, k, l, m)^{(i)}$ may be taken to have approximately a Poisson distribution with mean $\mu^{(i)}_{jklm}$. To model the dependence of the mean on house type and the three quantitative variables X_1, X_2, and X_3, a logarithmic link is assumed, taking

$$\begin{aligned}\log \mu^{(i)}_{jklm} = {} & \beta^{(i)}_{0j} + \beta^{(i)}_{1j}X_{1k} + \beta^{(i)}_{2j}X_{2l} + \beta^{(i)}_{3j}X_{3m} \\ & + \beta^{(i)}_{4j}X_{1k}X_{2l} + \beta^{(i)}_{5j}X_{1k}X_{3m} + \beta^{(i)}_{6j}X_{2k}X_{3m} \\ & + \beta^{(i)}_{7j}X_{1k}X_{2l}X_{3m} \qquad (1)\end{aligned}$$

This model ensures that the mean is positive and, taken with the Poisson assumption, a Poisson log-linear regression model. The model is general in form because it (*a*) includes interaction as well as linear effects for the variables X_1, X_2, and X_3 and (*b*) allows the regression coefficients to vary over both nationality and house type levels.

More concise model forms are also of interest. Although the order of model simplification is arbitrary, we shall adopt the procedure of first evaluating the house type effects within each nationality group by examining the reduced models

$$\begin{aligned}\log \mu^{(i)}_{jklm} = {} & \beta^{(i)}_{0j} + \beta^{(i)}_{1}X_{1k} + \beta^{(i)}_{2}X_{2l} + \beta^{(i)}_{3}X_{3m} \\ & + \beta^{(i)}_{4}X_{1k}X_{2l} + \beta^{(i)}_{5}X_{1k}X_{3m} + \beta^{(i)}_{6}X_{2k}X_{3m} \\ & + \beta^{(i)}_{7}X_{1k}X_{2l}X_{3m} \qquad (2)\end{aligned}$$

and

$$\begin{aligned}\log \mu^{(i)}_{jklm} = {} & \beta^{(i)}_{0} + \beta^{(i)}_{1}X_{1k} + \beta^{(i)}_{2}X_{2l} + \beta^{(i)}_{3}X_{3m} \\ & + \beta^{(i)}_{4}X_{1k}X_{2l} + \beta^{(i)}_{5}X_{1k}X_{3m} + \beta^{(i)}_{6}X_{2k}X_{3m} \\ & + \beta^{(i)}_{7}X_{1k}X_{2l}X_{3m} \qquad (3)\end{aligned}$$

Model 2 allows for house type to have a systematic effect on the average trip but takes the effects of X_1, X_2, and X_3 to be the same for all house types within each nationality groups. Model 3 implies that house type has no effect at all. Other reduced model forms of interest are obtained by setting subsets of the β terms on the right-hand sides of Models 1, 2, and 3 equal to zero.

Models are fitted by the method of maximum likelihood, and the statistical package GLIM provides a suitable way to perform the calculations (*8*). The goodness of fit of any model is measured by the deviance (*9,10*) and is given by

$$\begin{aligned}D = {} & 2\sum_j\sum_k\sum_l\sum_m\{Y^{(i)}_{jklm}\log[Y^{(i)}_{jklm}/N^{(i)}_{jklm}\mu^{(i)}_{jklm}] \\ & - [Y^{(i)}_{jklm} - \mu^{(i)}_{jklm}]\} \qquad (4)\end{aligned}$$

If the model is correct, D is approximately distributed as chi-square with degrees of freedom equal to the number of cells in the nationality group minus the number of parameters in the model. Changes in the deviances may be assessed using chi-square tests to test the goodness of fit of reduced models. Similar approaches can be used when the values for X_1, X_2, and X_3 are grouped into broader intervals; variable values are then put equal to the central values of the groups.

Finally, if the quantitative nature of X_1, X_2, and X_3 is ignored, the regression models may be replaced by ANOVA models. For example, Model 1 is replaced by the four-factor model

$$\begin{aligned}\log \mu^{(i)}_{jklm} = {} & \mu^{(i)} + A^{(i)}_j + B^{(i)}_k + C^{(i)}_l + D^{(i)}_m \\ & + (AB)^{(i)}_{jk} + (AC)^{(i)}_{jl} + (AD)^{(i)}_{jm} \\ & + (BC)^{(i)}_{kl} + (BD)^{(i)}_{km} + (CD)^{(i)}_{lm} \\ & + (ABC)^{(i)}_{jkl} + (ABD)^{(i)}_{jkm} + (ACD)^{(i)}_{jlm} \\ & + (BCD)^{(i)}_{klm} + (ABCD)^{(i)}_{jklm} \qquad (5)\end{aligned}$$

where A, B, C, and D are factors representing house type, number of children, car ownership, and number of adults, respectively. Models are again fitted by maximum likelihood, and assessments of fit are again based on changes in the deviances.

MODEL FITS FOR KUWAITI HOUSEHOLD DATA

Four house types were considered for Kuwaiti households: villas, NHA (government housing), apartments, and "others." As was indicated earlier, the untransformed trip rates could be taken to have approximately a Poisson distribution with a logarithmic link function for the means, the initial model being given by Model 1. The effect of house type was first examined by comparing Model 1 with the reduced Models 2 and 3 for $i = 1$.

For the ungrouped data, fits by maximum likelihood of Models 1, 2, and 3 for $i = 1$ gave deviances equal to 887.5, 904.1, and 905.1, with 971, 985, and 987 degrees of freedom, respectively. The χ^2 statistic for comparing Models 1 and 2 has value $904.1 - 887.5 = 16.6$ with 14 degrees of freedom. Similarly, the χ^2 statistic for comparing Models 2 and 3 has value 1.0 with 2 degrees of freedom. From χ^2 tables the upper 10 percent points are $\chi^2_{14}(0.9) = 21.06$ and $\chi^2_2(0.9) = 4.61$, so neither value is significant even at the 10 percent level. Thus, there is no evidence of any house type effect, and Model 3 is to be adopted.

Said et al. (*6*) pointed out that the models and associated statistical analyses are the same using the grouped data for

Kuwaiti households as using the ungrouped data. The only change is that the values of the explanatory variables are taken as the midpoints of the ranges selected for each variable. The broad findings are the same as for ungrouped data.

When reduced forms of Model 3 are examined, analysis using the grouped data indicates that only the three-variable interaction term $\beta_7 X_{1k} X_{2l} X_{3m}$ and the two-variable interaction term $\beta_5 X_{1k} X_{3m}$ can be excluded. The equation for the estimated cell mean trip rates was

$$\hat{\mu}^{(1)}_{jklm} = \exp(-0.483 - 0.064X_{1k} + 0.150X_{2l} + 0.120X_{3m} + 0.008X_{1k}X_{2l} - 0.006X_{2l}X_{3m}) \quad (6)$$

Using the ANOVA specification of GLM, if the house type effects are dropped and the chi-square tests are applied to the differences between the deviances of Model 5 with $i = 1$ and its reduced forms, the only significant interaction is $(CD)_{lm}$. The parameter estimates can be used to provide estimates of the cell mean trip rates using

$$\hat{\mu}^{(1)}_{jklm} = \exp(\hat{\mu} + \hat{B}_k + \hat{C}_l + \hat{D}_m + (\widehat{CD})_{lm}) \quad (7)$$

where from the computer fit of the model

$$\hat{\mu} = -0.311,$$
$$\hat{B}_k = (0, -0.170, -0.232, -0.454, -0.450),$$
$$\hat{C}_l = (0, 0.553, 0.479, 0.732),$$
$$\hat{D}_m = (0, 0.064, 0.711, 1.260), \text{ and}$$
$$(\widehat{CD})_{lm} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -0.083 & -0.237 & -0.880 \\ 0 & 0.494 & 0.165 & 0.050 \\ 0 & 0.612 & 0.300 & 0 \end{pmatrix}$$

These results for the regression and ANOVA models indicate that trip rates of Kuwaiti households increase with car ownership and number of adults but decrease with increasing number of children. The interaction effect of car ownership and number of adults is also significant in the two models. In the regression model the interaction effect of number of children and car ownership is also significant. The estimates of cell mean trip rates provided by the fitted regression and ANOVA models were close and in general agreed well with observed mean trip rates. The largest discrepancies were associated with outlying cells containing low frequencies.

MODEL FITS FOR ARAB HOUSEHOLDS

The effect of house type for Arab households ($i = 2$) has been tested with a procedure similar to that used for Kuwaiti households. When an underlying Poisson distribution with a logarithmic link was assumed for the trips, the deviances for the fits of Models 1, 2, and 3 using the ungrouped data were 511.4, 533.5, and 536.9, respectively. The χ^2 statistic for comparing Model 2 with Model 1 is therefore 22.1 with 15 degrees of freedom. This value, when compared with $\chi^2_{15}(95) = 24.9$, confirmed that the house type effect is significant.

Fits of the Poisson model for the means in Table 3 indicated that the most concise models for households in different housing types are of the following forms:

- For households in villas,

$$\hat{\mu}^{(2)}_{1klm} = \exp(-0.636 + 0.222X_{2l} + 0.117X_{3m}) \quad (8)$$

- For households in apartments,

$$\hat{\mu}^{(2)}_{2klm} = \exp(-0.127 - 0.10X_{1k} + 0.063X_{2l} + 0.148X_{3m} + 0.033X_{1k}X_{2l}) \quad (9)$$

- For households in other dwelling unit types,

$$\hat{\mu}^{(2)}_{3klm} = \exp(-0.193 - 0.057X_{1k} - 0.124X_{2l} + 0.231X_{3m} + 0.047X_{1k}X_{2l} - 0.018X_{1k}X_{3m}) \quad (10)$$

Comparisons among the models are difficult because they have markedly different forms. In all cases, the fitted models show the expected increase in mean trip rate with increasing number of adults. The sign and magnitude of the effect of car ownership depends strongly on whether the number of children has a significant effect on or interaction with other factors.

When the ANOVA specifications of GLM are used, the possible Poisson models for the means that were selected have the following forms:

- For households in villas,

$$\hat{\mu}^{(2)}_{1klm} = \exp(\hat{\mu} + \hat{C}_l + \hat{D}_m) \quad (11)$$

where

$$\hat{\mu} = -0.374,$$
$$\hat{C}_l = (0, -0.009, 0.422, 0.546), \text{ and}$$
$$\hat{D}_m = (0, 0.365, 0.642).$$

- For households in apartments,

$$\hat{\mu}^{(2)}_{2klm} = \exp(\hat{\mu} + \hat{B}_k + \hat{C}_l + \hat{D}_m + (\widehat{BC})_{kl}) \quad (12)$$

where

$$\hat{\mu} = .502,$$
$$\hat{B}_k = (0, -0.739, -0.807, -1.573),$$
$$\hat{C}_l = (0, -0.406, -0.052, 0.089),$$
$$\hat{D}_m = (0, 0.246, 0.761), \text{ and}$$
$$(\widehat{BC})_{kl} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0.656 & 0.505 & 1.393 \\ 0 & 0.450 & 0.457 & 1.279 \\ 0 & 0.528 & 0.652 & 0.651 \end{pmatrix}.$$

- For households in other dwelling unit types,

$$\hat{\mu}^{(2)}_{3klm} = \exp(\hat{\mu} + \hat{B}_k + \hat{C}_l + \hat{D}_m) \quad (13)$$

where

$$\hat{\mu} = .67,$$
$$\hat{B}_k = (0, -0.708, -0.978, -0.851),$$
$$\hat{C}_l = (0, -0.028, 0.164, 0.266), \text{ and}$$
$$\hat{D}_m = (0, 0.220, 0.689).$$

Compared with Models 8, 9, and 10, Models 11, 12, and 13 indicate that both regression and ANOVA approaches lead to fitted models having similar structure, with the exception of households in "other" dwelling unit types.

Table 5 gives the estimated mean trip rates of Arab households in apartments using the regression and ANOVA models. The table indicates that the two models produce similar estimates. The model estimates compare reasonably well with the observed trip rates of Table 3 for high-household-frequency cells. There are some discrepancies for outlying low-frequency cells.

MODEL FITS FOR ASIAN HOUSEHOLDS

Three house types were initially considered for Asian households. The chi-square test indicated that house type has no effect for these households. All Asian households in the home interview survey are therefore treated collectively in this section. The GLM analysis of trip rates using the data in Table 4 indicated that, when applying the regression model, only the main effects of X_1 (number of children in the household) and X_3 (number of adults in the household) were significant. The equation for the estimates of the cell means is

$$\hat{\mu}^{(3)}_{jklm} = \exp(0.086 - 0.109X_{1k} + 0.170X_{3m}) \quad (14)$$

In the case of the ANOVA model, the three main effects were significant, but there were no significant interactions. The estimated means are

$$\hat{\mu}^{(3)}_{jklm} = \exp(\hat{\mu} + \hat{B}_k + \hat{C}_l + \hat{D}_m) \quad (15)$$

where

$$\hat{\mu} = .306,$$
$$\hat{B}_k = (0, -0.143, -0.52),$$
$$\hat{C}_l = (0, -0.102, 0.228), \text{ and}$$
$$\hat{D}_m = (0, 0.430, 0.853).$$

The estimated mean trip rates are given in Table 6 on the basis of Models 14 and 15. The table indicates that the regression and ANOVA model estimates are consistent. The regression model trip rate estimates do not vary for the different car ownership groups, because Model 14 implies that this factor has no effect. This is unlike the ANOVA models, which indicate that some car ownership effects exist. The structure of the two model fits is consistent with the socioeconomic characteristics of Asian households in Kuwait, which are dominated by low-income working adults with low car ownership rates and significant reliance on public transport for work trip purposes.

TABLE 5 ESTIMATED MEAN TRIP RATES BASED ON FITS USING GROUPED DATA OF ARAB HOUSEHOLDS (APARTMENTS) (POISSON MODEL WITH LOGARITHMIC LINK)

Adults in the Household	Car Ownership	Number of Children: 0		1-3		4-6		9+	
		(i)	(ii)	(i)	(ii)	(i)	(ii)	(i)	(ii)
1 - 2	0	1.10	1.65	.90	.79	.60	.74	.42	.34
	1	1.17	1.10	1.02	1.01	.78	.81	.61	.92
	2	1.25	1.57	1.17	1.18	1.01	1.11	.90	1.17
	3+	1.35	1.81	1.36	1.46	1.39	1.55	1.41	0.88
3 - 5	0	1.59	2.11	1.30	1.01	.87	.94	.61	.44
	1	1.70	1.41	1.48	1.30	1.13	1.04	.89	1.18
	2	1.81	2.01	1.69	1.50	1.47	1.41	1.30	1.50
	3+	1.95	2.31	1.97	1.87	2.01	1.98	2.04	1.12
6+	0	2.38	3.54	1.95	1.69	1.30	1.58	.91	0.73
	1	2.53	2.36	2.21	2.17	1.69	1.74	1.33	1.97
	2	2.70	3.36	2.52	2.52	2.19	2.37	1.94	2.50
	3+	2.91	3.87	2.94	3.13	2.99	3.31	3.04	1.88

(i) Regression Model
(ii) ANOVA Model

TABLE 6 OBSERVED AND ESTIMATED MEAN TRIP RATES BASED ON FITS USING GROUPED DATA OF ASIAN HOUSEHOLDS (POISSON MODEL WITH LOGARITHMIC LINK)

Adults in the Household	Car Ownership	Number of Children: 0		1-3		4+	
		(i)	(ii)	(i)	(ii)	(i)	(ii)
1 - 2	0	1.41	1.36	1.13	1.18	.82	.81
	1	1.41	1.23	1.13	1.06	.82	.73
	2+	1.41	1.71	1.13	1.48	.82	1.01
3 - 5	0	2.15	2.09	1.73	1.81	1.25	1.24
	1	2.15	1.89	1.72	1.64	1.25	1.12
	2+	2.15	2.62	1.73	2.27	1.25	1.56
6+	0	3.46	3.19	2.78	2.76	2.00	1.89
	1	3.46	2.88	2.78	2.49	2.00	1.71
	2+	3.46	4.00	2.78	3.47	2.00	2.38

(i) Regression Model
(ii) ANOVA Model

USE OF DEVELOPED TRIP RATES IN PLANNING

The fitted models given in the preceding three sections can be used to construct charts giving estimates of the mean trip rates that allow planners to forecast total numbers of morning work trips for individual zones. Figure 2 shows such a chart based on estimates from the fitted ANOVA models. For convenience, only Arabs in apartments are included. In a more complete representation, Arab households in villas and "other" housing types would be needed.

The use of the chart is straightforward. Estimates of frequencies of households classified by nationality, car ownership, number of children, and number of adults are needed as input. Multiplication of these frequencies by the associated estimated means read from the chart and summation of the resulting products give the required estimate of trips in the zone.

Morning work trips in 1995 for two typical zones, Jahra and Hawalli, are estimated here for demonstration purposes. Jahra is typical of zones dominated by Kuwaiti households. The expected numbers of Kuwaiti and non-Kuwaiti households in 1995 are 15,726 and 9,270, respectively. These estimates are based on the cohort survival technique for population forecasting using the 1985 data base. The household size and relative proportions of households by size are assumed to remain stable over the 1985–1995 forecasting period. The cross-classification of these households is shown in Table 7.

Hawalli is typical of zones dominated by non-Kuwaiti households. The zone is expected to have 33,324 Arab non-Kuwaiti households in apartments, 3,200 Arab households in

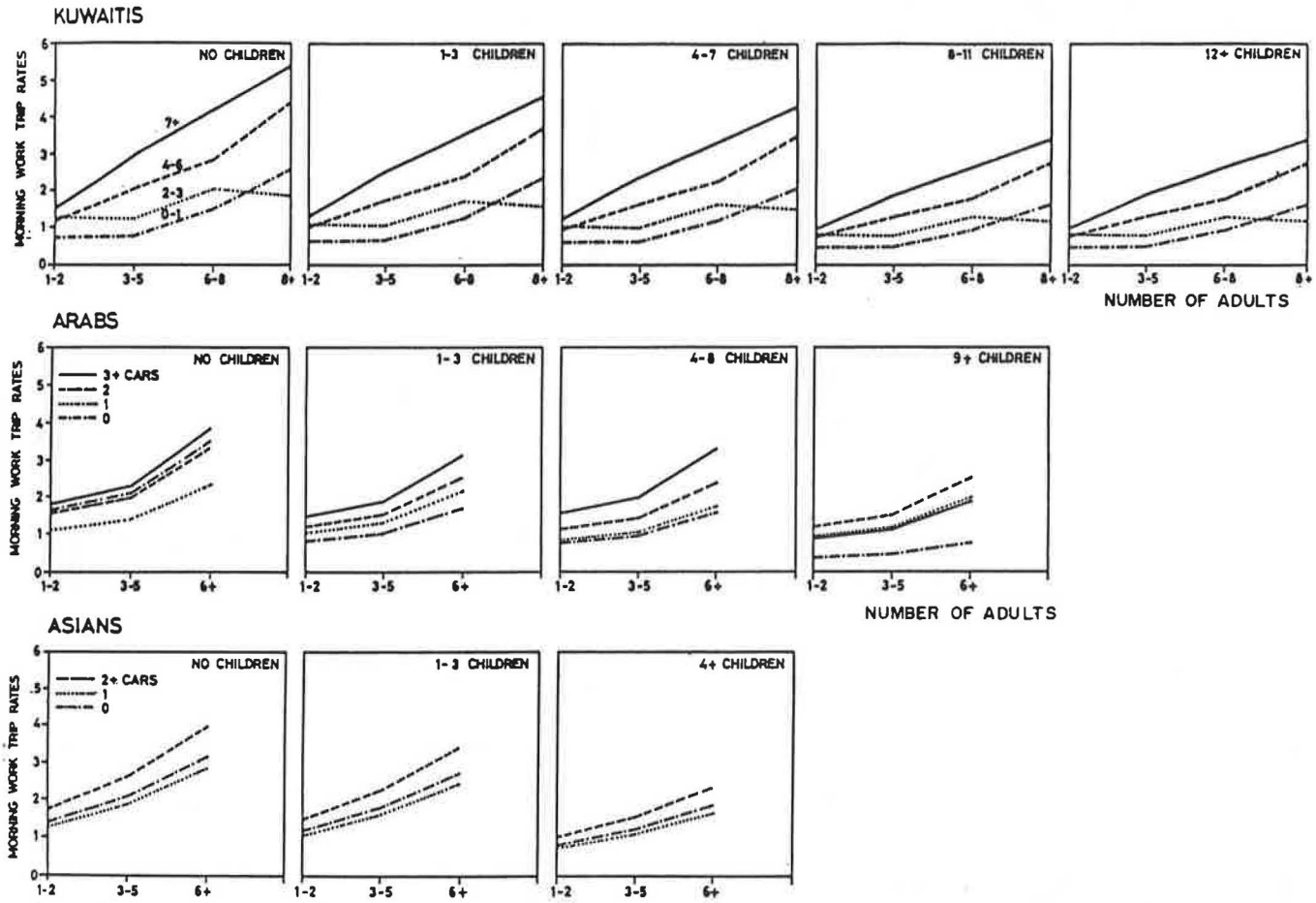

FIGURE 2 Work trip rates of households of different nationalities with levels of three explanatory variables based on fits of ANOVA models.

"other" housing types, and 800 Kuwaiti households. Table 8 shows the 1995 household frequencies for Arab households in apartments for this zone.

The expected numbers of morning work trips that can be generated by the two zones in 1995 are 29,060 and 51,250, respectively. These numbers were estimated on the basis of household frequencies and the appropriate trip rates from Figure 2 supplemented with a similar chart for Arab households in "other" housing types.

STABILITY OF ESTIMATED WORK TRIP RATES

Most trip generation studies assume a degree of stability over the forecasting period of the trip generation equation developed for the calibration year. The stability of the household level trip generation equations could be verified through (*a*) the stability of the estimated values of the regression coefficients for the household groups or the effects estimated when the ANOVA specification of GLM is used and (*b*) the stability of the ranges of the underlying explanatory variables in these equations.

The stability of the regression coefficients or the ANOVA effects is the subject of current research using data collected in a 1977 home interview survey along with the 1988 data. The differences between the magnitudes of the coefficients based on the two data sets will be studied.

Confirmation of stability in the ranges of underlying explanatory variables in the household-level trip generation equations is relevant to this investigation, because these equations can only be used with confidence over roughly the range of the explanatory variables used in the calibration. Large structural shifts in the magnitude of these variables clearly could introduce forecasting errors.

The regression and ANOVA models used three variables: number of children, number of adults, and car ownership. The first two variables are studied jointly through the household size variable, because there are no historical records of households classified by numbers of children and adults.

Household Size

The top of Figure 3 uses census data to show the percentage of households by size of household for Kuwaitis between 1970 and 1985. The figure indicates that household size distribution has been reasonably stable. This will probably be the case for the short and medium term. The trend of large household size because of increased number of children is being compensated for by the increasing presence of extended families where married sons remain within the parent household.

TABLE 7 HOUSEHOLD FREQUENCIES FOR TYPICAL KUWAITI ZONE IN 1995: JAHRA

Adults	Car Ownership	0	Number of Children 1-3	4-7	8-11	12-15
1 - 2	0 - 1	0	336	1193	796	0
	2 - 3	0	192	530	597	66
	4 - 6	0	0	0	0	0
	7 - 9	0	0	0	0	0
3 - 5	0 - 1	66	265	796	928	265
	2 - 3	0	336	2254	928	463
	4 - 6	0	133	265	192	66
	7 - 9	0	0	0	0	0
6 - 8	0 - 1	66	66	66	133	66
	2 - 3	0	398	663	597	463
	4 - 6	66	336	530	265	398
	7 - 9	0	66	0	0	0
9 - 12	0 - 1	0	0	66	0	0
	2 - 3	0	0	66	133	192
	4 - 6	0	66	66	398	66
	7 - 9	0	66	0	0	0

(b) Arab Households in Apartments

Adults	Car Ownership	0	Number of Children 1-3	4-8	9+
1 - 2	0	274	549	0	0
	1	0	823	1646	549
	2	0	0	549	0
	3+	0	0	0	0
3 - 5	0	823	0	0	0
	1	0	0	0	0
	2	0	0	274	0
	3+	0	274	0	0

(c) Arab Households in 'Others' Housing

1 - 2	0	0	27	239	80
	1	0	53	372	266
	2	0	27	53	80
	3+	0	0	0	0
3 - 5	0	0	0	159	159
	1	0	133	478	345
	2	27	53	133	186
	3+	0	0	27	53
6+	0	0	0	53	53
	1	0	0	80	53
	2	0	0	53	80
	3+	0	27	80	80

TABLE 8 HOUSEHOLD FREQUENCIES FOR TYPICAL NON-KUWAITI ZONE IN 1995: HAWALLI (ARAB HOUSEHOLDS IN APARTMENTS)

Adults in the Household	Car Ownership	0	Number of Children 1-3	4-8	9+
1 - 2	0	143	620	620	0
	1	810	3909	4624	143
	2	381	1287	1430	0
	3+	0	95	143	0
3 - 5	0	238	1049	1049	0
	1	1001	3289	4290	95
	2	715	1907	1287	48
	3+	143	620	48	0
6+	0	0	95	95	0
	1	191	953	286	95
	2	48	715	286	0
	3+	143	334	95	0

The bottom of Figure 3 shows household size data for non-Kuwaitis; data for separate non-Kuwaiti nationality groups are not available. These households showed some instabilities in household size over the first 5 years, although the scale of change was not large. Household size remained relatively stable from 1975 to 1985 and can be expected to remain so.

Household Car Ownership

Data on household car ownership in Kuwait are limited. Only the 1970 census contained this information. The two home interview studies of 1977 and 1988 had questions on car ownership for a sample of households. These studies, along with the 1970 census data, are used in this section.

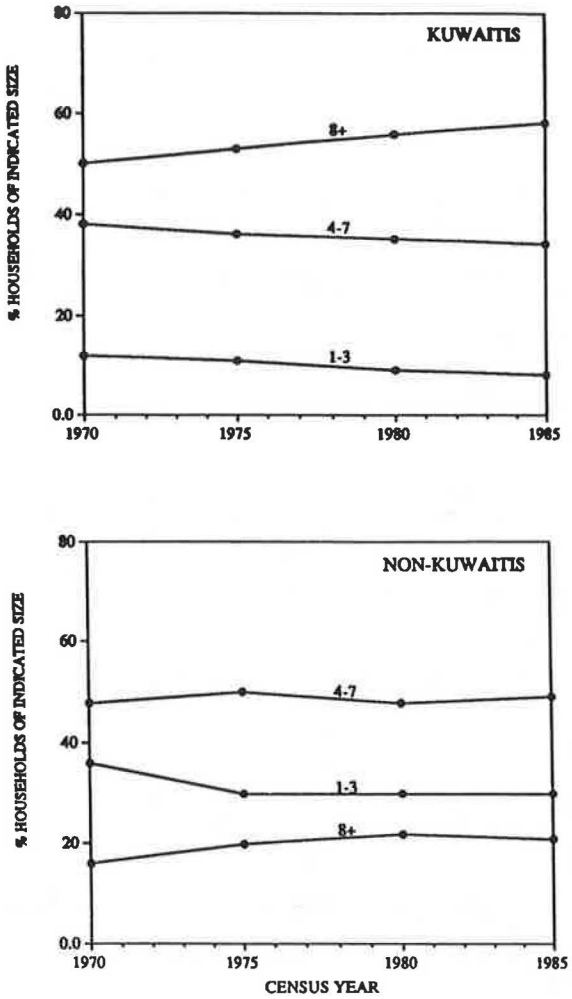

FIGURE 3 Percentage of Kuwaiti and non-Kuwaiti households by size, 1970–1985.

Figure 4 shows the trend in car ownership for Kuwaiti and non-Kuwaiti households. The top shows that for Kuwaiti households the proportion of one-car households declined from 70 percent in 1970 to 17 percent in 1980 and that the proportion of three- and four-or-more-car households rose noticeably between 1977 and 1988. This period corresponds to the era of rising national and per capita income because of the rise in the price of oil, Kuwait's main export, in the early 1970s. It is thought that current car ownership levels reflect saturation levels and that significant increases in the proportion of households with three or more cars are unlikely.

The bottom of Figure 4 shows reasonable stability in the proportions of one-, three-, and four-or-more-car households among non-Kuwaitis. Major shifts occurred in the proportions of zero- and two-car households. Zero-car households declined from 40 percent in 1970 to 15 percent in 1988, whereas the proportion of two-car households increased from 5 percent in 1970 to 20 percent in 1988. The socioeconomic characteristics of non-Kuwaitis and the regulations that govern the issue of operator licenses to various non-Kuwaiti occupational groups suggest that the current household car ownership structure will not change significantly in the short run.

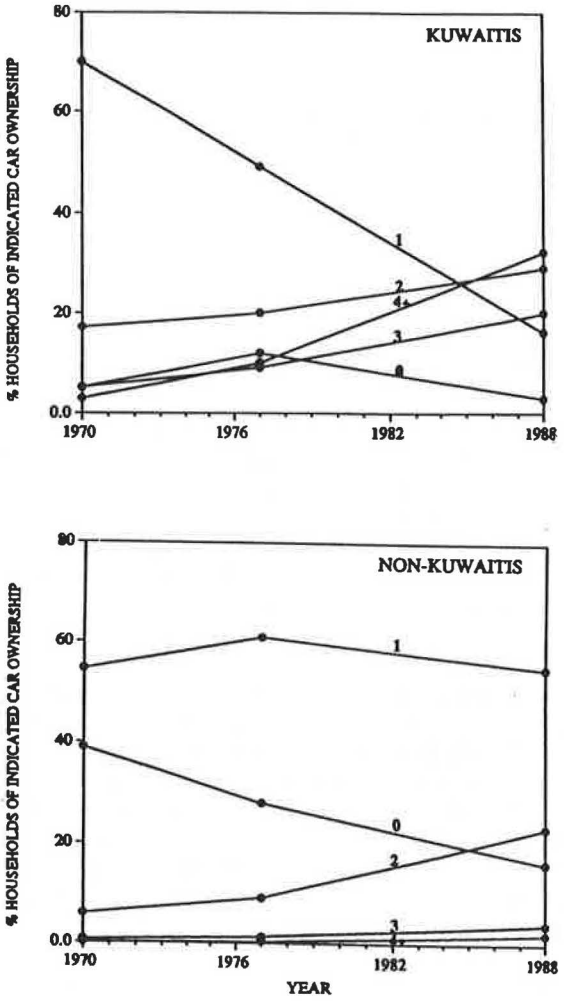

FIGURE 4 Percentage of Kuwaiti and non-Kuwaiti households by car ownership level, 1970–1988.

CONCLUDING REMARKS

The objective of this paper was to present the findings of a 3-year study of trip rates of households in Kuwait. The study led to a proposed general modeling approach that can be applied to a variety of urban areas. In the initial stages of the study, the intent was to refine one of the routinely used trip generation procedures. Further analysis indicated that several features in Kuwait, such as great variations in size in already-large households and variations in car ownership up to levels not common in many urban areas, would make the routine use of one of these techniques inappropriate. Other major variations among these households in characteristics such as income, labor force participation rates, and occupation and variations among subgroups based on nationality and house type were also shown. Situations like these may exist in other areas. Cities in neighboring countries in the Persian Gulf region provide good examples.

When the approach proposed in this paper is to be used for other urban areas, some investigative work will be warranted. Its purpose will be to (*a*) identify potential qualitative variables to be considered and verify their significance, (*b*) identify potential quantitative variables to be used in the analysis and select the most appropriate, (*c*) establish the relation of the variance-mean pattern of the individual household trips to aid in the selection of their underlying distribution, and (*d*) construct the groupings of the quantitative variables. Once these steps have been completed, model fits can be made using the GLM framework.

The specific conclusions of this paper are as follows:

1. There are marked variations among household groups in Kuwait.
2. House type is significant in the case of Arab households as a qualitative variable but not significant in the case of Kuwaiti and Asian households.
3. Regression and ANOVA models produce similar mean trip rate estimates, and these estimates agree with observed trip rates for high-frequency classification cells.
4. The number of adults and car ownership were found to be important variables in explaining variations in observed trip rates.
5. The number of children was found to be significant in most cases, with trip rates tending to decrease as the number of children increased.

ACKNOWLEDGMENTS

This research was funded by Kuwait University. Data were provided by Kuwait Municipality. Mrs. Reda Alfi typed the manuscript.

REFERENCES

1. G. M. Said. Work Trip Rates and the Socioeconomic Characteristics of Households in Kuwait. *Australian Road Research*, Vol. 20, No. 3, Sept. 1990.
2. P. Stopher and K. McDonald. Trip Generation by Cross-Classification: An Alternative Methodology. In *Transportation Research Record 944*, TRB, National Research Council, Washington, D.C., 1983, pp. 84–91.
3. G. M. Said and D. H. Young. A General Linear Model Framework for Estimating Work Trip Rates of Households in Kuwait. *Transportation Research A*, Vol. 24, No. 3, May 1990, pp. 187–200.
4. R. Dobson. The General Linear Model Analysis of Variance: Its Relevance to Transportation Planning and Research. *Socioeconomic Planning Science*, Vol. 10, 1976, pp. 231–235.
5. J. M. Rickard. Application of a Generalized Linear Modelling Approach to Category Analysis. *Transportation Planning and Technology*, Vol. 13, No. 2, 1989, pp. 121–129.
6. G. M. Said, D. H. Young, and H. K. Ibrahim. Statistical Models for Work Trips Rates of Kuwaiti Households. Submitted to *Journal of Kuwait University (Science)*, 1990.
7. B. G. Hutchinson and G. M. Said. Spatial Differentiation, Transport Demands and Transport Model Design in Kuwait. *Transport Reviews*, Vol. 10, No. 2, April 1990, pp. 91–112.
8. R. Baker and J. Nelder. *GLIM Manual (Release 3)*. Numerical Algorithms Groups, Oxford, United Kingdom, 1978.
9. A. J. Dobson. *An Introduction to Statistical Modelling*. Chapman and Hall, London, 1983.
10. P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983.

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

Circulator/Distributor Model for the Chicago Central Area

DAVID L. KURTH AND CATHY L. CHANG

The city of Chicago is evaluating alternative methods for providing for the distribution of commuters to and workers, visitors, and residents in the vibrant and growing central area. A detailed travel model has been calibrated to project ridership on the various circulator/distributor alternatives. The model is based on the Downtown People Mover System travel demand models developed for Los Angeles, Detroit, and Miami. The Chicago central area and the calibration of the circulator/distributor model are described. The calibrated model coefficients are compared with the coefficients used for the Los Angeles, Detroit, and Miami models. Finally, some brief recommendations regarding future circulator/distributor model development and research efforts are presented.

Chicago's central area is one of the most significant, highly concentrated, and exciting activity centers in the Midwest. Historically, the downtown coincided with the elevated loop structure. It is roughly bordered by Wacker Drive on the north and west, Michigan Avenue on the east, and the Congress Expressway on the south. This vibrant area is continually expanding. In the past 20 years significant new multiuse towers have punctuated the skyline. Growth is occurring along the North Michigan Avenue corridor and in areas south and west of the traditional Loop in office, retail, and residential space.

The existing transportation system was planned to serve destinations within the traditional Loop area. In this compact core area, most transit riders were able to walk from their alighting station to their destination. As the central area grows in shape and size, it becomes more and more difficult for the existing transit system to serve all destinations adequately. Expanded development patterns coupled with ever-increasing congestion add to travel times. It is becoming apparent that an expanded transit system is needed to serve the expanding central area, which now stretches from North Avenue on the north to Cermak Road on the south and from Halsted Street on the west to Lake Michigan on the east. It is approximately 4 mi long by 2 mi wide. Although the area is not completely developed (i.e., growth occurs in spots), these boundaries indicate the limit of current development trends.

The expanded central area is served by various modes of transportation. Commuter rail lines, rapid rail lines, and buses provide service to and through the area. In addition, taxis and private automobiles are prevalent. Because of the expansion of central area, it is no longer reasonable to expect all persons to walk from transit service or parking to their destinations, which may be as far as 2 mi. Thus, the concept of a central area circulator or downtown people mover (DPM) has evolved. The circulator system would provide quick and convenient access within the expanded central area. The proposed system would consist of either buses (the TSM alternative) or light rail transit (LRT).

The Chicago Central Area Circulator Study required a detailed model capable of projecting ridership for the extensive transit system in the central area and the proposed alternative network configurations. The study focused on modeling two major types of trips: those made from central area commuter rail stations to destinations within the central area (distributor trips) and those made wholly within the central area by residents, workers, and visitors to the central area (circulator trips). The modeling of these two types of trips led to the typical model form for modeling DPM systems (*1–3*).

In addition to these types of trips, there was concern about the proper assignment of trips that entered the central area on Chicago Transit Authority (CTA) bus and rapid rail lines. This concern, coupled with the modeling of the distributor trips, required an extensive interface with the regional transportation model maintained by the Chicago Area Transportation Study (CATS). CATS provided basic travel data for regional trips destined for the central area that used one of the six Metra commuter rail stations located in the central area and trips that entered the central area on CTA bus and rapid rail.

The remainder of this paper focuses on model form, model calibration and validation, and lessons learned from the calibration of the model. The final section discusses potential improvements and research for central area circulator/DPM models. The actual model results and recommendations regarding alternatives are not discussed. Readers who are interested in the results and recommendations should contact the City of Chicago Department of Planning for copies of reports written for the Chicago Central Area Circulator Alternatives Analysis.

MODEL FORM

Model form generally refers to the specific mathematical models and relationships used to estimate trip generation, trip distribution, mode choice, and assignment. These are important and will be discussed as appropriate. However, for modeling travel in the central area, the level of detail of the zone and network structure and the market segments modeled are of equal importance. If they are not well defined at the outset of a central area circulator study, it will be difficult to produce reasonable results even with the best travel models available.

Barton-Aschman Associates, Inc., 820 Davis Street, Evanston, Ill. 60201.

Zone Structure

A detailed zone structure was developed for the Central Area Circulator Study to properly analyze the trade-offs between walking, taking a taxi, and taking another transit vehicle or the circulator/distributor system from the line-haul transit system to the central area destination. Whereas the detailed zone structure was crucial for improving ridership forecasts, increases in the number of zones increased the difficulty of producing socioeconomic projections for those zones. Thus, there was also a practical trade-off restricting the level of detail used in the zone structure.

Figure 1 shows the zone structure used for the Central Area Circulator Study. The 8-mi^2 area modeled included 406 internal zones, 49 "transit external stations," and 51 "automobile external stations." Within the Chicago Loop, most of the zones were block-level zones. Outside the Loop, increasingly large zones were used. Transit external stations were established wherever transit lines crossed the boundary of the study area and at the six central area Metra commuter rail stations: Chicago & North Western Station, Union Station, LaSalle Street Station, and the Randolph Street, Van Buren Street, and Roosevelt Road Metra Electric Stations. The external transit stations were mode specific. If an express bus and several local bus routes crossed the boundary of the study area on the same street, two external stations were established—one for the express bus line and one for local bus lines. This process prevented spurious transfers between modes at external stations.

The commuter rail stations were handled as special cases in modeling travel to the central area. As will be discussed, detailed mode choice models were developed to estimate the number of trips by egress mode from the commuter rail stations to the central area destinations. Four of the six stations are terminals on commuter rail lines, and each of the commuter rail stations is a major transfer point. Commuters are forced to make a decision and are generally offered a number of choices for traveling to their destination at each of the stations. In contrast, bus passengers and rapid rail passengers have multiple points at which they can make a choice regarding travel from the main line-haul mode (crossing the study area boundary) to their destination. Because of the complexity of the choices, the trips made by bus and rapid rail passengers from external transit stations to central area destinations were handled simply through route choice (i.e., transit assignment).

Network Structure

Network coding was crucial to the accurate modeling of ridership on the alternative circulator/distributor systems. EMME/2 was used to code an integrated transportation network for the central area. It would have been possible to use other microcomputer- or mainframe-based transportation planning packages to perform the detailed coding; however, the coding was simplified greatly by using an interactive and integrated network editor. An integrated network was crucial because consistent highway, transit, and walk networks were required to build automobile and taxi paths, transit paths, and walk paths for the area being modeled.

The transportation network was coded as accurately as possible. This included coding of distances to the nearest 0.01 mi; coding an extensive walk network, including all sidewalks (streets) in the area, as well as pedestrian-only links; coding stair links to represent the time necessary to walk from the center of a subway or elevated platform to the street level; and coding access/egress links from the center platform of each commuter rail station to each possible exit from the station. This level of detail in network coding was necessary because walk was one of the possible modes considered. For example, had the network been coded to the nearest 0.1 mi, substantial differences in results could be obtained with little difference in actual travel times. Consider the estimation of walk travel times from a commuter rail station to two adjacent zones, one 0.24 and the second 0.26 mi from the station. If the network had been coded to only the nearest 0.1 mi, the distances to the two stations would have been coded as 0.2 and 0.3 mi. The modeled travel times using a 3-mph walk speed would have been 4.0 min to the first zone and 6.0 min to the second zone. Using a network coded to the nearest 0.01 mi would result in modeled travel times of 4.8 and 5.2 min to the two example zones.

Market Segments Modeled

Travel models were developed for two times of day: the morning peak period and midday. Two major types of trips were considered: the distribution of regional transit riders entering the central area on commuter rail to their destinations and the circulation of central area residents, workers, and visitors. Six main market segments resulted:

- The morning peak-period distributor market segment (regional commuter rail passengers making work trips to the central area),
- The morning peak-period circulator market segment (central area residents making work trips),
- The midday distributor market segment (regional commuter rail passengers making nonwork trips to the central area),
- The midday worker circulator market segment (central area workers making midday nonwork trips),
- The midday nonworker circulator market segment (central area visitors making midday nonwork trips), and
- The midday resident circulator market segment (central area residents making midday nonwork trips).

The afternoon peak-period distributor and circulator trips were considered to be the reverse of morning peak-period trips for the corresponding market segments.

In addition to these market segments, peak-period and midday trips entering the central area by bus and rapid rail (subway and elevated) were included in the assignment of transit trips. However, they were simply assigned from their "ports-of-entry" to their destinations. Modes used to reach their destinations were based strictly on path choice.

In the application of the models, trips to and from the central area were estimated by CATS using the regional travel model. CATS estimated new trip tables for each model alternative. The trip tables were all based on the same trip

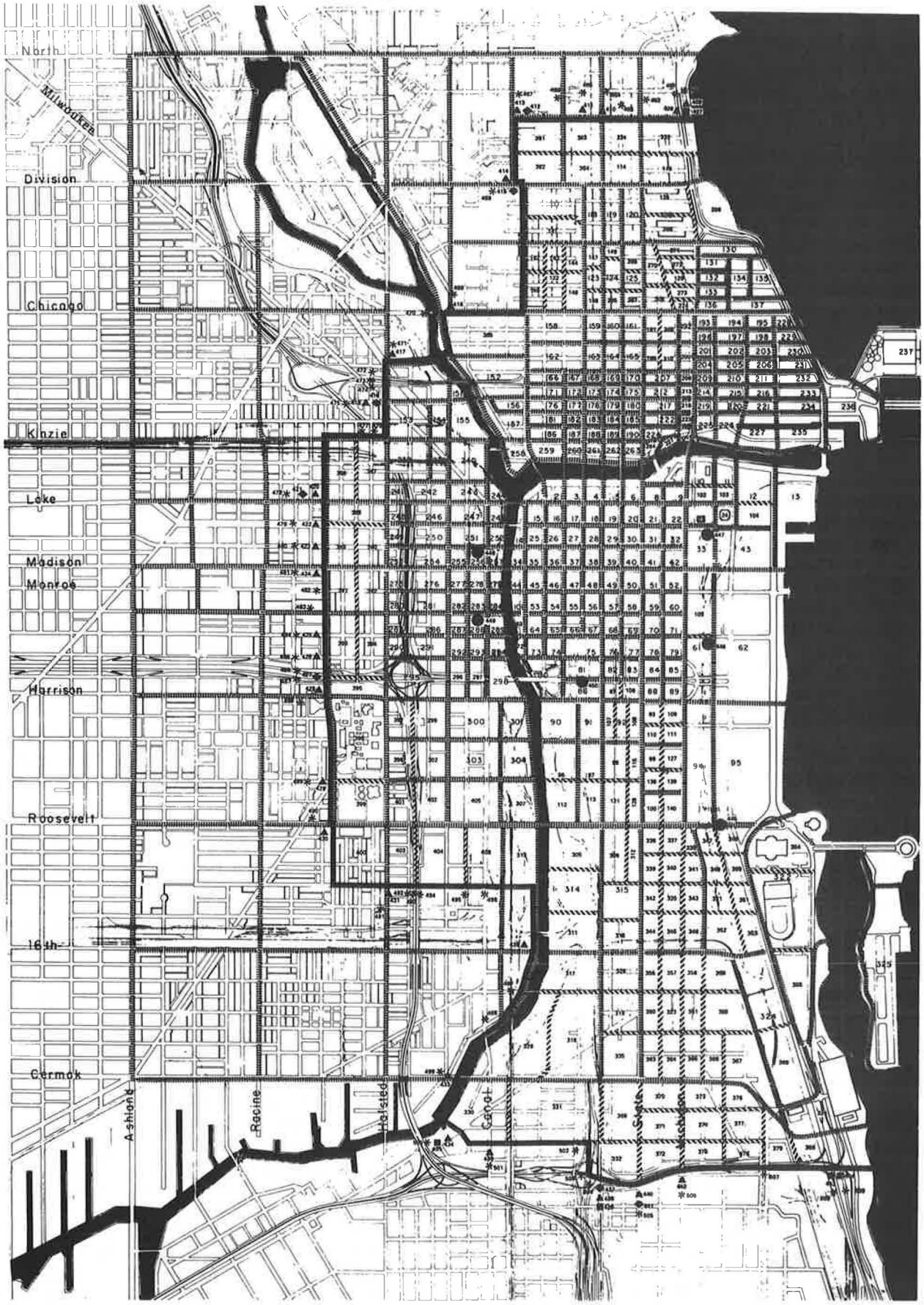

FIGURE 1 Structure of CBD distributor study zone.

distribution results. However, the "external-internal" trips to the central area varied for each alternative, because regional submode shares (bus, rapid rail, and commuter rail) to the central area were affected by the various alternatives.

Modes Considered

Four main modes were considered in the estimation of circulator/distributor travel in the central area: walk, transit, taxi, and automobile. The actual modes available depended on the market segment being considered. The walk, transit, and taxi modes were considered for all market segments. The automobile mode was considered for only those market segments that could have automobiles available—central area residents (peak-period and midday trips), central area workers (midday circulator trips only), and central area nonworkers (midday circulator trips only).

The actual mode considered for the central area circulator/distributor system was LRT. Figure 2 shows the full-build LRT system proposed as the most extensive of the alternatives for the central area circulator system. For most of the market segments, the proposed LRT system was not modeled as a separate, distinct mode in the mode choice models. Rather, the system was considered to be the same main transit mode as the rapid rail, local bus, shuttle bus, and express bus systems serving the central area. This was a departure from many previous DPM modeling efforts.

In all cases, only one set of transit travel paths (skims) was built from the coded network. The LRT system was used in these transit paths only if it was a part of the shortest path. However, EMME/2 determines transit travel times on the basis of multiple transit paths. If two or more efficient transit paths are available between two zones, the transit travel times posted on the transit skims are a composite of the efficient paths (*4*). This is done even if the different transit paths use different transit submodes (e.g., one path uses bus, a second uses an express bus, and a third uses LRT).

The use of the EMME/2 multipath transit model is a departure from previous DPM modeling efforts using UTPS or other all-or-nothing transit path–building algorithms. The EMME/2 multipath algorithm could slightly bias the results toward an LRT system. If two points were served by both bus and LRT and the LRT were a slightly slower mode, EMME/2 would still include the LRT in the impedance calculation and assign trips to the LRT. In contrast, all-or-nothing transit path–builders would ignore the LRT and assign all trips to the bus system. However, the reverse situation could also be true: the LRT might be the slightly faster mode for the interchange.

For midday worker circulator trips and midday nonworker circulator trips, LRT was considered more attractive than bus or rapid rail in the mode choice models. If the circulator was used on the shortest transit travel path, a positive circulator mode bias was added to the transit utility in proportion to the amount of in-vehicle travel time spent on LRT in comparison with the total transit in-vehicle travel time. The effect of this bias on LRT ridership was probably less than that obtained by considering LRT to be a separate, distinct mode, as in previous DPM modeling efforts.

Separate shortest–travel-time paths were built for the walk, transit, taxi, and, when necessary, automobile modes. When transit paths were built, steps were taken to minimize the number of walk-only paths. This resulted in some very short transit paths that, in most regional modeling efforts, would have been considered illogical. However, because walk was considered a competing mode in the mode choice models, this was a desirable result. The mode choice models were "allowed" to determine which short transit (or taxi or automobile) trips were not likely compared with the walk mode.

The building of separate paths obviated the need to consider fares in the path-building process. There was no need to exclude short, illogical transit trips from the path-building process for the reasons just mentioned. Fares and other travel costs were considered explicitly in the circulator/distributor models.

MODEL CALIBRATION PROCESS AND RESULTS

Disaggregate data were not available to perform a rigorous calibration of the logit models used to model the central area distributor and circulator mode use. Thus, an existing model was borrowed and adjustments were made to match observed aggregate data. DPM models from Los Angeles, Detroit, and Miami were considered as donor models. Both the Detroit and Miami models are based on the original Los Angeles DPM modeling work performed in the late 1970s. Apparently the Detroit model was calibrated in much the same way as the Chicago model—an adjustment of model constants and coefficients to match aggregate mode shares. The Miami model apparently used a slightly more rigorous calibration procedure. At the least, the calibration of the Miami model was based on observed travel behavior with a circulator/distributor system in place. However, the basic form of the Miami model was not changed from the form originally developed for Los Angeles.

The Detroit model was selected as the donor model for Chicago for several reasons. First, like Chicago, Detroit is a large northern city, and its climate resembles Chicago's more closely than Miami's does. Second, the implied values-of-time from the Detroit model appeared to be more reasonable than the implied values-of-time from Miami. Finally, because the Miami and Detroit models were both derived from the same root model, Los Angeles, they offered the same benefits in terms of model form.

Adjustments were made to alternative specific constants and to system coefficients. Good aggregate data existed to guide the adjustments. Most of the data were collected as part of ongoing monitoring processes conducted by CATS and Metra. The data included

- Egress mode shares from five of the six commuter rail stations;
- Egress mode shares by distance from the commuter rail stations;
- Average egress trip length from the commuter rail stations by mode;
- Mode shares for midday trips made by central area workers, nonworkers, and residents;
- Average trip lengths for midday trips made by central area workers, nonworkers, and residents; and
- Mode shares and average trip lengths for peak trips made by central area residents.

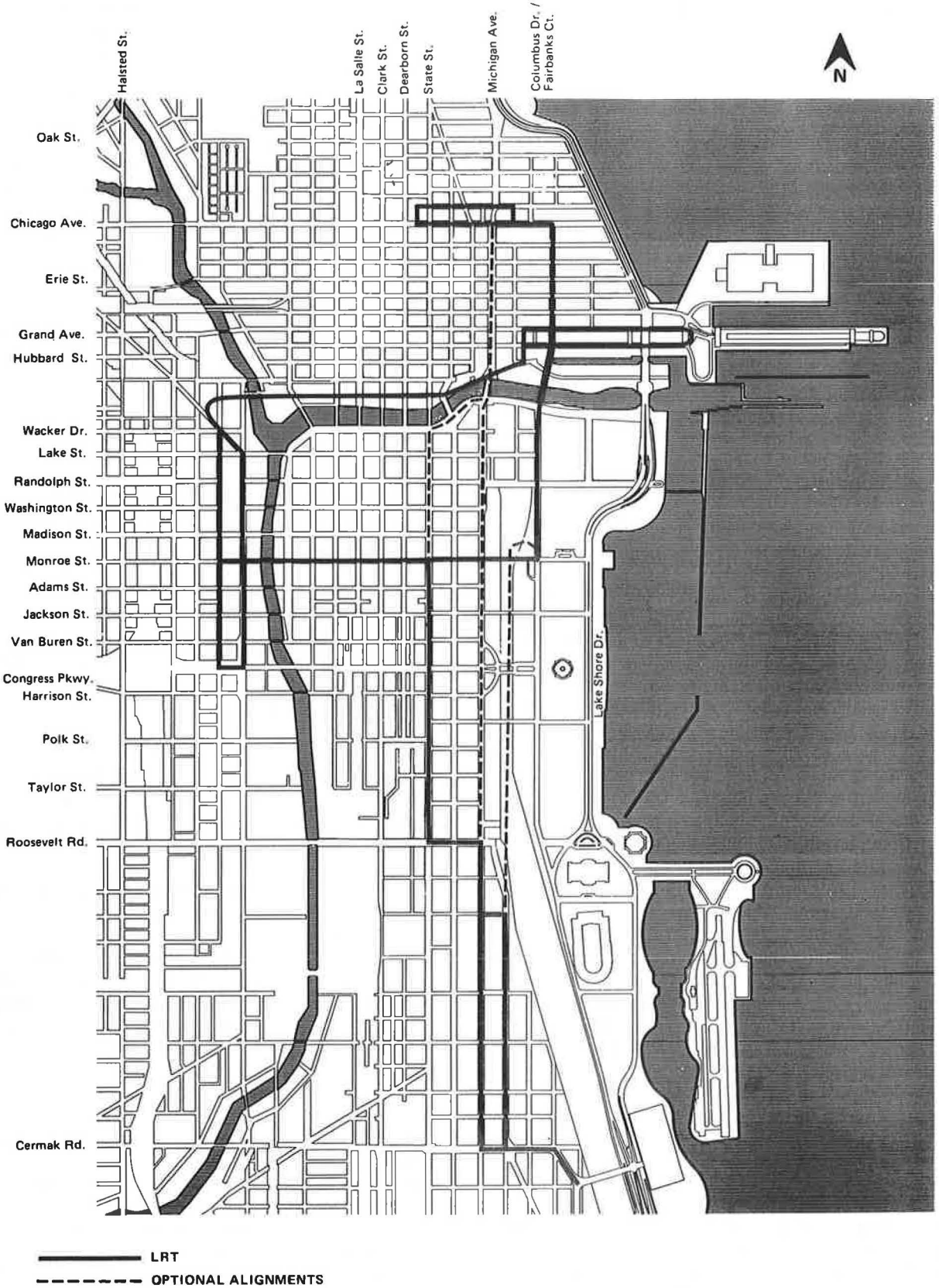

FIGURE 2 Full-build LRT system.

Two types of adjustments were made to system coefficients. The first adjusted the coefficients on travel cost to account for a different base year for measuring travel costs and for the local value-of-time. The value-of-time for central area commuters was assumed to be one-third of the average wage rate of the Metra commuters (as obtained from 1980 census journey-to-work data). The value-of-time for the circulator models for central area workers was assumed to be one-third of the wage rate for those workers. The average wage rate for central area workers is slightly lower than the average wage rate for central area Metra commuters. Finally, the value-of-time for central area nonworkers and midday resident trips was assumed to be one-sixth of the central area wage rate. The use of one-third and one-sixth of the average wage rate was based on guidance from UMTA.

The second type of coefficient adjustment was the modification or addition of a walk distance coefficient. This coefficient was adjusted so that modeled walk mode shares by distance range and average walk distance matched aggregate observed shares. An iterative approach was used to adjust the coefficient.

Alternative-specific constants were iteratively adjusted so that modeled aggregate mode shares matched observed aggregate mode shares. This process used the following formula to guide the adjustment of the constants:

$$C_1 = C_o + \ln \frac{P_1 \times (1 - P_o)}{P_o \times (1 - P_1)} \quad (1)$$

where

C_1 = revised constant,
C_o = original constant,
P_1 = desired share, and
P_o = share obtained using C_o.

Morning Peak-Period Distributor Mode Choice Model

The morning peak-period distributor mode choice model is a simple logit-based model of the form

$$P_m = \frac{\exp(U_m)}{\sum_j [\exp(U_j)]} \quad (2)$$

where

P_m = proportion of trips (for an interchange) using Mode m;
exp = the exponential function with the base of natural logarithms, e, as base; and
U_m = utility of Mode m.

Table 1 compares the calibrated coefficients for the Chicago morning peak-period distributor mode choice model with the models for Los Angeles, Detroit, and Miami. Some differences between the Chicago model and the other models warrant discussion. First, the coefficients for travel time and travel cost were set so that the implied value-of-time was equal to one-third of the average wage rate of central area workers who rode the Metra commuter rail service. This method for determining the coefficients for time and cost was consistent with the process used for Detroit. This method results in an implied value-of-time for central business district (CBD) workers that is substantially higher than the value-of-time for workers in Miami. It is slightly higher than the value-of-time for Detroit CBD workers (if the same year dollars were used) and substantially lower than the value-of-time used for the Los Angeles model.

TABLE 1 COMPARISON OF CENTRAL AREA DISTRIBUTOR MODEL COEFFICIENTS

Coefficient/Constant	Los Angeles	Detroit	Miami	Chicago
Walk Constant	2.29000	1.99000	-1.29000	2.74164
Transit Constant	0.20500	0.19860	-3.06200	-0.27072
Circulator Constant	0.00000	0.72500	0.00000	NA
Taxi Constant	NA	NA	NA	-3.13828
Travel Time (minutes)	-0.09790	-0.07419	-0.06370	-0.09000
Travel Cost (cents)	-0.00954	-0.02130	-0.02870	-0.01065
Walk Distance (miles)	NA	NA	NA	-3.00000
Walk Distance (transit paths)[a]	NA	NA	NA	-3.00000
Implied Value of Time	$6.16	$2.09	$1.33	$5.07
Year for Dollars	1975	1975	1986	1985

NA = not applicable.
[a] The walk distance coefficient for transit is applied only to the distance walked for transit access, egress, and transfer.

The second difference was the inclusion of taxi as a viable distributor mode. This was done by adding a taxi alternative-specific constant and adjusting the model constants to provide correct overall mode shares for the central area distributor portion of the model.

Perhaps the biggest difference between the Chicago model and the other models was the inclusion of walk distance as an explanatory variable. The need to add this variable resulted from an analysis of aggregate mode shares summarized by walk distance. Without the use of this explanatory variable, the model tended to greatly overpredict the walk mode share for walk travel times greater than 15 to 20 min. This overprediction of walk shares for the longer walk travel times resulted in an underprediction of transit use for the same walk travel time range. Figure 3 shows a comparison of modeled and observed mode shares by walk time for the walk and transit modes based on the calibrated model. As can be seen, walk shares still tend to be overpredicted and transit shares underpredicted. However, the relative magnitude of the over- and underpredictions was substantially reduced by the walk distance variable.

Walk distance can be directly converted to walk travel time, because a constant walk speed of 3 mph is used in the modeling process. Thus, it might be argued that the inclusion of walk distance in the utility function is equivalent to adding -0.15 to the coefficient of travel time for walk trips. This would result in raising the value-of-time for walk trips to $13.52/hr, which is similar to the value-of-time used for the Los Angeles model (if stated in the same year dollars). It is, however, more proper to consider this variable as a "fatigue factor," not as a change in the value-of-time. Alternatively, the inclusion of the walk distance factor can be viewed as an adjustment to account for the effects of out-of-vehicle travel time. Many regional mode choice models have found out-of-

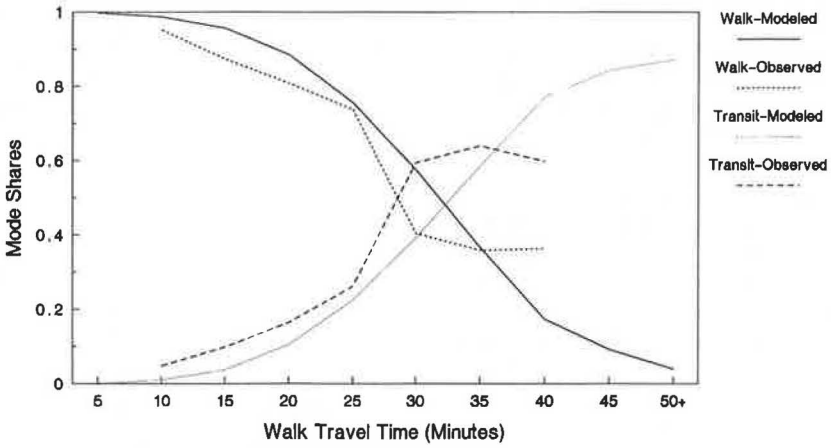

FIGURE 3 Comparison of modeled and observed mode shares for a.m. peak period.

vehicle travel time to be two to three times more onerous than in-vehicle travel time.

Table 2 gives the modeled and observed shares by distance range and the modeled and observed mode shares by commuter rail station. The observed shares are based on a 1985 Metra survey. As is obvious, the model reproduces the observed results closely for the North Western, Union, and LaSalle Street stations. Whereas the match was not as close for the Van Buren and Randolph stations, a separate 1989 survey suggested that walk mode shares for those stations were near 95 percent. This implies that the models were also working properly for those two stations.

Morning Peak-Period Circulator Trip Distribution—Mode Choice Model

The morning peak-period circulator model is a logit-based simultaneous trip distribution-mode choice model of the following form:

$$P_{j,m} = \frac{\exp(U_{j,m})}{\sum_j \sum_m [\exp(U_{j,m})]} \quad \text{for each production zone} \quad (3)$$

where $P_{j,m}$ is the proportion of trips to Destination Zone j using Mode m and $U_{j,m}$ is the utility of Destination Zone j for Mode m.

Table 3 gives the calibrated coefficients for the Chicago morning peak-period circulator trip distribution–mode choice model. The Los Angeles, Detroit, and Miami DPM models did not include such a market segment. Unlike Chicago, they apparently did not have a CBD population large enough to warrant such a model. The Chicago model used the coefficients of travel time and travel cost determined for the morning peak-period distributor model. This was based on the assumptions that the trip purpose being represented in both the peak-period distributor and circulator models was the work trip and that workers should have similar sensitivities to travel time and travel cost regardless of where they live. The destination zone trip density and logarithm of the destination zone area coefficients were taken from the midday circulator models for workers. The walk distance and walk distance to transit coefficients were adjusted to match observed average trip lengths as closely as possible. Finally, the

TABLE 2 COMPARISON OF MODELED AND OBSERVED EGRESS MODE SHARES FOR COMMUTER RAIL STATIONS

	Walk Share		Transit Share		Taxi Share	
	Modeled	Observed	Modeled	Observed	Modeled	Observed
Mode Shares by Walk Distance (time interval in minutes)						
0-5	100%	—	0%	—	0%	—
5-10	99	95%	1	5%	0	—
10-15	96	87	4	10	1	3%
15-20	87	81	11	16	1	3
20-25	76	74	23	26	2	—
25-30	58	41	39	59	3	—
30-35	37	36	59	64	5	—
35-40	17	36	77	60	5	4
40-45	9	—	84	—	7	—
45-50+	4	—	87	—	9	—
Mode Shares by Rail Station						
Roosevelt	40%	—	53%	—	7%	—
Van Buren	94	88%	5	12%	1	0%
Randolph	95	82	4	14	1	4
North Western	83	84	16	16	1	1
Union	82	81	17	18	1	1
LaSalle	92	91	7	6	1	2
Average	84%	84%	15%	15%	1%	1%

— Indicates that data were unavailable.

TABLE 3 CENTRAL AREA PEAK-PERIOD CIRCULATOR MODEL COEFFICIENTS

Coefficient/Constant	Chicago
Walk Constant	2.69269
Transit Constant	0.54637
Taxi Constant	-1.06622
Auto Constant	0.00000
Travel Time (minutes)	-0.09000
Travel Cost (cents)	-0.01065
Walk Distance (miles)	-4.70000
Walk Distance (transit paths)[a]	-4.70000
Destination Zone Trip Density (trip attr/acre)	0.00767
Ln of Destination Zone Area (in acres)	1.00000
Implied Values of Time	$5.07
Year for Dollars	1985

[a] The walk distance coefficient for transit is applied only to the distance walked for transit access, egress, and transfer.

alternative-specific constants were adjusted to match observed mode shares. Table 4 gives observed and modeled mode shares and average trip lengths by mode. The observed data were obtained from unpublished results of a CATS survey of central area residents.

Midday Distributor Mode Choice Model

The midday distributor mode choice model for Chicago has the same form as the morning peak-period distributor model. Like the peak-period circulator model, the midday distributor mode choice model for Chicago does not have a counterpart in the Los Angeles, Detroit, or Miami models. In 1985, approximately 16,000 trips were made to the central area on Metra outside the morning peak period. This was only 15 percent of the total daily trips made to the central area on Metra during the day, but it was not a trivial number of trips. To calibrate the model, the morning peak distributor model was chosen as a base. The coefficient of travel cost was doubled to cut the implied value-of-time in half for two reasons: (*a*) the midday trips were assumed to be nonwork trips and (*b*) travelers making nonwork trips value their time at approximately one-half the level of work-trip travelers. The alternative-specific constants were not modified from the values determined for the morning peak-period distributor model. Table 5 gives the coefficients and constants used for the midday distributor model.

No data were available to guide the adjustment of the midday distributor model constants and coefficients. However, the model results were compared with the peak-period model results for reasonableness. Table 6 gives the comparison. Note that the midday walk shares were consistently higher by distance range interval than the peak shares. This was expected because of the lower value-of-time used in the model. On the other hand, results by commuter rail station were mixed because of differences between the distributions of trips by commuter rail station for the peak and midday periods. The North Western, Union, and LaSalle Street stations were relatively farther from the midday attractors of trips than they were from the peak-period attractors. Thus, the midday walk shares from those stations were lower than the peak-period walk shares. The opposite was true for the Roosevelt, Van Buren, and Randolph Street stations.

TABLE 4 OBSERVED AND ESTIMATED MODE SHARES AND AVERAGE TRIP LENGTHS FOR PEAK-PERIOD CIRCULATOR MODEL

	Mode Share		Average Trip Length (minutes)[a]	
Mode	Observed	Modeled	Observed	Modeled
Walk	51.4%	51.4%	8.5	8.3
Transit	23.2	23.1	21.4	28.0
Taxi	12.6	12.7	20.1	24.9
Auto	12.9	12.8	19.0	37.4

[a] Average trip lengths based on walk travel times between zones for all modes.

TABLE 5 CENTRAL AREA MIDDAY DISTRIBUTOR MODEL COEFFICIENTS

Coefficient/Constant	Chicago
Walk Constant	2.74164
Transit Constant	-0.27072
Taxi Constant	-3.13828
Travel Time (minutes)	-0.09000
Travel Cost (cents)	-0.02130
Walk Distance (miles)	-3.00000
Walk Distance (transit paths)[a]	-3.00000
Implied Values of Time	$2.54
Year for Dollars	1985

[a] The walk distance coefficient for transit is applied only to the distance walked for transit access, egress, and transfer.

TABLE 6 COMPARISON OF MORNING PEAK-PERIOD AND MIDDAY DISTRIBUTOR MODE SHARES

	Walk Share		Transit Share		Taxi Share	
	Modeled	Observed	Modeled	Observed	Modeled	Observed
Mode Shares by Walk Distance (time interval in minutes)						
0-5	100%	100%	0%	1%	0%	0%
5-10	99	99	1	1	0	0
10-15	96	99	4	1	1	0
15-20	87	96	11	4	1	0
20-25	76	93	23	7	2	0
25-30	58	88	39	12	3	0
30-35	37	66	59	33	5	1
35-40	17	38	77	62	5	1
40-45	9	26	84	73	7	1
45-50+	4	13	87	87	9	1
Mode Shares by Rail Station						
Roosevelt	40%	66%	53%	34%	7%	1%
Van Buren	94	97	5	3	1	0
Randolph	95	98	4	2	1	0
North Western	83	82	16	17	1	0
Union	82	76	17	22	1	0
LaSalle	92	91	7	9	1	0
Average	84%	83%	15%	16%	1%	0%

Midday Circulator Trips—Central Area Workers

The midday circulator model for central area workers simultaneously determines trip generation, trip distribution, and mode choice. The model has the following form:

$$P_{f,j,m} = \frac{\exp(U_{f,j,m})}{\sum_f \sum_j \sum_m [\exp(U_{f,j,m})]} \quad \text{for each production zone} \tag{4}$$

where $P_{f,j,m}$ is the probability of making zero trips or making one trip to Zone j on Mode m and $U_{f,j,m}$ is the utility of making zero trips or making one trip to Zone j on Mode m.

Each of the combinations of alternatives has its own utility function that takes on unique values for each origin zone and in all cases, except for making zero trips, for each destination zone. The utility functions are

$$U(f = 0) = C_0 + A_1 * \text{origin employment density} \tag{5}$$

$$U(f = 1, j, \text{walk}) = C_w + A_2 * \text{walk time} + A_3 * \text{walk distance} + A_4 * \text{worker trip attraction density of Zone } j + A_5 * \ln(\text{zonal area in acres}) \quad (6)$$

$$U(f = 1, j, \text{transit}) = C_t + A_2 * \text{transit time} + A_6 * \text{transit walk distance} + A_7 * \text{transit fare} + A_4 * \text{worker trip attraction density of Zone } j + A_5 * \ln(\text{zonal area in acres}) \quad (7)$$

$$U(f = 1, j, \text{taxi}) = C_c + A_2 * \text{taxi time} + A_7 * \text{taxi fare} + A_4 * \text{worker trip attraction density of Zone } j + A_5 * \ln(\text{zonal area in acres}) \quad (8)$$

$$U(f = 1, j, \text{automobile}) = C_a + A_2 * \text{automobile time} + A_7 * \text{automobile cost} + A_4 * \text{worker trip attraction density of Zone } j + A_5 * \ln(\text{zonal area in acres}) \quad (9)$$

In Equations 5 through 9, C_0, C_w, C_t, C_c, and C_a are the alternative-specific constants for zero trips, walk, transit, taxi, and automobile, respectively, and $A_1, A_2, \ldots, A_7$ are calibrated model coefficients.

Table 7 compares the calibrated coefficients for the midday circulator model for central area workers with the models used in Los Angeles, Detroit, and Miami. There are several differences between the Chicago model and the others. First, the constant for zero trip making for Chicago is substantially higher than for Detroit and Miami but similar to the constant for Los Angeles. The constant was set to cause the model to match the observed percentage of zero trip makers for the central area in 1985: 39.7 percent. Table 8 gives the modeled-to-observed match for the other calibration measures—the mode shares and average trip lengths. Note that the observed mode shares and average trip lengths were obtained from a central area building survey performed by CATS in 1985.

The coefficient for walk distance is also substantially higher for the Chicago model than for the other cities. The coefficient was set to match an observed average walk trip length of 4.4 min for central area workers' midday trips.

The implied value-of-time for Chicago is substantially lower than for Los Angeles, similar to that for Detroit, and substantially higher than that for Miami. The variation in the values-of-time suggests that the worker model is not stable across urban areas. This is likely, because the model simultaneously projects trip frequency, trip distribution, and mode use. Indeed, one of the problems in calibrating the model for Chicago was the effect of the interaction of the variables—it was difficult to get the model to "settle down."

TABLE 7 COMPARISON OF MIDDAY CIRCULATOR MODEL COEFFICIENTS FOR CENTRAL AREA WORKERS

Coefficient/Constant	Los Angeles	Detroit	Miami	Chicago
Constant—No Trips	9.29400	4.80000	4.98160	11.50000
Walk Constant	3.03400	4.34000	5.03600	6.12823
Transit Constant	2.90000	2.43540	2.80200	0.39150
Circulator Constant	-0.81000	1.08500	-0.05400	0.79697
Taxi Constant	NA	NA	NA	-1.27064
Auto Constant	0.00000	0.00000	0.00000	0.00000
Travel Time (minutes)	-0.09190	-0.05226	-0.05980	-0.05226
Travel Cost (cents)	-0.00896	-0.01500	-0.04120	-0.00750
Walk Distance (miles)	-3.00000	-3.00000	-3.00000	-9.50000
Walk Distance (transit paths)[a]	-4.20000	-4.20000	-4.20000	-1.00000
Destination Zone Trip Density (trip attractions/acre)	0.00767	0.00767	0.00767	0.00767
Ln of Dest. Zone Area (in acres)	1.00000	1.00000	NA	1.00000
Employment Density (employ./acre) (for 0 trip makers)	0.0008552	0.0008552	0.0008552	0.0008552
Implied Value of Time	$6.15	$2.09	$0.87	$4.18
Year for Dollars	1975	1975	1986	1985

NA = not applicable.
[a] The walk distance coefficient for transit is applied only to the distance walked for transit access, egress, and transfer.

TABLE 8 MIDDAY CIRCULATOR MODE SHARES AND AVERAGE TRIP LENGTHS (CENTRAL AREA WORKERS)

	Mode Share		Average Trip Length (minutes)	
Mode	Observed	Modeled	Observed	Modeled
Walk	90.1%	89.5%	4.4	4.9
Transit	6.5	7.0	24.7	24.6
Taxi	1.6	1.7	—	19.0
Auto	1.7	1.8	—	25.5

— Indicates that data were unavailable.

Contrary to the other models used for Chicago, the midday central area worker model included a constant for the circulator mode that was different from the normal transit mode. Unlike the models for the other cities, the constant for the circulator makes it more attractive than transit when all travel impedances are equal. This was done in the Chicago model because the circulator (LRT) was not modeled as an explicit mode separate from transit, but rather as a transit submode. This procedure avoided the independence of irrelevant alternatives problem that would have been obvious if LRT had been considered a new mode. LRT could not be considered substantially different from the bus, subway, and elevated systems already in place in the central area.

Nevertheless, because of its visibility, accessibility, fare collection system, and other unique characteristics, it was felt that LRT would be more attractive than the existing transit system. This is especially true for intra-central-area trips. Specifically, it was assumed that at a point of indifference on an interchange, travelers would choose LRT 60 percent of the time and regular transit 40 percent of the time. This meant that the constant for LRT should be more positive than the constant for transit by the following amount: Delta = ln (0.6/0.4) = 0.405. Note that at the point of indifference (i.e., 50 percent choose bus and 50 percent choose transit), the delta value would be zero.

In the actual application of the model, the added attractiveness of LRT was added to the utility of transit in proportion to the amount of in-vehicle travel time spent on LRT. In other words, if LRT was used for 50 percent of the in-vehicle travel time on an interchange, the utility of transit was only 0.2025 more than what the transit utility would have been if only bus had been used for the entire trip. The full difference (0.405) was added only if LRT was the only transit mode used for the entire interchange.

The LRT attractiveness difference can be equated with travel time or travel cost by dividing by the coefficients of travel time or travel cost, as appropriate. The maximum LRT attractiveness difference is equivalent to 7.8 min of travel time or 54 cents of travel cost.

During the calibration and validation of the model, it was discovered that the trip distribution portion of the model was sensitive to the destination zone trip density. There is a substantial variation in trip density in the central area. It was discovered that several zones were attracting a major portion of the trips from all other zones in the central area. To solve this problem, maximum trip densities were set at two standard deviations above the mean trip attraction density for the central area.

Midday Circulator Trips—Central Area Nonworkers

The midday circulator model for central area nonworkers is a simultaneous trip distribution–mode choice model. It is of the same form as the model used for peak-period circulator trips. However, in the midday circulator model for nonworkers, the destination zone trip density is based on non-home-based trips, not home-based work trips as in the morning peak-period circulator model. Table 9 compares the calibrated coefficients for the Chicago model with the models used for Los Angeles, Detroit, and Miami.

The coefficient for walk distance is substantially higher for the Chicago model than for the other cities. This is the same situation that occurred for the midday circulator model for central area workers. The coefficient was set to match an observed average trip length of 4.4 min. The average walk trip length for nonworkers was assumed to be identical to the average trip walk length for central area workers. Table 10 compares modeled and observed mode shares and average trip lengths for central area nonworkers. As with the midday circulator model for workers, the observed nonworker travel characteristics were obtained from the 1985 central area building survey performed by CATS.

When the implied values-of-time are compared across the three cities for the circulator model for central area nonworkers, the same patterns emerge as in the circulator model for central area workers. Specifically, the Chicago value-of-time is substantially lower than the value-of-time used for Los Angeles, similar to the value-of-time used for Detroit, and substantially higher than the value-of-time used for Miami.

TABLE 9 COMPARISON OF MIDDAY CIRCULATOR MODEL COEFFICIENTS FOR CENTRAL AREA NONWORKERS

Coefficient/Constant	Los Angeles	Detroit	Miami	Chicago
Walk Constant	2.92200	2.87680	3.82400	5.58921
Transit Constant	1.31800	4.35300	1.01100	2.78463
Circulator Constant	-3.15500	1.93800	-1.03800	3.19010
Taxi Constant	NA	NA	NA	-1.00428
Auto Constant	0.00000	0.00000	0.00000	0.00000
Travel Time (minutes)	-0.08780	-0.16900	-0.05810	-0.05226
Travel Cost (cents)	-0.01096	-0.09657	-0.04280	-0.01500
Walk Distance (miles)	-3.00000	-3.00000	-3.00000	-9.00000
Walk Distance (transit paths)[a]	-4.20000	-4.20000	-4.20000	-9.00000
Destination Zone Trip Density (trip attractions/acre)	0.00378	0.00378	0.00378	0.00378
Ln of Dest. Zone Area (in acres)	1.00000	1.00000	NA	1.00000
Implied Value of Time	$4.83	$1.05	$0.81	$2.09
Year for Dollars	1975	1975	1986	1985

NA = not applicable.
[a] The walk distance coefficient for transit is applied only to the distance walked for transit access, egress, and transfer.

Midday Circulator Trips—Central Area Residents

The midday circulator model for central area residents is a simultaneous trip distribution–mode choice model. It is of the same form as the model used for peak-period circulator trips and the midday circulator model for nonworkers. For the midday circulator model for central area residents, the destination zone trip density is based on home-based nonwork trips. Table 11 gives the calibrated coefficients for the Chicago

TABLE 10 MIDDAY CIRCULATOR MODE SHARES AND AVERAGE TRIP LENGTHS (CENTRAL AREA NONWORKERS)

	Mode Share		Average Trip Length (minutes)	
Mode	Observed	Modeled	Observed	Modeled
Walk	92.7%	92.2%	4.4	4.9
Transit	3.5	3.7	24.7	24.7
Taxi	0.9	1.0	—	17.9
Auto	3.0	3.2	—	38.7

— Indicates that data were unavailable.

TABLE 11 MODEL COEFFICIENTS FOR MIDDAY CIRCULATOR TRIPS (CENTRAL AREA RESIDENTS)

Coefficient/Constant	Chicago
Walk Constant	6.35276
Transit Constant	3.51171
Taxi Constant	-0.69088
Auto Constant	0.00000
Travel Time (minutes)	-0.05226
Travel Cost (cents)	-0.01500
Walk Distance (miles)	-10.00000
Walk Distance (transit paths)[a]	-10.00000
Destination Zone Trip Density (trip attr/acre)	0.00378
Ln of Destination Zone Area (in acres)	1.00000
Implied Values of Time	$2.09
Year for Dollars	1985

[a] The walk distance coefficient for transit is applied only to the distance walked for transit access, egress, and transfer.

TABLE 12 MIDDAY CIRCULATOR MODE SHARES AND AVERAGE TRIP LENGTHS (CENTRAL AREA RESIDENTS)

	Mode Share		Average Trip Length (minutes)	
Mode	Observed	Modeled	Observed	Modeled
Walk	92.0%	92.1%	4.9	5.3
Transit	4.0	3.8	24.7	28.0
Taxi	1.0	0.9	—	24.9
Auto	3.0	3.2	—	37.4

— Indicates that data were unavailabie.

model. Table 12 gives the observed and modeled mode shares and average trip lengths by mode for the midday resident trips. Los Angeles, Detroit, and Miami did not use a comparable model for their DPM models.

SUMMARY

A useful central area circulator/distributor model has been calibrated for Chicago based on the DPM modeling methodology originally performed for Los Angeles. A number of interesting lessons were learned during the calibration of the models. This led to the following conclusions and recommendations regarding DPM models.

First, detail is critical. A great amount of detail was used in defining the Chicago central area zone structure and transportation network. If the model calibration were performed again for Chicago, additional detail would probably be used in describing the transit system. There are difficulties in acquiring the data necessary to develop detailed networks and the socioeconomic data for detailed zones. However, the detail is crucial to properly model the utilities of the different choices available in circulator/distributor models.

In future model calibration efforts, attempts should be made to move away from simultaneous model forms. Whereas simultaneous model forms might be theoretically satisfying, they are very difficult to control in practice. In addition, when the circulator/distributor models are transferred from one urban area to another, it can be easy to "forget" the distribution parts of the models in attempts to match mode shares. Although they are not the subject of this paper, simultaneous trip distribution–mode choice models make it difficult or impossible to isolate the effects of system changes in alternatives analyses.

Attention should be paid to average trip lengths and mode shares by average trip length in the calibration of circulator/distributor models. It is interesting to compare the various models from Los Angeles, Detroit, Miami, and Chicago and speculate on the effects of the different coefficients on the model results. In many ways, the Chicago models are similar to the Los Angeles models. The Los Angeles models had a high value of travel time. On the basis of the need to add the coefficient of walk distance to the Chicago model, it is likely that the high value-of-time in the Los Angeles model was compensating for the disutility of walk distance. In contrast, if a model with low values-of-time had been used in Chicago (similar to the Miami models), the coefficient of walk distance would have been even more important to keep the modeled travelers from walking "forever."

Another change that might be appropriate for circular/distributor models would be to disaggregate travel-time into out-of-vehicle and in-vehicle travel time as is done in many regional mode choice models. This change would make the circulator/distributor models more consistent with the regional model and decrease the importance of the walk distance coefficient.

This model development effort highlighted the need for additional research into the modeling of circulator/distributor trips. The circulator/distributor models for Detroit, Miami, and Chicago are all derivatives of the Los Angeles model developed in the late 1970s. The Detroit and Chicago models adjusted the Los Angeles model constants and coefficients to match observed aggregate travel characteristics. The Miami model was apparently recalibrated in a more rigorous manner; however, the basic structure was not changed from the structure originally developed for Los Angeles. Experience in the development of the Chicago model suggests that changes in the basic model structure used for circulator/distributor models could improve both the understandability and the reasonableness of the models. However, such changes will require a concerted effort to collect and analyze circulator/distributor mode choice data from cities with circulator/distributor systems.

ACKNOWLEDGMENTS

The authors would like to thank Robert Kunze and Richard Hazlett of the City of Chicago Department of Planning for their invaluable assistance throughout the duration of this project. They were especially helpful in obtaining base data and providing timely suggestions. We would also like to thank Dean Englund and Joyce Stenzel of the Chicago Area Transportation Study for their help in obtaining the regional network and their continuing assistance.

The preparation of this report was financed in part by UMTA, U.S. Department of Transportation, under the Urban Mass Transportation Act of 1964, as amended.

REFERENCES

1. *Planning for Downtown People Movers*. Office of Planning Methods and Support, UMTA, U.S. Department of Transportation, 1979.
2. Charles River Associates, Inc. *An Independent Forecast of Detroit DPM Ridership*. Joint Center for Urban Mobility Research, Rice Center, Houston, Tex., 1986.
3. K. G. Brooks and M.-H. Sung. Miami Downtown People Mover Demand Analysis Model. In *Transportation Research Record 1167*, TRB, National Research Council, Washington, D.C., 1988.
4. *EMME/2 User's Manual—Software Version: 4.0*. INRO Consultants, Inc. Montreal, Canada, 1989.

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

Evaluating the Sensitivity of Travel Demand Forecasts to Land Use Input Errors

JIT N. BAJPAI

The sensitivity of the urban transportation planning process (UTPP) to differences (or errors) in socioeconomic input is examined using the Dallas–Fort Worth area as a case study. The sensitivity analysis indicated that the final output of the UTPP, link volumes, is sensitive to errors in the district-level forecasting of population and employment. Planners undertaking corridor-level studies should be most concerned about the reliability and accuracy of district-level socioeconomic forecasts, particularly for districts directly served by the corridor. Because local transportation facilities are most sensitive to errors in zone-level inputs, site-specific studies may be severely affected by data errors at the traffic zone level. Greater attention should be paid to suburban areas where the potential for introducing large input errors is high. Because of the potential for large assignment errors, expansion of a district-to-district trip table to a zone-to-zone trip table should be avoided. Travel demand models must be applied using traffic zone–level data, not district-level data.

The Urban Transportation Planning Process (UTPP), comprising a sequence of models, is commonly used to analyze and predict traffic volumes on a transportation network on the basis of anticipated changes in socioeconomic variables such as population, employment, vehicle availability, income, and household size. To produce an accurate geographic distribution of trip making, these models are usually applied at the traffic zone level using the regional transportation system network- and zone-level socioeconomic inputs. Agencies responsible for the preparation of zone-level data, in general, develop inputs in two steps: first, from region to district and, second, from district to zone (*1*). Because events that influence the location of economic activities and population are difficult to predict, allocation errors are inevitable though of different magnitude in the two steps of allocation.

Large errors in the forecast of socioeconomic variables are likely to produce highly inaccurate traffic forecasts for decision makers responsible for investment in transportation infrastructure. For instance, a large disparity between the forecast and actual traffic can cause substantial misallocation of public resources due to under- or overdesign of a facility. A recent comparative analysis of 10 urban rail transit projects (*2*) indicated that overestimates of future population and employment in downtown areas were sufficiently large to contribute significantly to overestimation of future ridership on some rail projects built with federal funds. Acknowledging this, it becomes necessary for a transportation planner to understand the implications of various types and magnitudes of socioeconomic input errors on the prediction of travel demand. This issue is examined by analyzing the sensitivity of travel demand forecasts to land use input errors.

Prior works on the propagation of errors in the travel prediction process mostly considered errors caused by the structure of models, input data size, and aggregation procedures. For example, one of the early works on the sensitivity of the four-step travel prediction process to model specification errors and sampling variation in data used was conducted by CONSAD in 1968 (*3*). Later, efforts were made to analyze errors in prediction with disaggregate choice models. Koppelman (*4*) identified major sources of errors in prediction, including model error and aggregation error, and suggested ways to improve prediction models.

The sensitivity analysis reported in this paper differs from earlier efforts. The focus of this analysis was on evaluating only the impact of changes in the input of socioeconomic data on the UTPP outputs; other inputs of the process (network data, travel data, and zone structure and the models) were kept the same. The same travel demand model was repeatedly applied to avoid the effect of changes in the structure of the model on traffic forecasts. The Dallas–Fort Worth area travel demand model was selected for the sensitivity tests. Five scenarios representing various types and magnitudes of errors in district and subdistrict (zone) allocations were tested.

APPROACH

Selected Travel Demand Model

The Dallas–Fort Worth area travel demand model was selected for the sensitivity tests. The selection considered the sophistication of and familiarity with travel demand models as well as staff skill levels in their maintenance and operation. However, the foremost factor was the willingness of the North Central Texas Council of Governments (NCTCOG) to assist in the analysis. NCTCOG currently uses the DRAM/EMPAL (*5*) models to allocate regional land use forecasts for the nine-county Dallas–Fort Worth area to 170 forecast districts. The district forecasts are subsequently allocated to almost 5,691 traffic survey zones (TSZs). Travel demand simulation models produce assignments for transit and highway networks using the TSZ-level socioeconomic variable inputs (households; median income; and employment by basic, retail, and service categories) and mode-specific network attributes (*6*). The model

COMSIS Corporation, Suite 1100, 8737 Colesville Road, Silver Spring, Md. 20910. Current affiliation: World Bank, 1818 H Street, N.W., Washington, D.C. 20433.

generates interzonal trip tables for four income categories (low, low-middle, high-middle, and high) and four trip purposes (home-based work, home-based nonwork, non–home based, and other). The four-step modeling process (trip generation, trip distribution, mode choice, and network assignment) includes three sophisticated multinomial logit type mode choice models for each of the three trip purposes (excluding "other" purpose).

The trip distribution model uses a standard gravity formulation technique and a second-order Bessel curve as the travel decay function for each of the seven trip purposes (home-based work for three income groups, home-based nonwork, non–home based, and other). Minimum time of travel between zones is used as the impedance measure in the distribution models.

For roadway traffic assignment a capacity-restrained assignment model with an incremental loading procedure is applied. A generalized cost function of time, distance, and toll is used in building the minimum paths between zones. The model uses the upper and lower bounds of the total number of trips (defined as 20,000 and 100,000 trips) and the critical volume-to-capacity ratio (set at 0.8) as three parameters for controlling link updating. As the network becomes congested, the trips assigned between each successive link updating decrease from the upper bound toward the lower bound. Two separate volume-delay equations are used for high- and low-capacity facilities. Further distinction is made between the daily and peak-hour assignment models. As the volume-to-capacity ratio exceeds 0.7, the delay rises exponentially to the maximum allowable delay.

Because all sensitivity tests were performed for the base year, the mode shares (transit and highway) represented the observed shares in the base year. In other words, mode-split models were not used while running the entire model chain. The primary intent was to examine the effect of allocation errors on highway trip assignments only. The daily assignment procedure was applied, and estimates of daily link volumes were converted to hourly units using factors of 0.10 and 0.12 for high- and low-capacity facilities, respectively.

Scenarios of Land Use Input Errors

Traffic zone–level inputs are usually prepared in two steps: first, from region to district and second, from district to zone. Most agencies use two separate methods for each of the two steps of forecasting. The difference between the input forecasts and reality, referred to here as errors in allocation/forecasting, can occur at either step, though the errors are of different magnitude and nature. "Nature of error" means the pattern of error distribution in the geographical space. Errors can be concentrated in a few locations, reflecting a geographical bias, or distributed in a random fashion. The contemporary phenomenon of rapid suburban growth, for example, can easily produce underprediction of employment in the suburbs and overprediction in the central city. This can occur if the forecasting/allocation method used for district forecasts fails to anticipate, for example, the magnitude and trend of suburbanization in high-technology service jobs. Similarly, one or more biased parameters in an analytical method can produce allocation errors of almost random nature.

To illustrate the effect of errors (or changes) in socioeconomic input data on the UTPP outputs, five scenarios, representing various types and magnitudes of errors in district and subdistrict allocations, were formulated. In addition, a separate test scenario was developed to examine the impact of assigning a zone-to-zone trip interchange table that was produced by disaggregating a district-to-district trip table instead of using a trip table generated directly by the application of travel demand models at the zone level. The intent behind this scenario was to examine whether the practice among certain agencies of expanding a trip table to develop a smaller spatial unit level trip interchange table is accurate. In practice, a district-level trip table, an output of either a trip distribution model or a land use simulation model such as DRAM/EMPAL (*5*), POLIS (*7*), or PLUM (*8*), is often disaggregated into a zonal trip using a proportioning method.

Although many test scenarios could be evaluated, the following six tests addressed the issues of land use allocation mentioned previously within the available resources.

- Errors in district level allocation were evaluated by Test A (introduce ± 20 percent random errors into land use forecasts at the district level) and Test B (introduce ± 40 percent random errors into land use forecasts at the district level).
- Errors in subdistrict level allocation were evaluated by Test C (introduce geographical bias into land use allocation at the district level), Test D (introduce ± 40 percent random errors into land use forecasts at the zone level), and Test E (distribute district forecasts uniformly among zones composing a district).
- Trip table disaggregation was evaluated by Test F (disaggregate district-level trip table to zone level).

For Tests A, B, and D, the distribution of error was created by randomly selecting one-third of districts or zones for positive, one-third for negative, and one-third for no errors. The overall magnitude of positive error was kept equal to negative error and distributed in proportion to population and employment. For Test C, employment for the downtown district and the districts in the southwestern sector of Dallas was reduced by 20 percent, and the same magnitude of employment increase was allocated among districts located in the northwestern sector (Figure 1) in the same proportion as existing employment. For developing uniform distribution under Test E, traffic zone–level data were prepared by dividing the district forecasts by the number of zones (i.e., P/N, where P and N represent the population and the number of zones in a district, respectively). For Test F, an 800 × 800 trip interchange table was first collapsed into a 147 × 147 table and then expanded back to an 800 × 800 table using zonal shares of households and employment in each district. The number of households and total employment were used as the proportioning factors for trip productions and attractions, respectively.

Sensitivity of Travel Models to Zone-Level Inputs

Changes introduced in zonal socioeconomic inputs affect all four steps of travel forecasting: trip generation, distribution, mode choice, and assignment. Even though the population

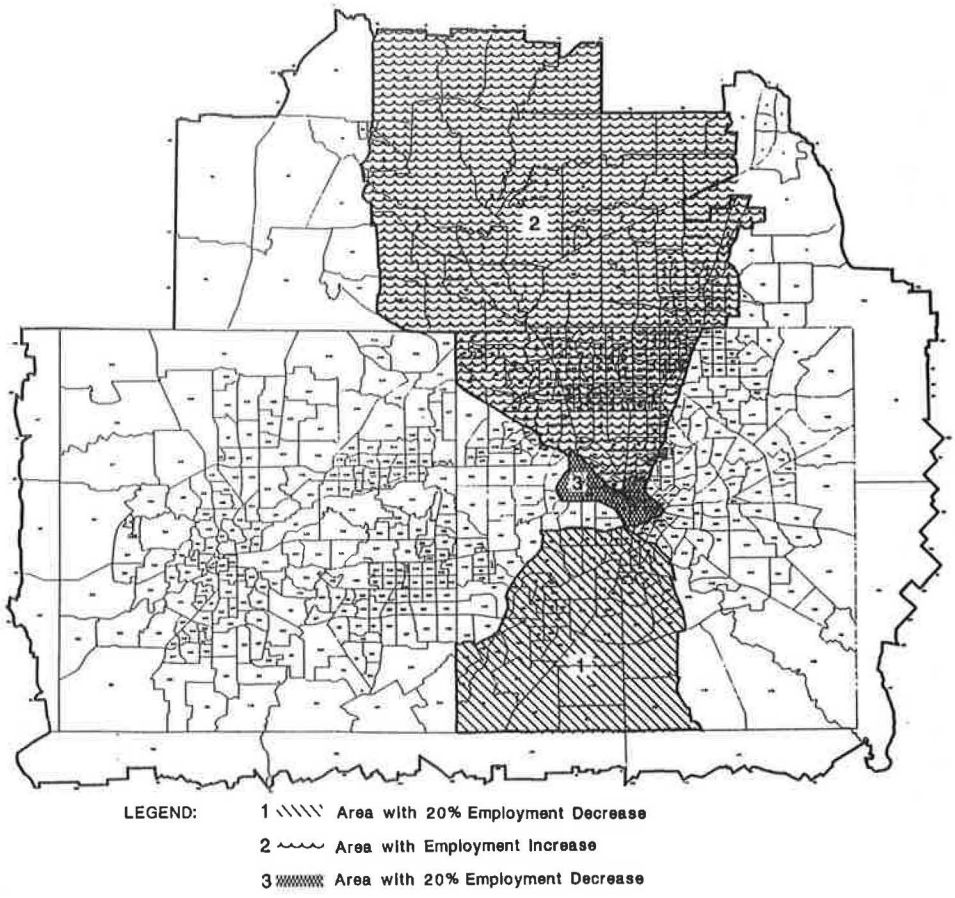

FIGURE 1 Introduction of geographical bias into land use allocation (Test C).

and employment control totals remain fixed, changes in their allocation among zones cause variation in estimates of zone-level trip productions and attractions. The level of variation depends on the magnitude of change introduced in the zone-specific variables and their sensitivity to trip generation rates. Because the trip production model of the Dallas–Fort Worth area is a function of four income and six household-size categories, any change in zonal population would affect the estimate of trip productions for a given percentage distribution of households by income and household size. For example, the net effect of a population decline in low-income zones and a rise in high-income zones would be an increase in systemwide travel due to the positive relationship between household income and trip productions.

The Dallas–Fort Worth area trip attractions are defined as the number of person trips per employee and are stratified by five area types, four employment types, and, in the case of home-based work trips, purpose by income quartile. For each zone the estimate of trip attractions by purpose varies with the magnitude of change in zonal employment, assuming that the existing shares of employment categories and household income quartiles remain the same. Though the trip-balancing procedure guarantees that regional production and attraction totals by purpose remain equal, their geographical distributions are disturbed under each land use scenario.

The trip distribution model is sensitive to person trips estimated by the trip generation model to and from each zone, plus the zone-to-zone travel time. Because interzonal travel times remain fixed across scenarios, the trip distribution pattern is principally influenced by the change in the estimates of zone-level productions and attractions. In the case of Dallas–Fort Worth, the distribution model for work trips is further stratified by household income to capture the income effect on commuting trip length. Low-income households are more sensitive to travel impedance compared with higher-income households in the Dallas–Forth Worth area. Therefore, trip lengths and the resulting trip distribution patterns are also affected by the spatial distribution of various income-offering jobs and wage-earning households.

Because mode shares (proportion of transit and highway) between zones represent the observed shares in the base year, under each test scenario the estimate of highway trips between zones principally depends on total trip estimates between zones (output of the trip distribution stage). The effect of fixed mode shares on total highway trips can be pronounced for zone pairs that experience a large change in the estimate of total trips compared with the base and exhibit high transit share in the base year [e.g., radial travel to central business district (CBD) from low-income areas].

The effect of changes in zonal inputs thus propagates through each of the three steps of travel demand estimation and, finally, a highway trip table is produced. The roadway assignment model uses an incremental capacity-restrained procedure to load the vehicle trip table onto the road network.

Because the assignment process is sensitive to congestion, a significant change in the orientation of vehicle trips can trigger traffic diversion. The output of this step is road link volumes, the final outcome of the four-step process. Hence, a comparison of the base year and individual test scenario specific link volumes illustrates the magnitude of overall traffic impact caused by a particular scenario. In other words, changes in link volumes manifest the accumulated effect of the zone-level input changes on link-level travel.

Traffic Impact Measures

For each test scenario, the effect of introducing a certain type and magnitude of socioeconomic input error on the final output of UTPP is measured at both the systemwide level and the micro level. The output measures of individual test runs and the run without any error (base run) are compared to illustrate the magnitude of the effect.

At the system level, root mean square error (RMSE), average trip length, and total number of trips are considered as aggregate output measures of UTPP. The final UTPP output is traffic volumes assigned on a particular transportation system network or a group of links. To check the accuracy of UTPP, however, the assigned link volumes are generally compared with the ground counts (or ridership counts in the case of transit). RMSE is usually calculated to indicate the overall goodness of fit between traffic counts and model assigned traffic volumes. It is measured in the following manner:

$$\text{RMSE} = \sqrt{\frac{\sum (\text{count} - \text{assigned volume})^2}{(N - 1)}}$$

where N is the number of traffic count stations.

Because of the large number of trip interchanges in an area, average trip length is a commonly used summary measure of trip distribution patterns. Similarly, the total number of trips is a simple measure of trip generation model output.

To examine the micro-level effects of each sensitivity test, lane error, a measure reflecting the difference between the test case assigned link volume and the base case (without error) link volume, is calculated for individual links. Lane error for a link is defined as

$$\text{Lane error} = \frac{\text{test case link volume} - \text{base link volume}}{\text{link capacity} * \text{number of lanes}}$$

Link capacity is expressed in terms of vehicles per hour per lane. It varies with the type of facility (freeways, major arterial, minor arterial, and collector) and area type (downtown, suburb, etc.). Lane error is a measure of discrepancy in traffic volume easily understood by transportation planners and highway engineers. A road planning rule of thumb suggests that a forecast error not exceeding one half-lane (positive or negative) is tolerable without a high probability of under- or overdesigning a facility. In other words, it reflects the magnitude of public resource misallocation that may occur because of the error in traffic forecasts.

FINDINGS OF SENSITIVITY TESTS

Errors in Distict Level Land Use Forecasts

Tests A, B, and C represent three scenarios of district level forecasting/allocation errors. Table 1 gives the magnitude of population and employment moved under each scenario (i.e., the magnitude of disturbance introduced in the land use allocation). Test B (± 40 percent random error) introduces the largest allocation errors, followed by Tests A (± 20 percent random error) and C (geographical bias). Test C causes minimum misallocation compared with the other two tests; only 4.3 percent of total regional jobs are moved.

Tables 2 and 3 summarize the results of the sensitivity test runs. Regionwide output comparisons between the base run (without error) and the three district-level tests (A, B, and C) indicated no significant variation in total number of trips by purpose, average trip length by purpose, or percent of RMSE (Table 2). The explanation for this could be that, because of no change in the regional control totals and the smoothing out of the effects of positive and negative errors, the values of the aggregate sensitivity measures of each test appear similar to the base run.

At the micro level, however, the overall number of links affected by more than half-lane error varied across the three test runs depending on the magnitude of activity allocation errors (see Table 3). For example, about 13.6 percent of network links experienced more than half-lane error (positive or negative) under Test B (± 40 percent random error). Under Test A (± 20 percent random error) and Test C (geographical bias), however, 5.36 and 2.68 percent, respectively, of links were affected by the same magnitude (see Table 3). Similarly, for each facility type, the proportion of links severely affected (more than half-lane error) consistently increases with the magnitude of input error across all three test runs.

The results of individual tests indicate that the effect of allocation error is not uniform across all types of road facilities. Under each of the three tests, lane error is more pro-

TABLE 1 MAGNITUDE OF POPULATION AND EMPLOYMENT MOVED UNDER DISTRICT-LEVEL ALLOCATION TESTS

Test Scenarios of Errors in District Level	Population Moved		Employment Moved	
	Amount	% of Region	Amount	% of Region
Test A: (± 20% Random Errors)	228,392	6.92	153,987	7.43
Test B: (± 40% Random Error)	456,820	13.82	307,991	14.86
Test C: (Geographical Bias)	0	0	89,930	4.34

TABLE 2 AGGREGATE OUTPUTS OF SENSITIVITY TEST RESULTS

		Errors in District Level			Errors in Subdistrict Level	
	Base No Error	Test A ± 20% Random Error	Test B ± 40% Random Error	Test C Geo-graphi-cal Bias	Test D ± 40% Random Error	Test E Uniform Distri-bution
Output						
Total Person Trips	12,468,578	12,476,562	12,471,145	12,465,871	12,500,563	12,473,757
Total Vehicle Trips	9,638,826	9,662,462	9,662,935	9,654,295	9,674,076	9,749,958
Average Trip Length by Purpose (in miles)						
HBW-Low Income	8.83	8.8	8.79	8.86	8.82	8.66
HBW-Mid. Low Income	10.52	10.46	10.47	10.54	10.52	10.43
HBW-Mid. High Income	11.65	11.58	11.54	11.63	11.61	11.61
HBW-High Income	12.10	12.03	11.98	11.95	12.03	12.01
Home Based Non-Work	6.29	6.16	6.06	6.27	6.04	6.26
Non-Home Based	6.90	6.75	6.70	6.79	6.78	7.01
Other	10.72	10.58	10.51	10.65	10.64	10.76
% RMS	60.2	61.23	64.38	59.19	61.6	60.77

TABLE 3 DISTRIBUTION OF HIGHWAY LINKS BY FACILITY TYPE AND LANE ERROR

		Percentage of Links by Lane Error (±)									
		Errors in District Level						Errors in Subdistrict Level			
		Test A		Test B		Test C		Test D		Test E	
		±20% Random Error		±40% Random Error		Geographical Bias		±40% Random Error		Uniform Distribution	
Facility Type	Number of Links	0.5 Lane	>1 Lane	0.5 Lane	>1 Lane	0.5 Lane	>1 Lane	0.5 Lane	>1 Lane	0.5 Lane	>1 Lane
Freeway	1,966	0.00	0.00	1.02	0.00	0.00	0.00	0.00	0.00	0.05	0.00
Major Arterial	1,789	12.75	2.73	22.47	16.60	8.50	1.34	11.29	5.76	19.17	13.53
Minor Arterial	3,751	9.65	2.11	16.55	12.16	3.82	1.01	11.95	3.17	16.82	13.60
Collector	5,964	2.16	0.39	5.42	2.65	0.96	0.18	3.79	1.46	6.56	4.59
Freeway Ramps	2,107	0.24	0.00	1.04	0.19	0.14	0.00	0.10	0.00	1.04	0.05
Frontage Roads	1,369	2.33	0.08	5.12	1.16	1.82	0.15	1.97	0.22	5.26	1.39
Overall	16,946	4.46	0.90	8.16	5.49	2.24	0.44	5.34	1.84	8.62	6.17

nounced in the links of major and minor arterials compared with other facility types (see Table 3). Freeways and freeway ramps are affected the least, suggesting that district-level allocation errors have limited effect on high-capacity facilities serving mainly regional and interregional movements. Only in the case of Test B, where very large error (± 40 percent random error) is introduced, did almost 20 (i.e., 1.02 percent as shown in Table 3) freeway links display more than half-lane error. Facilities serving local and within-district movements, such as collectors, are affected much less compared with major and minor arterials. It appears that errors in link volumes gradually accumulate from local facilities to higher-level facilities, such as major and minor arterials that principally serve interdistrict travel.

Congested links [links with volume/capacity ratio (V/C) larger than 1.0] are less sensitive to increases in traffic due to traffic diversion, so a comparison of V/C by facility type was undertaken as shown in Table 4. The comparison indicates that, with the increase in magnitude of input errors, the proportion of congested links (V/C more than 0.8) increases, but the increase is low across all facility types. Because almost 60 percent of freeway links are not congested, low sensitivity of freeways, as observed earlier, appears to be less influenced by the saturation of freeway links. However, in areas of freeway congestion such as the northern Dallas, some traffic diversion from freeways may have occurred, but only in the case of traffic increase. Highway links most affected by the input errors fall below the 0.8 V/C ratio range.

To illustrate the impact of geographical bias in allocation, Test C results are further stratified by three geographical areas: areas with employment increase, areas with employment decrease, and areas with no change (see Table 5). Most road links experiencing more than half-lane error are situated in areas where allocation errors are made. Links with positive

TABLE 4 DISTRIBUTION OF HIGHWAY LINKS BY FACILITY TYPE AND V/C RATIO

	Distribution of Links (in %)			% Change in the Distribution of Links Compared to Base					
				Errors in District Level					
	Base Case			Test A +/- 20 Percent Random Error			Test B +/- 40 Percent Random Error		
Facility Type	< 0.8	0.8 - 1.0	> 1.00	< 0.8	0.8 - 1.0	> 1.00	< 0.8	0.8 - 1.0	> 1.00
Freeway	59.92	10.89	29.20	0.10	-0.31	0.20	-1.93	0.15	1.78
Major Arterial	47.18	11.85	4.97	-2.24	0.17	-0.11	-2.40	-0.11	2.52
Minor Arterial	75.23	8.93	15.84	0.05	0.56	-0.61	-0.64	1.04	-0.40
Collector	89.40	3.24	7.36	-0.15	0.03	0.12	0.07	-0.17	0.10
Freeway Ramps	79.69	6.03	14.29	-0.33	0.09	0.24	-1.00	0.90	0.09
Frontage Roads	81.67	4.60	13.73	-0.37	0.15	0.22	-1.17	0.07	1.10
Overall	76.55	6.75	16.69	-0.11	0.14	-0.04	-0.81	0.30	0.52

TABLE 5 LANE ERRORS BY GEOGRAPHICAL AREAS UNDER TEST C (GEOGRAPHICAL BIAS)

Facility Type	Number of Links	Areas with No Change		Areas w/Employment Increase		Areas w/Employment Decrease	
		+0.5 Lane	->0.5 Lane	+0.5 Lane	->0.5 Lane	+0.5 Lane	->0.5 Lane
Freeways	1966	0	0	0	0	0	0
Major Arterial	1789	7	13	89	2	4	61
% of Links		0.39	0.73	4.97	0.11	0.22	3.41
Minor Arterial	3751	6	1	86	2	0	86
% of Links		.16	0.03	2.29	0.05	0.00	2.29
Collector	5964	2	11	41	4	0	10
% of Links		0.03	0.18	0.69	0.07	0.00	0.17
Freeway Ramps	2107	0	0	0	0	0	3
% of Links		0.00	0.00	0.00	0.00	0.00	0.14
Frontage Roads	1369	2	6	12	1	0	6
% of Links		0.15	0.44	0.88	0.07	0.00	0.44
Total	16946	17	31	228	9	4	166
% of Links	100.00	0.10	0.18	1.35	0.05	.02	0.98

(overestimation of traffic) and negative lane error are concentrated in areas with employment increase and decrease, respectively. For instance, out of 455 road links with greater than half-lane error, 228 links serving areas of employment increase indicated positive error, and 166 links situated in areas where employment is reduced indicated negative error. As observed earlier, the most affected links are concentrated in the categories of major and minor arterials. Overall, although the percentage of links severely affected appears low (2.68 percent with more than half-lane error) due to a small magnitude of geographical bias in district inputs (Test C), the affected number of links is high enough (176 major and 181 minor arterial links) to cause misallocation of public resources.

Errors in the Disaggregation of District-Level Inputs to Zone Level

Tests D (± 40 percent random error) and E (uniform distribution) are extreme cases of subdistrict allocation errors. Test E presents a case where zone-level forecasts are prepared with no consideration given to zonal capacity, zoning policy, or other major factors influencing the attractiveness of a zone for development (e.g., transportation accessibility, availability of public services, existing development, etc.). Test D, however, reflects a case of large random error in allocation. To illustrate the level of disturbance caused under each of the two scenarios, R^2 values for zonal population and employment are estimated by comparing the inputs for the base (no error) and individual test runs separately (Table 6). In this case, the R^2 value reflects the strength of association between the base case inputs and particular test run inputs. A value of 1 represents a perfect match between the two sets, and 0 means no match. The higher values of R^2 (0.917 for population and 0.915 for employment) observed under Test D (± 40 percent random error) clearly indicate that this test does not cause as large a deviation from the base case allocation as Test E (uniform distribution). Actually, uniform distribution under Test E causes an extremely large error in zonal inputs.

TABLE 6 CORRELATION BETWEEN BASE YEAR INPUTS AND SUBDISTRICT ALLOCATION TEST INPUTS

Subdistrict Allocation Tests	R-Square Population	Employment
Test D: (±40% Random Error)	0.917	0.915
Test E: (Uniform Distribution)	0.683	0.714

A comparison between the base and test runs for total trips by purpose, areawide RMSE, and trip lengths by purpose indicates no visible impact of input errors on systemwide output measures (see Table 2). However, the lane error between the two tests varied with the overall magnitude of errors introduced under each test. For instance, large allocation errors caused under Test E (uniform distribution) in comparison with Test D (±40 percent random error) led to the observed large difference in percentages of links severely affected [7.2 and 14.8 percent of links with more than half-lane error under Tests D and E, respectively (see Table 3)].

The test results also suggest that, much like the district-level allocation tests, the magnitude of lane errors in each road facility class varied noticeably in all categories. Major and minor arterials are the most affected facilities. A comparison of the results of Tests D and B, where ±40 percent errors are introduced in a random fashion at the subdistrict and district levels, respectively, indicates that the percentage of links showing more than half-lane error is much higher for district-level error (Test B) than for subdistrict-level error (Test D). This is true across all road facility types (see Table 3). For example, compared with almost 17 percent of major arterial links affected by more than half-lane error under Test D (±40 percent random error at the district level), 39 percent of major arterial links are affected under Test B (±40 percent random error at subdistrict level). However, if the results of Test E (uniform distribution) and Test B are compared, a more or less similar magnitude of lane errors across all facility types is seen, except for collectors. This is because, under Test E, an extremely large error is introduced in zonal inputs. Moreover, collectors are expected to be more affected by errors in zonal inputs because they mainly serve local trips.

In summary, the traffic effects of errors in subdistrict allocation are similar to district allocation errors in the facilities most affected and the association between the magnitude of input errors and traffic volume errors. The degree of sensitivity, however, appears to vary substantially. For example, major and minor arterials and expressways are more susceptible to errors in district-level allocation than subdistrict allocation. Local roads, on the other hand, are more affected by errors in zone-level than in district-level inputs.

In the real world, the likelihood of experiencing large subdistrict allocation errors of the magnitude of Test E (uniform distribution) is extremely low for two reasons. First, most subdistrict allocation procedures distribute the incremental growth of a district among its zones and assume almost no change in existing developments (except in cases where large renovations or revitalization schemes are planned). Second, these procedures, in general, take into account major factors influencing development in a zone (e.g., availability of public services, zoning, accessibility, etc.). Areas with the greatest potential for large deviations from anticipated growth are undeveloped areas usually located at the fringe of a city. In these areas, there are always uncertainties linked to market forces influencing the location of activities. Moreover, there is often more than one site competing for new development.

Disaggregation of District-Level Trip Table to Zone Level

Under Test F, an 800 × 800 trip interchange table is first collapsed into a 147 × 147 table and then expanded back to an 800 × 800 table using the number of households and total employment in zones as the proportioning factors for trip productions and attractions, respectively. Such a trip table expansion procedure assumes that the interzonal accessibility, usually measured in terms of travel time or cost, does not influence the magnitude of interzonal interactions. In reality, however, an inverse relationship between the magnitude of interactions among areas and their spatial separation (or impedance) is common. To check the validity of this assumption, the trip length frequency curves for the original trip table and Test F are compared in Figure 2.

The comparison indicates that the average trip length increased from its original value of 13.62 to 15.30 min after application of the disaggregation procedure, as evident from the trip length frequency curves shown in Figure 2. The possible explanation for this is that because of the omission of the accessibility effect from the zone-level trip distribution, the trip interchange volume between zones that are highly accessible to each other is likely to be underestimated, and volumes between zones that are farther apart will be overestimated. The result is an increase in the proportion of longer trips.

As expected, because of the increase in the share of longer trips, the highway assignment produces significant positive lane error within each class of links (see Table 7). Both major and minor arterials experienced substantial increases in traffic, almost as high as under Test B (±40 percent random errors in district level). Actually, the proportion of links with more than one lane error is extremely high under this scenario compared with all earlier tests. For instance, the percentage of major arterial links with more than half-lane error under Tests B and F are 39.07 and 40.75, respectively (compare Tables 3 and 7). But the percentage of links with more than one lane error is 16.6 and 23.53 for Tests B and F, respectively. It is obvious that the preceding kind of trip table stratification procedure can cause high prediction of trip volumes on major facilities, leading to their overdesign.

Tests results indicate that a procedure used for the disaggregation of a district-level trip to zone level must account for interzonal accessibility. Exclusion of accessibility variation among zones may produce large errors in UTPP outputs. The magnitude of overall error will greatly depend on the size of districts—the magnitude of expansion (i.e., ratio of total number of zones and districts). In general, the smaller the size of the districts, the lower the potential for introducing large errors in the trip table splitting process. This is true because the smaller districts would account for greater variations in interzonal accessibility than the larger districts.

In the light of these findings, it is recommended that travel demand models (for instance, UTPP) be applied at the traffic zone level so that interzonal trip tables are produced directly and the need for trip table stratification is avoided. In the

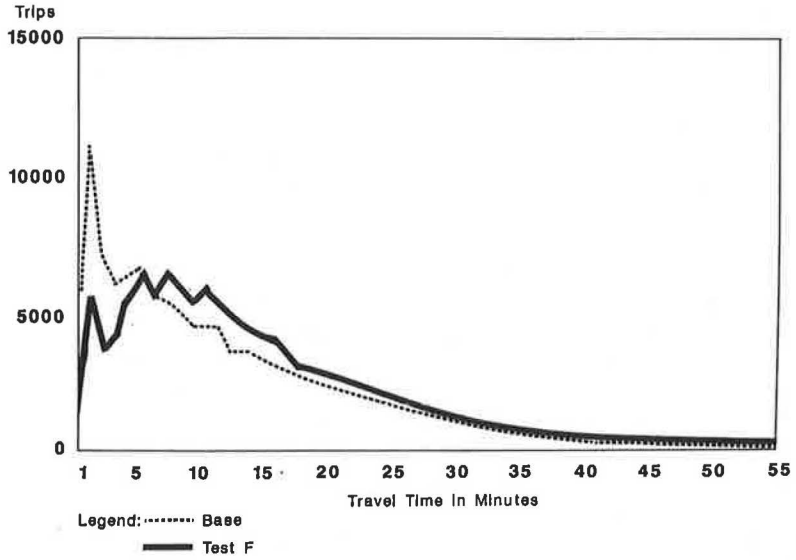

FIGURE 2 Comparison of trip length frequency curves.

TABLE 7 DISTRIBUTION OF HIGHWAY LINKS BY LANE ERROR UNDER TEST F (SPLITTING OF DISTRICT-LEVEL TRIP TABLE)

	Percentage of Links			
	Negative Error		Positive Error	
Facility Type	0.5 Lane	> 1 Lane	0.5 Lane	> 1 Lane
Freeway	0.00	0.00	0.51	0.00
Major Arterial	2.35	0.95	14.87	22.58
Minor Arterial	1.52	0.41	13.01	13.74
Collector	0.89	0.32	4.59	3.30
Freeway Ramps	0.05	0.00	0.81	0.05
Frontage Roads	1.02	0.00	6.22	6.90
Overall	0.99	0.29	6.44	6.76

case of site-level analysis (or local area network planning), traffic zones are sometimes further divided into smaller spatial units, and the zone-to-zone trip table is once again disaggregated to develop a new trip table. Because of the splitting of smaller zones, the potential for large errors is reduced, although it is not fully eliminated.

SUMMARY OF FINDINGS

To illustrate the effect of changes in socioeconomic input data on the estimation of network traffic volumes, five scenarios representing various types and magnitudes of changes in district and subdistrict allocations, referred to here as errors, were tested (see Table 8). The Dallas–Fort Worth area was selected as the test case. Three out of five scenarios reflected the random nature of allocation errors. These scenarios were created by randomly selecting one-third of districts or zones for positive, one-third for negative, and one-third for no errors. One district-level allocation scenario replicated geographical bias in land use allocation that may occur because of the rapid flight of jobs to the suburbs. The test represents the contemporary problem of land use forecasting in urban areas that are witnessing unexpectedly high suburban growth and decline or modest growth in the central city. The scenario illustrated the impact of employment underprediction in one of the suburban sectors and overprediction in the CBD and one of its adjacent sectors. Among subdistrict allocation scenarios, one test showed the effect of distributing district-level inputs equally (or uniformly) among the zones of each district. It represented a special case of subdistrict allocation where zone-specific land use forecasts are completely insensitive to zone capacity, zoning policy, and major factors influencing the potential for development in a zone (existing development, availability of utilities, accessibility, etc.).

In addition, a separate test was performed to examine the impact of assigning a zone-to-zone trip interchange table that is produced by disaggregating a district-to-district trip table instead of using a trip table generated directly by the application of travel demand models at the zone level.

For each scenario the demand sensitivity of facilities was measured in terms of the proportion of links experiencing large differences in traffic volumes (increase or decrease by more than half-lane capacity) after the introduction of input

TABLE 8 FINDINGS OF SENSITIVITY TESTS

Type	Allocation Error		Traffic Impact on Facilities		
	Nature	Magnitude	Expressway	Major/Minor Arterial	Collectors
From Region to District	Random	Small	Not significant	Moderate	Low
	Random	Large	Low	High	Low/Moderate
	Geographical Bias	Small	Not Significant	Moderate but concentrated in areas with input errors	Low
From District to Zones	Random	Large	Not Significant	Moderate	Low/Moderate
	Uniform	Large	Low	High	Moderate
Splitting of Trip Table			Low	High	Moderate/High

errors. The findings of the case study are summarized in Table 8. Broad conclusions drawn from the sensitivity tests are as follows:

1. The severity of traffic prediction errors increases with the overall magnitude of activity allocation error but not uniformly across all types of road facilities. Major and minor arterials are most sensitive to input errors, followed by local roads and expressways.
2. Errors in district and zone-level forecasts influence each facility type differently. Facilities serving major interdistrict movements, such as major and minor arterials and expressways, are more sensitive to errors in district-level allocation than subdistrict (or zone-level) allocation. Local roads, on the other hand, are affected more by the errors in zone-level rather than district-level inputs.
3. A small magnitude of geographical bias in district-level forecasts may not produce significant systemwide impacts, but it can severely affect facilities near districts with input errors.
4. Subdistrict allocation procedures must take into consideration the zonal capacity and factors influencing the attractiveness of a zone for development. Insensitivity to these may produce large errors in zonal inputs and, in turn, traffic forecasts across all types of facilities.
5. The practice of trip table expansion (or splitting) to develop trip interchange tables for smaller spatial units introduces large errors in traffic forecasts. This is mainly due to the omission of an accessibility factor influencing trip interchanges.

ACKNOWLEDGMENT

The work reported in this paper was performed under a project sponsored by the National Cooperative Highway Research Program of the National Research Council. The author acknowledges the assistance and contribution of COMSIS staff, including James Ryan, Michael Doherty, Arthur B. Sosslau, David M. Levinsohn, and Gordon Schultz. Sincere thanks are extended to the staff of the Transportation and Energy Department of NCTCOG, who contributed significant time and effort in undertaking the sensitivity runs.

REFERENCES

1. J. N. Bajpai. *NCHRP Report 328: Forecasting the Basic Inputs to Transportation Planning at the Zonal Level*. TRB, National Research Council, Washington, D.C., 1990.
2. D. H. Pickrell. *A Comparative Analysis of Forecasts and Actual Ridership and Costs for Ten Federally Supported Urban Rail Transit Projects*. Urban Mass Transportation Administration, U.S. Department of Transportation, 1988.
3. CONSAD Research Corporation. *Sensitivity Analysis of the Urban Travel Forecasting Process*. Bureau of Public Roads, U.S. Department of Transportation, 1968.
4. F. S. Koppelman. Methodology for Analyzing Errors in Prediction with Disaggregate Choice Models. In *Transportation Research Record 592*, TRB, National Research Council, Washington, D.C., 1976.
5. S. H. Putman. *Integrated Urban Models*. Pion Limited, London, 1983.
6. *Multimodal Transportation Analysis Process (MTAP): A Travel Demand Forecasting Model*. North Central Texas Council of Governments, 1990.
7. P. Prastacos. Urban Development Models for the San Francisco Region: From PLUM to POLIS. In *Transportation Research Record 1046*, TRB, National Research Council, Washington, D.C., 1985.
8. W. Goldner et al. *Projective Land Use Model—PLUM: Theory and Application*. Institute of Transportation Engineering, University of California, Berkeley, 1972.

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

Forecasting Intercity Rail Ridership Using Revealed Preference and Stated Preference Data

TAKAYUKI MORIKAWA, MOSHE BEN-AKIVA, AND KIKUKO YAMADA

A methodology for incorporating revealed preference (RP) and stated preference (SP) data in discrete choice models is presented. The methodology is applied to intercity travel mode choice analysis. New mode shares for each origin-destination pair resulting from changes in service levels are predicted. The combined estimation technique with RP and SP data is developed to promote advantages of the two complementary data sources. The empirical study of intercity travel demand demonstrates the practicality of the methodology by accurately reproducing observed aggregate data and by applying a flexible operational prediction method.

Travel demand models are usually estimated with observations of actual behavior, or revealed preference (RP) data, using the methods of discrete choice analysis [e.g., Ben-Akiva and Lerman (*1*)]. However, in estimating individual choice models RP data may be deficient for the following reasons:

1. RP data do not provide information on preferences for nonexisting services;
2. The choice set considered by the decision maker may be ambiguous;
3. Some service attributes are measured with error; and
4. Some attributes are highly correlated or lack variability, or both.

The drawbacks can be alleviated to a great extent in a survey with hypothetical choice scenarios and fully controlled alternatives. Such experimental data are called stated preference (SP) data, and they have been used by a number of travel demand researchers [e.g., Louviere et al. (*2*), Bates (*3*), and Hensher et al. (*4*)] as well as in marketing research [e.g., Green and Srinivasan (*5*) and Cattin and Wittink (*6*)]. However, the applications of SP data in practical transportation studies are still limited because of the uncertain reliability of elicited preferences under hypothetical scenarios. Advantages and disadvantages of RP and SP data and potential biases specific to SP data are discussed in detail by Ben-Akiva et al. (*7*).

Because RP and SP data have complementary characteristics, this paper explores the idea of using both types of data simultaneously. The methodology includes explicit consideration of the unknown reliability of SP data, and its objective is to yield more reliable travel demand models than those produced by separate or sequential SP and RP analyses. The following context explains the main idea of the paper. Trade-offs among certain attributes often cannot be estimated accurately from available RP data. For instance, high correlation between travel cost and travel time in RP data may yield insignificant parameter estimates for their coefficients. However, an SP survey with a design based on low or zero correlation between these attributes may provide additional information on their trade-offs. Although the SP responses may not be valid for forecasting actual behavior due to their unknown bias and error properties, they often contain useful information on trade-offs among attributes. SP data also add critically important information on preferences in the introduction of new services, such as a new type of high-grade passenger car in rail service. RP data alone cannot provide the information needed to assess the impact of such a new service.

In previous papers the authors have proposed a methodology for statistically combining RP and SP data in estimating travel demand models (*8*, *9*). The key features of the methodology are

- Bias correction (explicit response models for SP data that include both preference and bias parameters),
- Efficiency (joint estimation of preference parameters from all the available data), and
- Identification (estimation of trade-offs among attributes and the effects of new services that are not identifiable from RP data).

The objective of this paper is to demonstrate the effectiveness of the combined RP/SP estimation method by an application to predict intercity rail ridership in conjunction with changes in service quality. The changes in service considered include the introduction of a high-grade passenger car, which could not be evaluated by analyzing RP data only.

METHODOLOGY

Model Specification

Two different model types are considered: RP and SP models. The RP model represents market behavior by some appropriate structure (e.g., random utility model with discrete

T. Morikawa, Department of Civil Engineering, Nagoya University, Chikusa-ku, Nagoya, 464-01, Japan. M. Ben-Akiva, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139. K. Yamada, Social System Department, II, Mitsubishi Research Institute, Inc., 2-3-6 Otemachi, Chiyoda-ku, Tokyo, 100, Japan.

choices), whereas SP response is modeled by the SP model. As discussed earlier, although SP data might not be valid for forecasting market behavior due to unknown bias and random error properties, they often contain useful information on trade-offs among attributes and preferences for nonexistent services. Thus, the role of SP data is illustrated by the following framework:

RP model:

$$u_{in}^{RP} = \beta' \mathbf{x}_{in}^{RP} + \alpha' \mathbf{w}_{in}^{RP} + \varepsilon_{in}^{RP}$$
$$= v_{in}^{RP} + \varepsilon_{in}^{RP}$$
$$i = 1, \ldots, I_n^{RP}, n = 1, \ldots, N^{RP} \quad (1)$$

$$d_n^{RP}(i) = \begin{cases} 1 & \text{if Alternative } i \text{ is chosen by Individual } n \text{ in the RP data} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

SP model:

$$u_{in}^{SP} = \beta' \mathbf{x}_{in}^{SP} + \gamma' \mathbf{z}_{in}^{SP} + \varepsilon_{in}^{SP}$$
$$= v_{in}^{SP} + \varepsilon_{in}^{SP}$$
$$i = 1, \ldots, I_n^{SP}, n = 1, \ldots, N^{SP} \quad (3)$$

$$d_n^{SP}(i) = \begin{cases} 1 & \text{if Alternative } i \text{ is chosen by Individual } n \text{ in the SP data} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where

u_{in} = utility of Alternative i to Individual n,
v_{in} = systematic component of u_{in},
ε_{in} = random component of u_{in},
$d_n(i)$ = choice indicator of Alternative i for Individual n,
$\mathbf{x}_{in}, \mathbf{w}_{in}, \mathbf{z}_{in}$ = vectors of explanatory variables of Alternative i for Individual n, and
α, β, γ = vectors of unknown parameters.

The superscript RP or SP indicates the data type.

In this framework, it is assumed that the SP response is a "choice" or the most preferred alternative presented to the respondent. Even when the SP response is given by other formats, such as preference ranking or pairwise comparison with categorical response, the SP model can be based on the same random utility model. A different response format only requires a slightly different estimation method.

The term represented by $\gamma'\mathbf{z}$ is specific to the SP model and may include SP biases and effects of hypothetical new services that are included only in the SP survey. The appearance of β in both models implies that the trade-offs among the attributes in the vector $\mathbf{x}$ are the same in both actual market behavior and the SP tasks.

The level of random noise in the data sources is represented by the variance of the disturbance term ε. If RP and SP data have different noise levels, this can be expressed by

$$\text{Var}(\varepsilon_{in}^{RP}) = \mu^2 \text{ Var}(\varepsilon_{in}^{SP}) \qquad \forall\, i, n \quad (5)$$

If SP data contain more random noise than RP data, μ will lie between 0 and 1. μ is also known to represent the "scale" of the model coefficients.

Assuming independently and identically distributed (i.i.d.) Gumbel disturbance terms in the RP model, a logit model is obtained with the choice probability given by

$$P_n^{RP}(i) = \frac{\exp(v_{in}^{RP})}{\sum_{j=1}^{I_n^{RP}} \exp(v_{jn}^{RP})} \quad (6)$$

An i.i.d. Gumbel assumption for the SP utility disturbances leads to the following SP logit model, which includes the scale parameter μ:

$$P_n^{SP}(i) = \frac{\exp(\mu \cdot v_{in}^{SP})}{\sum_{j=1}^{I_n^{SP}} \exp(\mu \cdot v_{jn}^{SP})} \quad (7)$$

Model Estimation

The unknown parameter vectors, α, β, and γ and the scale parameter μ are jointly estimated using both RP and SP data. The log-likelihood functions for the RP and SP data sets are given by

$$L^{RP}(\alpha, \beta) = \sum_{n=1}^{N^{RP}} \sum_{i=1}^{I_n^{RP}} d_n^{RP}(i) \ln P_n^{RP}(i) \quad (8)$$

$$L^{SP}(\beta, \gamma, \mu) = \sum_{n=1}^{N^{SP}} \sum_{i=1}^{I_n^{SP}} d_n^{SP}(i) \ln P_n^{SP}(i) \quad (9)$$

Separately maximizing Equations 8 and 9 yields maximum likelihood estimators of the RP and SP models, respectively. In that case the scale parameter μ and the coefficients are not separable in the SP model.

By maximizing the sum of Equations 8 and 9 we can force the β coefficients to be the same in the RP and SP models. Thus, the combined RP/SP estimator is obtained by maximizing the joint log-likelihood function:

$$L^{RP+SP}(\alpha, \beta, \gamma, \mu) = L^{RP}(\alpha, \beta) + L^{SP}(\beta, \gamma, \mu) \quad (10)$$

This estimator fully uses the information contained in both RP and SP data as discussed above. If the random terms of the RP and SP models for the same individual are assumed to be statistically independent, maximizing Equation 10 will yield the maximum likelihood estimator of all the parameters. If the random terms are not independent, this estimator is consistent, but the standard errors of the estimates calculated in the usual way are incorrect (*10*).

Because the joint log-likelihood function (Equation 10) is not linear in parameters due to the introduction of μ, the estimation cannot be carried out using ordinary MNL software packages for logit models. If the response format of the SP data is choice, a program to estimate a nested logit model

may be used. Alternatively, the following sequential estimation method using ordinary software packages may yield consistent but less efficient estimates.

Step 1. Estimate the SP model (Equation 3) by maximizing Equation 9 using the SP data to obtain $\widehat{\mu\beta}$ and $\widehat{\mu\gamma}$. Define $y_{in}^{RP} = \mu\beta' \mathbf{x}_{in}^{RP}$ and calculate the fitted value $\hat{y}_{in}^{RP} = \widehat{\mu\beta}' \mathbf{x}_{in}^{RP}$ for the RP observations.

Step 2. Estimate the following RP model with the fitted value $\hat{y}_{in}^{RP}$ included as a variable to obtain $\hat{\lambda}$ and $\hat{\alpha}$:

$$\hat{u}_{in}^{RP} = \lambda \hat{y}_{in}^{RP} + \alpha' \mathbf{w}_{in}^{RP} + \varepsilon_{in}^{RP} \quad (11)$$

where $\lambda = 1/\mu$. Calculate $\hat{\mu} = 1/\hat{\lambda}$, $\hat{\beta} = \widehat{\mu\beta}/\hat{\mu}$, and $\hat{\gamma} = \widehat{\mu\gamma}/\hat{\mu}$. The accuracy of $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ can be improved by Step 3.

Step 3. Multiply $\mathbf{x}^{SP}$ and $\mathbf{z}^{SP}$ by $\hat{\mu}$ to obtain a modified SP data set. Pool the RP data and the modified SP data and then estimate the two models jointly to obtain $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$.

In this paper the joint estimator is employed. It was implemented in a special program written in GAUSS.

Prediction with the RP/SP Models

For prediction only the RP model is used because our concern is actual behavior, not experimental response. Therefore, the systematic utility component used for prediction is given by

$$\hat{v}_{in} = \hat{\beta}' \mathbf{x}_{in} + \hat{\alpha}' \mathbf{w}_{in} \quad (12)$$

Note that $\hat{\beta}$ in Equation 12 is estimated using both RP and SP data. If some hypothetical services presented in the SP questions are to be included for predicting demand, the corresponding term in the SP model should be added to Equation 12, as follows:

$$\hat{v}_{in} = \hat{\beta}' \mathbf{x}_{in} + \hat{\alpha}' \mathbf{w}_{in} + \hat{\bar{\gamma}}' \bar{\mathbf{z}}_{in} \quad (13)$$

where $\bar{\mathbf{z}}_{in}$ is a subvector of $\mathbf{z}_{in}$ representing hypothetical attributes relevant to the policy changes and $\hat{\bar{\gamma}}$ is an estimate of the parameters on $\bar{\mathbf{z}}_{in}$.

Terms from the RP and SP utility functions can be combined, as shown in Equation 13, because the scale of the utilities is adjusted between the RP and SP models by introducing the scale parameter μ.

CASE STUDY—ESTIMATION OF INTERCITY MODE CHOICE MODELS

Description of Survey Data

The survey was conducted to assess intercity rail ridership in conjunction with a planned replacement of regular cars by high-grade cars on trains. The alternative travel modes in the study corridor are express bus (or coach) service and private cars. The corridor connects two districts between which it takes 2 to 3 hr by rail and 4 to 6 hr by bus and car. Currently the corridor is covered by 26 daily trains, of which four have high-grade cars. Because there is no difference in rail fare between regular and high-grade trains, the high-grade trains are always fully booked. The rail operator is considering the upgrading of additional trains and would like to know how many new rail passengers will be attracted from the competing modes.

A survey of passengers traveling in the corridor was conducted using pure choice–based random sampling for the three competing modes. The questionnaire asked for the socioeconomic characteristics of the traveler, the attributes of the chosen mode, and availability of alternative modes. Level-of-service attributes, such as travel time and cost for the chosen and unchosen modes, were calculated using network data for the reported origin and destination of the trip.

Each respondent was also asked for a preference ranking of the three alternative modes under the following hypothetical scenarios: for rail passengers, Scenario 1 (status quo) and Scenario 2 (better access to the bus terminal); and for bus and car passengers, Scenario 1 (status quo), Scenario 2 [increase in frequency of high-grade trains (13 services daily)], Scenario 3 (reduction in rail line-haul travel time by 10 percent), and Scenario 4 (reduction in rail line-haul travel time and increase in frequency of high-grade trains). Respondents were asked to rank in order the three travel modes under each scenario.

The numbers of usable responses were 274, 89, and 82 from rail, bus, and car passengers, respectively. Those who said that they had no available modes other than the chosen one are assumed to be "captive" to the chosen mode. One hundred thirty-three respondents were captive to rail and 17 and 40 to bus and car, respectively. Captives are excluded from the calibration data set, but they are included in the prediction of aggregate ridership.

Estimation Results

Three models were estimated: RP model, SP model, and combined RP/SP model, each of which was estimated by maximizing the corresponding log-likelihood function (see Equations 8 through 10). The independent variables include line-haul travel time for rail and bus and total travel time for car (in hours), travel time for access and egress trips for rail and bus (in hours), travel cost per person (in thousands of yen), and business trip dummy (1 if the trip is associated with a business purpose, 0 otherwise). The last variable interacts with travel time and cost.

Because pure choice–based sampling was employed, the estimates of the alternative specific constants should be adjusted by the following correction formula (*11*):

$$\hat{\beta}_0^i = \hat{\beta}_0^i - \log \frac{H_i}{W_i} \quad (14)$$

where

$\hat{\beta}_0^i$ = adjusted estimate of the constant for Alternative i,
$\hat{\beta}_0^i$ = estimate of the constant for Alternative i through the exogenous sample maximum likelihood,
H_i = share of Alternative i in the sample (for SP models sample share must reflect the repetitions of the SP questions for each respondent), and

W_i = market share of Alternative i in the population.

The RP model estimated from the RP data is given in the first column of Table 1. The value of line-haul travel time for a business trip is approximately 1,500 yen per hour, or \$10/hr.

Estimation of the SP model used the SP data from the bus and car passengers to analyze their intention to switch to rail. A choice data set was created by taking the first ranked alternative as the preferred one, or chosen one. Because few respondents had the full choice set (i.e., three alternatives), information on the second ranking was not used. A dummy variable that indicates the increase in frequency of high-grade trains was added to the rail utility.

The second column of Table 1 gives the estimates of the SP model. The high-grade train dummy has a significantly positive coefficient. The rail and bus constants are significantly different from those of the RP model, which may be ascribed to the use of only the bus and car passengers' SP data or to some SP biases. The value of line-haul travel time for business trips is approximately 400 yen per hour, or \$3/hr.

The third column of Table 1 gives the estimation result of the combined RP/SP model. The parameters are calibrated through the joint estimation method. Alternative specific constants are estimated separately from the RP and SP data because the two models show significant differences in those constants. This implies that alternative specific constant terms belong to $\alpha'\mathbf{w}$ and $\gamma'\mathbf{z}$ in the framework of Equations 1 and 3.

The high-grade train dummy has a significantly positive coefficient. The value of line-haul travel time for business trips is approximately 560 yen per hour, or \$4/hr. The scale parameter μ is 1.33, but it is not significantly different from 1.0, which suggests that the variances of the random terms in the RP and SP models are approximately the same.

PREDICTION FROM ESTIMATED MODELS

In this section, two types of aggregation techniques, sample enumeration and representative individual, are applied to the estimated model to predict demand for policy changes.

Sample Enumeration Method

The fitted values of systematic utilities are given by Equation 13, and then the fitted choice probabilities are calculated by substituting these values in the MNL form.

Aggregated demand in the population can be obtained by the sample enumeration method as follows. It is assumed that the ratio of captives for each mode in the population is the same as in the sample. $C(i)$ is defined as the number of captives in the population. The predicted aggregate demand of Alternative i is calculated by Equation 15:

$$\begin{aligned} N(i) &= C(i) + N \cdot S(i) \\ &= C(i) + N \sum_{j=1}^{I} W_j \frac{1}{N_{sj}} \sum_{n=1}^{N_{sj}} \hat{P}_{nj}(i) \\ &= C(i) + N \sum_{j=1}^{I} \frac{N_j}{N} \frac{1}{N_{sj}} \sum_{n=1}^{N_{sj}} \hat{P}_{nj}(i) \\ &= C(i) + \sum_{j=1}^{I} \frac{N_j}{N_{sj}} \sum_{n=1}^{N_{sj}} \hat{P}_{nj}(i) \\ &= C(i) + \sum_{j=1}^{I} E_j \sum_{n=1}^{N_{sj}} \hat{P}_{nj}(i) \qquad i = 1, \ldots, I \end{aligned} \tag{15}$$

where

N_{sj} = number of observations choosing Alternative j in the estimation sample,
N_j = observed number of individuals choosing Alternative j in the population,
N = total number of noncaptive individuals in the population,
$S(i)$ = predicted share of Alternative i,
W_j = observed share of Alternative j in the population,
$\hat{P}_{nj}(i)$ = predicted choice probability of Alternative i for Individual n sampled on Alternative j, and
E_j = an expansion factor defined by N_j/N_{sj}.

Table 2 gives predicted aggregate demand by this method under the same four scenarios as used in the SP questions. Observed aggregate numbers are obtained from on-off counts

TABLE 1 ESTIMATION RESULTS (t-STATISTICS IN PARENTHESES)

Variables	RP Model	SP Model	RP/SP Model
Rail constant (RP)	1.66 (5.4)		1.40 (5.1)
Bus constant (RP)	-1.43 (-5.0)		-1.59 (-5.9)
Rail constant (SP)		0.706 (2.4)	0.906 (4.0)
Bus constant (SP)		-3.37 (-1.6)	-3.24 (-1.9)
High-grade train dummy		0.702 (3.1)	0.520 (2.4)
Line-haul travel time × business trip	-0.458(-1.7)	-0.370 (-0.6)	-0.270 (-1.4)
Terminal travel time × business trip (Rail and Bus)	-0.973(-1.8)	0.232 (0.3)	-0.143 (-0.5)
Total travel cost	-0.402(-5.5)	-0.336 (-4.7)	-0.294 (-4.3)
Business trip dummy × total travel cost	0.102 (0.7)	-0.551 (-1.2)	-0.187 (-1.6)
Scale parameter μ			1.33 (3.6)
N	255	434	689
$L(0)$	-191.35	-332.26	-524.61
$L(\hat{\beta})$	-149.25	-271.18	-427.59
ρ^2	0.220	0.184	0.185
$\bar{\rho}^2$	0.189	0.163	0.166

TABLE 2 PREDICTED ANNUAL TRIPS AND MODAL SHARES BY SAMPLE ENUMERATION (DIFFERENCE FROM VALUES UNDER SCENARIO 1 IN PARENTHESES)

		Rail		Highway Bus		Car	
Observed	Annual Trips	5,365,865		117,237		1,808,940	
	Modal Share	73.6%		1.6%		24.8%	
Scenario 1		5,357,431		113,046		1,821,565	
(status quo)		73.5%		1.5%		25.0%	
Scenario 2		5,646,818	(+289,387)	80,992	(-32,054)	1,564,232	(-257,333)
(increase in high-grade trains)		77.4%	(+3.9%)	1.1%	(-0.4%)	21.5%	(-3.5%)
Scenario 3		5,369,599	(+12,168)	111,719	(-1,327)	1,810,724	(-10,841)
(reduction of rail time)		73.7%	(+0.2%)	1.5%	(-0.0%)	24.8%	(-0.2%)
Scenario 4		5,656,751	(+299,320)	80,112	(-32,934)	1,555,179	(-266,386)
(Scenarios 2 + 3)		77.6%	(+4.1%)	1.1%	(-0.4%)	21.3%	(-3.7%)

for rail and bus trips and screen-line counts for car trips. The observed and predicted numbers under Scenario 1 (status quo) match perfectly because the full set of alternative specific constants estimated from the RP data are used in the predicted utilities. This desirable property of MNL models is obtained by separately estimating alternative specific constants from RP and SP data and using the RP constants for prediction. The table shows that high-grade trains significantly increase rail ridership.

Representative Individual Method

Another aggregation technique employed here is the representative individual method. This method approximates aggregate shares by the choice probabilities of the "representative" individual. The representative individual can be created by calculating averages of attributes in the sample or assigning appropriate attribute values. This method is very operational when the model is transferred to places where disaggregate data are unavailable. However, aggregate predictions by this method have an aggregation bias.

The fitted utility functions are also calculated by Equation 13 with "representative" attribute values. This case study predicts prefectural level origin-destination (O-D) trip tables between the two districts. Each O-D pair is treated as a market segment, and average attribute values for each O-D pair in the sample are used for representative individuals.

Table 3 gives the observed aggregate O-D table, and the predicted one is given in Table 4. The tables agree fairly well, which can be ascribed to good parameter estimates under the proposed method. Although not shown in this paper, predicted O-D tables under different scenarios were calculated.

CONCLUDING REMARKS

The method of combined estimation of discrete choice models from RP and SP data was presented. An empirical case study of intercity travel demand analysis demonstrated the practi-

TABLE 3 OBSERVED O-D TABLE (ANNUAL RIDERSHIPS AND SHARES)

	A1			A2			A3		
	Rail	Bus	Car	Rail	Bus	Car	Rail	Bus	Car
B1	191,768	0	37,230	414,524	2,664	79,570	191,768	1,332	29,565
	(83.7%)	(0.0%)	(16.3%)	(83.4%)	(0.5%)	(16.0%)	(86.1%)	(0.6%)	(13.3%)
B2	1,349,818	0	527,425	1,265,649	27,980	604,805	567,890	13,320	109,135
	(71.9%)	(0.0%)	(28.1%)	(66.7%)	(1.5%)	(31.9%)	(82.3%)	(1.9%)	(15.8%)
B3	475,767	0	146,730	621,082	66,610	218,635	287,600	5,329	55,845
	(76.4%)	(0.0%)	(23.6%)	(68.5%)	(7.3%)	(24.1%)	(82.5%)	(1.5%)	(16.0%)

TABLE 4 PREDICTED O-D TABLE USING REPRESENTATIVE INDIVIDUAL METHOD

	A1			A2			A3		
	Rail	Bus	Car	Rail	Bus	Car	Rail	Bus	Car
B1	181,179	0	45,819	408,038	7,722	81,498	179,301	1,348	42,016
	(80.0%)	(0.0%)	(20.0%)	(82.1%)	(1.5%)	(16.4%)	(80.5%)	(0.6%)	(18.9%)
B2	1,453,364	0	417,541	1,366,426	35,672	496,336	592,188	11,131	87,027
	(77.7%)	(0.0%)	(22.3%)	(72.0%)	(1.9%)	(26.1%)	(85.8%)	(1.6%)	(12.6%)
B3	467,374	0	155,122	659,621	58,552	188,154	300,924	6,383	41,466
	(75.1%)	(0.0%)	(24.9%)	(72.7%)	(6.5%)	(20.8%)	(86.3%)	(1.8%)	(11.9%)

cality of the method. The case study predicted rail ridership under hypothetical scenarios, such as introduction of high-grade trains.

When RP and SP data were used simultaneously to estimate the mode choice model, alternative specific constants were estimated separately from each data set. Using the MNL estimates of the constants from the RP data enables us to reproduce the aggregate shares through the sample enumeration method. Aggregation by the representative individual method also accurately reproduced the observed O-D table. This is an encouraging result for using the combined estimation method and predicting demand under hypothetical scenarios.

The work presented in this paper and two previous studies (*8*, *9*) has shown the effectiveness and practicality of combined estimation with RP and SP data. This paper provided further evidence. However, more empirical work in different contexts may be needed to justify the methodology conclusively. In addition, the authors are developing more efficient estimators that explicitly treat potential correlation between the random utilities of RP and SP models for the same individual.

REFERENCES

1. M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis*. MIT Press, Cambridge, Mass., 1985.
2. J. J. Louviere, D. H. Henly, G. Woodworth, R. J. Meyer, I. P. Levin, J. W. Stones, D. Curry, and D. A. Anderson. Laboratory-Simulation Versus Revealed-Preference Methods for Estimating Travel Demand Models. In *Transportation Research Record 794*, TRB, National Research Council, Washington, D.C., 1981, pp. 42–51.
3. J. Bates. Stated Preference Technique for the Analysis of Transportation Behavior. *Proc., 3rd WCTR*, Hamburg, West Germany, 1983, pp. 252–265.
4. D. A. Hensher, P. O. Barnard, and T. P. Thruong. The Role of Stated Preference Methods in Studies of Travel Choice. *Journal of Transport Economics and Policy*, Vol. 22, No. 1, 1988, pp. 45–58.
5. P. E. Green and V. Srinivasan. Conjoint Analysis in Consumer Research: Issues and Outlook. *Journal of Consumer Research*, Vol. 5, 1978, pp. 103–123.
6. P. Cattin and D. R. Wittink. Commercial Use of Conjoint Analysis: A Survey. *Journal of Marketing*, Vol. 46, 1982, pp. 44–53.
7. M. Ben-Akiva, T. Morikawa, and F. Shiroishi. Analysis of the Reliability of Stated Preference Data in Estimating Mode Choice Models. *Selected Proc., 5th WCTR*, Vol. 4, Yokohama, Japan, 1989, pp. 263–277.
8. M. Ben-Akiva and T. Morikawa. Estimation of Switching Models from Revealed Preferences and Stated Intentions. *Transportation Research A*, Vol. 24A, No. 6, 1990, pp. 485–495.
9. M. Ben-Akiva and T. Morikawa. Estimation of Travel Demand Models from Multiple Data Sources. *Proc., 11th International Symposium on Transportation and Traffic Theory* (M. Koshi, ed.), Elsevier, 1990, pp. 461–476.
10. T. Amemiya. *Advanced Econometrics*. Harvard University Press, Cambridge, Mass., 1985.
11. C. Manski and S. Lerman. The Estimation of Choice Probabilities from Choice-Based Samples. *Econometrica*, Vol. 45, 1977.

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

Stochastic Process Approach to the Estimation of Origin-Destination Parameters from Time Series of Traffic Counts

GARY A. DAVIS AND NANCY L. NIHAN

The origin-destination (OD) matrix gives the volume of traffic from each of a region's origins to each of its destinations and is a fundamental input to transportation planning and network design activities. Because the traditional methods of estimating the OD matrix—surveys and trip generation/distribution modeling—tend to be expensive, cumbersome, and inaccurate, researchers have sought to develop methods for estimating the OD matrix from observations of traffic volumes on the region's road network. For simple linear networks, such as single intersections or freeway sections, OD estimators with desirable statistical properties can be developed using least-squares methods, but for general networks it has not yet been possible to produce consistent estimators of OD parameters using traffic count data alone. It is believed that the link counts on a traffic network are generated by a stochastic process that is parameterized by the means and variances of the separate OD flows. By using a tractable approximation to the traffic-generating process, it is possible to develop both maximum likelihood and method of moments estimators of OD parameters, and the estimators have desirable consistency and asymptotic normality properties. Simulation studies suggest that the maximum likelihood estimator, though efficient in its use of data, is computationally demanding, whereas the method of moments estimator is not computationally demanding but is statistically inefficient.

In transportation modeling, the origin-destination (OD) matrix is an array whose rows index the locations on a network where trips originate and whose columns index the locations where trips terminate. The entry at the intersection of a row and a column gives, for some predetermined time interval, the number of trips between that particular OD pair. The OD matrix is the fundamental summary of a region's demand for travel, so it is an important input to any transportation planning activity. Historically, OD matrices have been estimated using some combination of survey methods and trip generation/distribution modeling, but the data collection needed for these approaches tends to be time-consuming and expensive, and the result is often of unreliable accuracy (*1*). Because the OD matrix can be viewed as an input to a traffic assignment process the outputs of which are the traffic volumes on the network's links, an alternative approach to OD matrix estimation is to start with observed link volumes and somehow "invert" the traffic assignment to obtain the OD matrix. This approach appears increasingly attractive as the proliferation of automatic traffic surveillance and control systems makes automatic traffic count data more readily available. The clearest formal statement of this OD estimation problem to date has been given by Cascetta and Nguyen (*2*), although related work appears in Maher (*3*), Bell (*4*), and Spiess (*5*).

Before considering methods of OD estimation based on traffic count data, it is useful to review desirable properties of parameter estimators. An estimator is said to be consistent if the probability of large discrepancies between it and the true parameter value approaches zero as the amount of data used approaches infinity. Thus for large amounts of data, a consistent estimator is likely to be close to the true parameter value, and consistency can be regarded as a minimal necessary condition for an estimator to be useful. Consistency is a property of point estimators, but in addition to generating a point estimate of a parameter, it is often desirable to be able to compute confidence bounds for the estimate or to test hypotheses concerning parameter values. For a large class of estimators, including many maximum likelihood (ML) estimators, a useful theory of inference can be developed on the basis of the fact that, as the amount of data becomes large, the estimator tends to have a normal distribution. This property is called asymptotic normality. Because a parameter estimator is a random variable, it has a variance, and larger variances indicate greater uncertainty concerning the true parameter value. If the variance of a particular estimator about the true parameter value is lower than that of any other estimator, that estimator is called efficient.

PROBLEM FORMULATION

Imagine indexing the region's OD pairs by the single index $j = 1, \ldots, m$, and let d_j denote the demand for travel between OD Pair j. We can also imagine that counts are available from a total of p of the network's links. Define $\mathbf{y}$, a p-dimensional vector whose kth element y_k denotes the traffic count on the kth link containing a traffic counter, and $\mathbf{q}_j$, a p-dimensional vector whose kth element q_{jk} denotes the probability that a trip between OD Pair j uses Link k. The link use probabilities $\mathbf{q}_j$ are assumed to be constant, consistent with the assumption that the traffic assignment is in equilib-

G. A. Davis, Department of Civil and Mineral Engineering, University of Minnesota, 500 Pillsbury Drive SE, Minneapolis, Minn. 55455. N. L. Nihan, Department of Civil Engineering, FX-10, University of Washington, Seattle, Wash. 98195.

rium. Computation of these probabilities thus requires first solving an equilibrium assignment problem. Following Cascetta and Nguyen (*2*), the relation between the link counts **y** and the OD volumes d_j can then be given as a linear regression of the form

$$\mathbf{y} = \sum_{j=1}^{m} (d_j \mathbf{q}_j) + \mathbf{e} \tag{1}$$

where **e** denotes an error vector accounting for discrepancies between the link counts and their expected values.

If the rank of the matrix $[\mathbf{q}_1, \ldots, \mathbf{q}_m]$ is greater than or equal to m, the preceding regression problem is well posed and least-squares estimates of the OD parameters d_j can be readily computed. For networks of realistic size, however, the number of OD elements exceeds the number of links in the network, so that Equation 1 leaves the d_j underdetermined. A natural solution to this problem is to expand Equation 1 by collecting a time series of link counts $\mathbf{y}(1), \mathbf{y}(2), \ldots$ made at times $t = 1, 2, \ldots$ and replacing Equation 1 by an extended version:

$$\mathbf{y}(1) = \sum_{j=1}^{m} d_j \mathbf{q}_j + \mathbf{e}(1)$$

$$\mathbf{y}(2) = \sum_{j=1}^{m} d_j \mathbf{q}_j + \mathbf{e}(2) \tag{2}$$

However, for constant $\mathbf{q}_j$, it is easy to verify that the rank of the extended regression matrix will always equal the rank of $[\mathbf{q}_1, \ldots, \mathbf{q}_m]$, and the identifiability problem encountered above cannot be solved, even with an infinite sequence of observations. Thus in this formulation of the problem, link count data cannot provide a method of consistently estimating the OD parameters. This difficulty has been well recognized, and the focus of research on OD estimation has been on combining link count data with other data (such as from surveys) to provide usable procedures (*2–7*).

The situation for general networks contrasts with that for linear networks, such as single intersections, freeway sections, and transit routes. Here only one route connects each OD pair, and it is possible to obtain counts of the total traffic departing each origin (input counts) and arriving at each destination (output counts). Several papers have established that, given time-series observations of the input and output counts, least-squares–based estimators may be used to estimate the proportions of an origin's traffic that terminate at the various destinations, even when the number of OD parameters exceeds the number of count locations. This approach is originally due to Cremer and Keller (*8*), whereas a recent paper by Nihan and Davis (*9*) gives conditions and a proof for the strong consistency of ordinary least-squares in estimating intersection turning-movement proportions. Thus for linear networks, consistent estimators of OD parameters are readily constructed from least-squares algorithms, whereas for general networks, Equation 1 permits consistent estimation only if the number of counted links exceeds the number of OD pairs.

The source of this discrepancy lies in how the information available in a set of link counts is used. Equation 1 essentially states that the observed traffic counts are equal to the sum of a mean value and an error term, where the mean value vector is a linear function of the OD parameters. Even though the link count mean is consistently estimable from stationary observations, the consistency is not inherited by the estimates of the OD parameters because of the noninvertibility of the function relating the mean link flows to the OD parameters. In linear networks, on the other hand, the statistics of interest are not the mean values of the counts, but the covariance matrix between the input counts and the output counts. This matrix is also consistently estimable, and, given certain conditions on the process generating the input counts, an invertible relationship between the OD parameters and the covariance matrix elements can be established, permitting consistent estimation of the OD parameters. This suggests that if not only the mean values but also the covariance properties of link counts can be expressed as well-behaved functions of the OD parameters, consistent estimators of these parameters requiring only link count data can be constructed. In fact, it has been known for quite some time that the OD flows and the covariance of the link volumes have a well-defined and plausible connection (*10*), but it has been difficult to find an appropriate use for this knowledge. In large part this is because the process that generates traffic counts has nontrivial dynamics, so that a time series of link counts must be viewed as a realization of a stochastic process rather than as the result of random sampling. The temporal dependencies among the link counts often invalidate the use of classical statistical procedures, whereas attempts to develop dynamic, stochastic models of traffic assignment have either been restricted to simple networks (*11*) or have produced intractable models (*12*). However, recent work has investigated this problem in some detail (*13,14*) and established that under conditions similar to those needed to justify the use of stochastic user equilibrium (SUE) traffic assignment methods, a stochastic traffic generation model similar to that used by Cascetta (*12*) can be approximated by a stationary linear stochastic process driven by normally distributed noise as the number of travelers in the system becomes large. The parameters of this process are in turn well-defined functions of the OD parameters, and the approximation can be used to develop both ML and method of moments (MOM) approaches to OD parameter estimation. The approximation model is first presented in some detail, and the way the model is parameterized by the OD parameters is emphasized. OD estimation based on this model is then described.

A STOCHASTIC TRAFFIC MODEL

Before one can develop and validate statistical procedures, one must have an explicit model of the probabilistic mechanisms that are assumed to generate the available data. Equation 1 provides a model for traffic counts, but for the reasons described is not sufficiently rich for developing a useful statistical theory. Fortunately, Equation 1 has been used not so much because of its inherent validity, but because more realistic alternatives are lacking. Development of better alternatives requires more detailed consideration of how traffic counts are generated by the underlying trip generation and assignment processes, and these are poorly understood. One

could wait until more detailed knowledge is available (presumably from laboratory studies) and then use it to construct models of traffic generation for actual systems, much as well-established principles of mechanics are used in structural design. A more direct strategy would be to formulate plausible models based on available knowledge with the aim of eventually using real-world systems as one's laboratory, and this is the strategy followed here. Thus, rather than attempting to construct a comprehensive stochastic model of traffic generation, just enough probabilistic structure will be added to the standard assumptions concerning traffic generation so that a tractable stochastic process model of traffic counts can be derived. These assumptions lead to a simple but, from an estimation standpoint, intractable stochastic process model. For large traveling populations, however, it is possible to approximate this intractable model by a more tractable linear time-series model, from which OD estimators are readily constructed. The justification for this approximation uses standard mathematics but is long and technical. Because it is described elsewhere (*13,14*), it is not included here. However, familiarity with the structure of the resulting models helps one to see the straightforward way in which OD parameter estimators are constructed and how the properties of the estimators follow from the properties of the models.

We begin by treating the OD flows not as constants but as random variables, and in particular assume that $d_j(t)$, the flow between OD Pair j on Day t is a binomial random variable with parameters n_j and p_j. The outcomes for each OD pair and for each day are assumed to be independent, and the means and variances of the OD flows are thus given by

$$\bar{d}_j = n_j p_j$$

$$\sigma_j^2 = n_j p_j (1 - p_j) \tag{3}$$

The values of the OD parameters $\bar{d}_j$ and σ_j^2 are what we desire to estimate from link count observations. (If the OD flows are treated as constants rather than random variables, we simply deal with the special case where $\sigma_j^2 = 0, j = 1, \ldots, m$.) Now assume we have a total of n links in our network, and let

$\mathbf{x}(t)$ = the n-dimensional vector whose kth element $x_k(t)$ gives the traffic volume on Link k on Day t;
$c_k(\mathbf{x})$ = a differentiable function that gives the cost of traversing Link k as a function of the traffic volume vector $\mathbf{x}$;
$g_k(t)$ = the traveling population's anticipated cost of traversing Link k on Day t;
$p_{jr}[\mathbf{g}(t)]$ = differentiable functions giving the probability that a traveler between OD Pair j uses the rth route connecting j, as a function of the current anticipated cost vector $\mathbf{g}(t)$; and
δ_{jrk} = 1 if Link k lies on Route r connecting j and 0 otherwise.

Then the underlying traffic generation model, Model A, can be expressed in recursive form as follows:

0. Given initial anticipated costs $\mathbf{g}(0)$ and link volumes $\mathbf{x}(0)$, let $t = 1$.
1. Generate the $d_j(t)$ as binomial outcomes with parameters n_j and p_j, $j = 1, \ldots, m$.
2. Let $g_k(t) = (1 - \alpha)g_k(t - 1) + \alpha c_k[\mathbf{x}(t - 1)]$, $0 \leq \alpha \leq 1$.
3. Generate the route flows $d_{jr}(t)$ as multinomial outcomes with parameters $[d_j(t), p_{jr}(\mathbf{g}(t))]$.
4. Let $x_k(t) = \Sigma_j \delta_{jrk} d_{jr}(t)$, $k = 1, \ldots, n$.
5. Let $t = t + 1$ and go to Step 1.

The process described in Model A is a first-order Markov process. The recursion in Step 2 allows the anticipated cost to be a weighted average of the actual historical link costs, with α controlling the relative importance of recent costs. Adjustment mechanisms of this sort have appeared elsewhere (*11,12*), whereas the idea that route selection is a multinomial process appears to date back to Daganzo (*10*). Model A is easy to simulate, at least for small networks, but the convolutional nature of the link volumes expressed in Step 4 makes derivation of the probability distribution of the link volumes a difficult practical problem. Similar problems have been encountered in statistical mechanics, population biology, mathematical sociology, and so forth (*15,16*). They have often been successfully dealt with by using tractable approximations that become increasingly accurate as the size of the population increases. Intuitively, the approximation of Model A can be based on the fact that, conditional on what has happened on Day $t - 1$, the link volumes on Day t are the result of a large number of independent, individual route choice decisions, so that analogs of the Strong Law of Large Numbers and the Central Limit Theorem ought to apply. This intuition is given formal substance elsewhere (*13,14*). The result is that for large traveling populations, and in the vicinity of a stable user equilibrium assignment, the link count process generated by Model A can be approximated by a vector-valued autoregressive moving average process. In many cases the moving average component of this approximation can be neglected, giving the following first-order, vector autoregressive [VAR(1)] model, Model B:

$$\mathbf{x}(t) - \bar{\mathbf{x}} = \mathbf{F}[\mathbf{x}(t - 1) - \bar{\mathbf{x}}] + \mathbf{a}(t) \tag{4}$$

In Model B, $\mathbf{x}$ is a stochastic user equilibrium assignment satisfying

$$\bar{\mathbf{x}} = \sum_{j=1}^{m} \mathrm{d}_j \mathbf{q}_j[\mathbf{c}(\bar{\mathbf{x}})] \tag{5}$$

$\mathbf{F}$ is a weighted Jacobian matrix of the right hand side of Equation 5:

$$\mathbf{F} = \alpha \sum_{j=1}^{m} \bar{d}_j \left[\frac{\partial \mathbf{q}_j[\mathbf{c}(\mathbf{x})]}{\partial \mathbf{x}} \right]_{\bar{\mathbf{x}}} \tag{6}$$

The $\mathbf{a}(t)$ are independent, identically distributed normal random vectors with mean vector equal to $\mathbf{0}$ and covariance matrix $\mathbf{Q}$ given by

$$\mathbf{Q} = \sum_{j=1}^{m} (\bar{d}_j[\tilde{\mathbf{Q}}_j - \mathbf{q}_j \mathbf{q}_j^T] + \sigma_j^2 \mathbf{q}_j \mathbf{q}_j^T) \tag{7}$$

As defined earlier, the $\mathbf{q}_j$ appearing in Equation 7 denote the vectors of link use probabilities, whose elements are given by

$$q_{jk} = \sum_r (\delta_{jrk}) p_{jr}[\mathbf{c}(\bar{\mathbf{x}})] \quad (8)$$

The matrices $\mathbf{Q}_j$ appearing in Equation 7 are defined so that the klth element $Q_{j,kl}$ gives the probability that a trip between OD Pair j uses both Links k and l:

$$\bar{Q}_{j,kl} = \sum_r (\delta_{jrk})(\delta_{jrl}) p_{jr}[\mathbf{c}(\bar{\mathbf{x}})] \quad (9)$$

The advantages of replacing Model A with Model B derive from the fact that the VAR(1) process is arguably the most tractable and well understood of Markov processes. VAR(1) processes are parameterized by the quantities $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ and, for a set of observations $\mathbf{x}(t)$, $t = 1, \ldots, N$, the ML estimates of $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ are easily calculated using least-squares methods (*17*). We are not interested in all VAR(1) processes, however, but only those that can result from underlying traffic assignment processes, and thus must restrict our attention to VAR(1) processes whose parameters satisfy Equations 5 through 7. Because Equation 5 expresses the mean vector $\mathbf{x}$ as an implicit function of the OD flow means, the question arises as to whether the relationship between the VAR parameters $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ and the OD parameters $\bar{d}_j$ and σ_j^2 is well defined. If the VAR parameters are differentiable functions of the OD parameters, we can consider taking the derivative of the VAR likelihood function with respect to the OD parameters and solving for the OD parameter values making these derivatives equal to zero, producing ML estimates. If no such functions exist, the estimation task is much more difficult. Fortunately, it can be shown (*13*, chapter 4) that if (*a*) the link cost functions $c_k(.)$ and the route choice probability functions $p_{jr}(.)$ are continuously differentiable and (*b*) the rank of the matrix $\mathbf{I} - (1/\alpha)\mathbf{F}$ is n when the d_j are the true values of the OD parameters, then at least in a neighborhood of the true values of $\bar{d}_j$, σ_j^2, the VAR parameters are continuously differentiable functions of the OD parameters. Although an explicit representation of this function cannot be given, implicit differentiation can be used to obtain the necessary derivatives. It can also be readily verified that when the link cost functions have the BPR form, while the route choice probabilities are given by the multinomial logit formula, conditions (*a*) and (*b*) are satisfied (*13*). This well-behaved relationship between the OD parameters and the VAR parameters is the basis of OD estimation.

Essentially Model B claims that if one had time-series data consisting of, say, morning peak-period traffic volumes for all links in a network collected over several months, the data should be accurately modeled as a VAR(1) process. The mean of the data set should be equal to the SUE assignment and the regression and covariance matrices, respectively, should be equal to the $\mathbf{F}$ and $\mathbf{Q}$ given in Equations 6 and 7. If traffic counts are only available from a subset of the network links, the time-series model describing these partial counts will no longer necessarily be VAR(1) but will still be a stationary linear model whose parameters are at least in theory computable from $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ (*18*). Thus Model B makes some strong claims concerning the nature of traffic count data that are, at least in principle, testable from data collected by automatic surveillance and control systems. A limited amount of published empirical work suggests that Model B is not implausible. For instance, Cascetta (*12*) provides evidence that an SUE assignment is as reasonable a forecast of traffic volumes as the more standard DUE assignment, whereas in earlier work (*19,20*) we found that once seasonal trends have been accounted for, stationary linear models provide reasonable representations of freeway volume counts. However, more adequate empirical testing of traffic assignment methods requires accurate estimation of OD parameters, because otherwise one cannot distinguish between poor fit caused by a poor model and poor fit caused by ignorance of the actual OD patterns (*21,22*). By treating OD estimation as a problem in system identification, these aspects—model estimation and model validation—can be integrated.

ESTIMATION METHODS

Before turning to the construction of estimators for the OD parameters, it is necessary to resolve two technical issues. The first concerns the placement of the traffic counters generating the available data. In most practical cases, not all network links contain traffic counters, although the advent of automatic surveillance and control systems makes the availability of traffic counts more widespread. The estimation theory for VAR(1) processes, however, assumes that, subject to a limitation to be discussed shortly, observations from all links are available. This assumption allows the estimation theory for Markov processes developed by Billingsley (*23*) to be applied to this problem, resulting in relatively straightforward estimation methods whose asymptotic properties are readily established. On the other hand, when we assume that only a subset of the links have counters, it is well known from systems theory that the stochastic process describing the observations will no longer be Markov, even though the underlying process is. Computation of the likelihood function for this case requires employment of the Kalman filter, and establishing the asymptotic properties of the resulting estimators is an unsolved problem. Thus in this paper we concentrate on the case where a full set of traffic counts is available, saving the partial count case for later research.

Having made the case for a full set of traffic counts, we next note that the link flows on a network contain linear dependencies because of conservation of flow requirements. This means that the link flow vector $\mathbf{x}$ can be partitioned into $\mathbf{x}_1$ and $\mathbf{x}_2$, with a linear relationship $\mathbf{x}_2 = B\mathbf{x}_1$ existing between $\mathbf{x}_1$ and $\mathbf{x}_2$. The linear dependence causes the covariance matrix $\mathbf{Q}$ to be singular, which in turn causes both practical and theoretical problems in OD estimation (*13*). This is easily solved by working with the independent set of link counts $\mathbf{x}_1$ and deleting the appropriate rows and columns from $\mathbf{Q}$. Computation of the corresponding matrix $\mathbf{F}$ is readily done using the chain rule for vector-values functions. Thus from here on it is assumed that we have available a full set of linearly independent counts, and the quantities $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ are computed for this linearly independent set.

With these cautions noted, let

$$\hat{\mathbf{x}}(t) = \bar{\mathbf{x}} + \mathbf{F}[\mathbf{x}(t-1) - \bar{\mathbf{x}}] \quad (10)$$

denote the predicted value for $\mathbf{x}(t)$. ML estimates for the model parameters are then obtained by minimizing the scaled log-likelihood function

$$L = \log|\mathbf{Q}| + (1/N) \sum_{t=1}^{N} [\mathbf{x}(t) - \hat{\mathbf{x}}(t)]^T \mathbf{Q}^{-1} [\mathbf{x}(t) - \hat{\mathbf{x}}(t)] \quad (11)$$

with respect to the parameters of interest. For the VAR parameters $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$, the ML estimates have a closed-form solution, which can be readily computed using standard regression methods. It is also well known that these estimates are consistent, asymptotically normally distributed, and asymptotically efficient (*17,23*). For the OD parameters $\bar{d}_j$ and σ_j^2, one can still take derivatives of Equation 11 using the chain rule and implicit differentiation, set the derivatives equal to zero, and then in principle solve the resulting equations for the ML estimates. The likelihood equations will in this case not admit a closed-form solution, so that numerical solution methods must be employed. However, given some technical conditions, it can still be established that Billingsley's results on ML estimation for Markov chains apply to this situation and that the solutions to the likelihood equations are consistent, asymptotically normally distributed, and asymptotically efficient (*13*). Thus the classical results concerning ML estimators apply to the ML estimates of the OD parameters, though with some numerical difficulties in actually computing these estimates.

Alternatively, Equations 5 through 7 defining the VAR parameters can be written in a vectorized form:

$$\bar{\mathbf{x}} = \sum_{j=1}^{m} \bar{d}_j \mathbf{q}_j$$

$$v(\mathbf{F}) = \alpha \sum_{j=1}^{m} \bar{d}_j v\left(\left[\frac{\partial \mathbf{q}_j}{\partial \mathbf{x}}\right]\right)$$

$$vh(\mathbf{Q}) = \sum_{j=1}^{m} [\bar{d}_j vh(\bar{\mathbf{Q}}_j - \mathbf{q}_j\mathbf{q}_j^T) + \sigma_j^2 vh(\mathbf{q}_j\mathbf{q}_j^T)] \quad (12)$$

where $v(.)$ denotes the vector operator (the stacking of the columns of a matrix on top of each other) and $vh(.)$ denotes the vector half-operator (the stacking of the columns of the lower diagonal of a symmetric matrix on top of each other). Given estimates of $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ and knowledge of the parameter α, Equation 12 defines a system of linear equations with the OD parameters as the unknowns. When the number of individual paramaters in $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ exceeds the number of OD parameters, these equations can be solved using the Moore-Penrose pseudoinverse (or equivalently, ordinary least-squares) to obtain estimates of the OD parameters. Because the ML estimates for the VAR parameters are also the MOM estimators, the second approach gives a MOM estimator of the OD parameters. Under the conditions that the link cost and route choice functions are continuously differentiable, the function relating the MOM OD estimates to the MOM VAR estimates will also be continuously differentiable, so that the consistency and asymptotic normality shown by the VAR estimators will be inherited by the MOM OD estimators (*24*, p. 388). The asymptotic efficiency of the VAR estimators, however, is not guaranteed to transfer to the MOM OD estimators.

MONTE CARLO EVALUATION OF ESTIMATORS

Both the ML and the MOM estimators assume the VAR(1) Model B is an accurate description of the process generating the link volumes. Model B is in turn derived as a large-population approximation to the Markov process described in Model A. With infinite populations and infinite amounts of data, the asymptotic theory characterizes the properties of these estimators. However, for a given finite population and finite data sequences, the asymptotic properties may not be operative, and simulation studies are required to verify the practical usefulness of the asymptotic results. We have already described some asymptotic properties of the ML and MOM estimators based on large-population and large-sample approximations. To gain some appreciation of how the ML and MOM estimators would perform on finite data sets generated by a Model A process, we conducted the following simulation study. A FORTRAN program, SIMFLO, was written that could simulate the Model A process for networks small enough that enumeration of the network routes was feasible. (This was necessary so that the multinomial simulation described in Step 3 of Model A could be performed.) SIMFLO assumes that the link costs are given by the BPR functions with the form

$$c_k(x_k) = a_k\left[1 + b_k\left(\frac{x_k}{f_k}\right)^4\right] \quad (13)$$

The route choice probabilities are given by the logit formula

$$p_{jr} = \frac{e^{-\Theta c_{jr}}}{\sum_s e^{-\Theta c_{js}}} \quad (14)$$

Two FORTRAN estimation programs were also written. Program MARKOD computes numerical ML estimates of the OD parameters using the quasi-Newton routine E04JBF, obtained from the NAG subroutine library (*25*). Program MOMOD computes MOM estimates using standard least-squares procedures. Although we conducted tests with several hypothetical networks, space limitations permit the description of only one simulation experiment. However, the results of all experiments were consistent with the results that we now present. Figure 1 shows a 14-link network with three origin nodes (O1, O2, and O3) and three destination nodes (D1, D2, and D3), for a total of nine OD pairs. If one imagines collecting weekday peak-period volume data for a network, a time series of length 150 would correspond to 7 to 8 months of observations, and this was considered a reasonable upper bound on the duration over which OD patterns might be taken as being constant. From such a time series, one could obtain one set of estimated OD parameters, so to sample the statistical properties of the OD estimators, 50 separate time series, each of length 150, were generated using SIMFLO. The 50 time-series were then input to MARKOD and MOMOD to obtain a sample, of size 50, of the ML and MOM estimates. Although in principle it is possible to treat the weighting parameter α and the logit parameter Θ as unknown (MARKOD permits this option), the relationship between $\bar{\mathbf{x}}$, $\mathbf{F}$, and $\mathbf{Q}$ and the unknown parameters shown in Equation 12 is no longer linear, and computation of the MOM estimators will be more difficult. Because the primary goal of this study is assessment of the estimates of the OD parameters $\bar{d}_j$ and σ_j^2, α and Θ will be treated as known a priori. Estimation using MARKOD was done on the Cyber 180/855 computer at the University of Washington, and estimation using

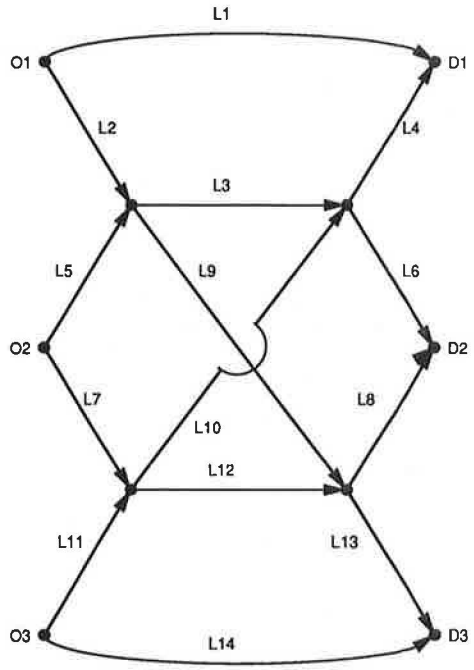

FIGURE 1 Diagram of 14-link network used in simulation study.

MOMOD was done on the Cyber 170/845 computer at the University of Minnesota.

Table 1 gives statistical information concerning the properties of the ML estimator. The column labeled "True" gives each OD parameter's true value; the column labeled "Mean" gives the average, across the 50 simulations, of the estimated values produced by the ML method; and the column labeled "S.D." gives the standard deviation of these estimates about their average. The column labeled "t" gives a computed *t*-statistic testing the hypothesis that the average estimate equals its corresponding true value, which is a test for bias in the estimator. The column labeled "r" gives the normal-score correlation test for normality (*26*), which tests the asymptotic normality of the estimates. Table 2 gives similar information for the MOM estimator. One asterisk indicates statistical significance of a test at the .05 level, whereas two asterisks indicate statistical significance at the .01 level.

From the "r" columns, it can be seen that both the ML and the MOM methods produce estimates that tend to be normally distributed, indicating that a data series of length 150 appears sufficient to obtain asymptotic normality. Looking at the "t" columns, we see an undeniable tendency for both estimators to generate biased estimates, indicating that with data sequences of length 150 there is still some gap, on the average, between the estimate and the true value. Inspection of the "True" and "Mean" columns in both tables indicates that for the ML estimator the biases tend to be less than 10 percent of the true values, whereas for MOM the biases are somewhat worse. A comparison of the "S.D." columns in Tables 1 and 2 indicates that the ML estimates have markedly less variance than the MOM estimates, a result consistent with the asymptotic efficiency property of the ML estimator. The variability of the MOM estimates of the OD variances is in fact so large that it calls into question the practical value of such estimates.

TABLE 1 ESTIMATION SUMMARY FOR ML METHOD

Parameter	True	Mean	S.D.	t	r
d_1	450.	449.2	2.1	2.76**	.991
d_2	300.	300.1	23.5	0.03	.985
d_3	301.	301.	23.9	0.00	.989
d_4	352.	362.3	25.4	-2.86**	.996
d_5	490.	471.7	8.4	15.5**	.995
d_6	150.	157.8	24.5	-2.23*	.988
d_7	100.	90.9	24.9	2.60*	.994
d_8	200.	217.1	25.1	-4.89**	.993
d_9	700.	691.9	3.3	17.23**	.978
$\sigma_{1_2}^2$	45.	47.0	10.6	-1.33	.994
$\sigma_{2_2}^2$	75.	75.4	12.8	-0.22	.994
$\sigma_{3_2}^2$	42.1	44.3	9.7	-1.59	.992
$\sigma_{4_2}^2$	42.2	40.1	9.9	1.07	.989
$\sigma_{5_2}^2$	147.	141.3	22.4	1.78	.991
$\sigma_{6_2}^2$	37.5	38.2	10.9	-0.47	.994
$\sigma_{7_2}^2$	50.	49.6	10.8	0.27	.990
$\sigma_{8_2}^2$	40.	39.6	12.0	0.24	.973*
σ_9^2	210.	202.0	27.5	2.05*	.988

CPU seconds/estimation = 2504
(Cyber 180/855)

TABLE 2 ESTIMATION SUMMARY FOR MOM

Parameter	True	Mean	S.D.	t	r
d_1	450.	449.99	25.87	0.003	.987
d_2	300.	286.07	54.98	1.790	.985
d_3	301.	315.54	45.15	-2.270*	.973
d_4	352.	341.97	61.27	1.158	.987
d_5	490.	504.80	99.10	-1.057	.996
d_6	150.	128.00	70.58	2.204*	.986
d_7	100.	104.95	45.61	-0.760	.990
d_8	200.	189.44	63.75	1.170	.993
d_9	700.	716.89	60.03	-1.989	.994
$\sigma_{1_2}^2$	45.	28.31	34.40	3.427**	.993
$\sigma_{2_2}^2$	75.	67.63	32.29	1.612	.986
$\sigma_{3_2}^2$	42.1	34.41	23.83	2.280*	.988
$\sigma_{4_2}^2$	42.2	30.80	45.22	1.781	.988
$\sigma_{5_2}^2$	147.	126.23	31.08	4.720**	.990
$\sigma_{6_2}^2$	37.5	34.47	45.79	0.467	.993
$\sigma_{7_2}^2$	50.	45.38	23.02	1.417	.995
$\sigma_{8_2}^2$	40.	39.77	26.88	0.060	.984
σ_9^2	210.	187.56	52.80	3.000**	.988

CPU seconds/estimation = 0.89
(Cyber 170/845)

Finally, at the bottom of Tables 1 and 2 we display the average CPU time required for the two estimation methods. Whereas the MOM method is very efficient from a numerical standpoint, the ML procedure proved to be computationally burdensome, suggesting that ML estimation for networks of realistic size may prove difficult even on a supercomputer.

SUMMARY AND CONCLUSION

We have considered here the problem of estimating the parameters describing the OD demand on a traffic network from time-series observations of the network's link volumes. The existence of such methods, coupled with the availability of automatic traffic-counting technology, holds the promise of being able to obtain timely, inexpensive estimates of existing travel demand and of being able to detect and track changes in demand over time or in response to transportation initiatives. Because the dynamics of the processes generating traffic cannot be neglected, we argue that OD estimation is more properly seen as statistical inference on stochastic processes rather than an example of classical statistical procedures. After developing a tractable stochastic process model that is a plausible approximation of the traffic count process, we use the

model to develop both ML and MOM estimators of the process's OD parameters. Both estimators have desirable consistency and asymptotic normality properties, but only the ML estimator can claim asymptotic statistical efficiency. Simulation studies indicate that the ML estimator is superior from a statistical standpoint, but the numerical labor needed to compute ML estimates makes application to networks of realistic size problematical. On the other hand, the MOM estimates, though clearly inferior statistically, are numerically efficient enough to be considered for networks of realistic size. At this point, the trade-off between statistical and numerical efficiency is unforgiving. Thus, although it is possible to begin developing a statistical theory for making inferences about OD parameters from traffic count data, numerical issues stand between the theory and its general usefulness.

The obvious research need then is to develop methods that preserve the statistical efficiency of the ML method but reduce its computational burden. The minimization procedure implemented in Program MARKOD can be viewed as a variant of the reduced-gradient algorithm, and it has been recently reported that reduced-gradient algorithms tend to be numerically slower than competitors such as sequential quadratic programming (SQP) (*27*). A promising line of investigation would be to reformulate the ML estimation problem as an SQP and use a state-of-the-art SQP code to solve it. We are currently pursuing this course. Another need is to develop methods that do not require a full set of traffic counts. Numerical computation of ML estimators for this case is relatively straightforward (although some limited numerical experiments indicate that the CPU demands tend to be greater than those for the full-count case). What is problematic is demonstrating that these estimators have desirable asymptotic properties, such as consistency and asymptotic normality. Because the numerical difficulties shown by the full-count case also tend to be shown by the partial count case, the main obstacle to practical use is the lack of a numerically efficient computation method.

ACKNOWLEDGMENTS

This research was supported in part by the Valle Foundation through a Scandinavian Study Scholarship at the Royal Institute of Technology, Stockholm, Sweden awarded to Gary Davis and in part by the National Science Foundation through a grant awarded to Nancy Nihan. The authors would like to thank Mr. Konstantinos Koutsoukos for programming and running some of the simulation results.

REFERENCES

1. S. Atkins. Transportation Planning Models—What the Papers Say. *Traffic Engineering and Control*, Vol. 27, 1986, pp. 460–467.
2. E. Cascetta and S. Nguyen. A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research*, Vol. 22B, 1988, pp. 437–455.
3. M. Maher. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian Statistical Approach. *Transportation Research*, Vol. 17B, 1983, pp. 435–447.
4. M. Bell. The Estimation of an Origin-Destination Matrix from Traffic Counts. *Transportation Science*, Vol. 17, 1983, pp. 198–217.
5. H. Spiess. A Maximum Likelihood Model for Estimating Origin-Destination Matrices. *Transportation Research*, Vol. 21B, 1987, pp. 395–412.
6. S. McNeil and C. Hendrickson. A Regression Formulation of the Matrix Estimation Problem. *Transportation Science*, Vol. 19, 1985, pp. 278–292.
7. E. Cascetta. Estimation of Trip Matrices from Traffic Counts and Survey Data: A Generalized Least-Squares Estimator. *Transportation Research*, Vol. 18B, 1984, pp. 289–299.
8. M. Cremer and H. Keller. Dynamic Identification of Flows from Traffic Counts at Complex Intersections. In *Eighth International Symposium on Transportation and Traffic Theory* (V. F. Hurdle et al., eds.), University of Toronto Press, Toronto, 1983, pp. 121–142.
9. N. L. Nihan and G. A. Davis. Application of Prediction-Error Minimization and Maximum Likelihood To Estimate Intersection O-D Matrices from Traffic Counts. *Transportation Science*, Vol. 23, 1989, pp. 77–90.
10. C. F. Daganzo. Some Statistical Problems in Connection with Traffic Assignment. *Transportation Research*, Vol. 11, 1977, pp. 385–389.
11. J. L. Horowitz. The Stability of Stochastic Equilibria in a Two-Link Transportation Network. *Transportation Research*, Vol. 18B, 1984, pp. 13–28.
12. E. Cascetta. A Stochastic Process Approach to the Analysis of Temporal Dynamics in Transportation Networks. *Transportation Research*, Vol. 23B, 1989, pp. 1–17.
13. G. A. Davis. *A Stochastic, Dynamic Model of Traffic Generation and its Application to the Maximum Likelihood Estimation of Origin-Destination Parameters*. Doctoral dissertation. University of Washington, Seattle, 1989.
14. G. A. Davis and N. L. Nihan. A Dynamic, Stochastic Approach to Traffic Assignment. Submitted to *Operations Research*, 1990.
15. Karmeshu and R. K. Pathria. A Stochastic Model for Highway Traffic. *Transportation Research*, Vol. 15B, 1981, pp. 285–294.
16. T. G. Kurtz. Strong Approximation Theorems for Density Dependent Markov Chains. *Stochastic Processes and Their Applications*, Vol. 6, 1978, pp. 223–240.
17. T. W. Anderson. *The Statistical Analysis of Time Series*. John Wiley and Sons, New York, 1971.
18. P. E. Caines. *Linear Stochastic Systems*. John Wiley and Sons, New York, 1988.
19. N. L. Nihan and G. A. Davis. Estimating the Impacts of Ramp-Control Programs. In *Transportation Research Record 957*, TRB, National Research Council, Washington, D.C., 1984, pp. 31–32.
20. G. A. Davis and N. L. Nihan. Using Time-Series Designs to Estimate Changes in Freeway Level of Service, Despite Missing Data. *Transportation Research*, Vol. 18A, 1984, pp. 431–438.
21. M. Florian. An Introduction to Network Models Used in Transportation Planning. In *Transportation Planning Models* (M. Florian, ed.), Elsevier, Amsterdam, 1984, pp. 137–152.
22. Y. J. Stephanedes, I. Z. Argiopoulos, and P. Michalopoulos. On-line Traffic Assignment for Optimal Freeway Corridor Control. *ASCE Journal of Transportation Engineering*, Vol. 116, 1990, pp. 744–755.
23. P. Billingsley. *Statistical Inference for Markov Processes*. University of Chicago Press, Chicago, 1961.
24. C. R. Rao. *Linear Statistical Inference and Its Applications*, 2nd ed. John Wiley and Sons, New York, 1973.
25. *NAG Fortran Library Manual*, Mark 9, NAG (USA), 1976.
26. J. J. Filliben. The Probability Plot Correlation Coefficient Test for Normality. *Technometrics*, Vol. 17, 1975, p. 111.
27. P. Gill, W. Murray, M. Saunders, and M. Wright. Constrained Nonlinear Programming. In *Optimization*, Vol. 1 (G. L. Nemhauser et al., eds.), Elsevier Science Publishers, Amsterdam, 1989.

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

Route-Specific and Time-of-Day Demand Elasticities

YUPO CHAN

In assessing user response to cost and service changes, demand elasticities are useful tools. Current compilations of demand elasticities, however, are not helpful to scheduled transportation operators. The reason is that the range of an elasticity is too wide, and there is no practical guideline for picking an appropriate value within the range. Furthermore, they are often compiled for systemwide and average-day operations, whereas most analyses need to be performed on a route-by-route basis during peak or off-peak periods. A methodology to address this problem is presented with the objective of providing demand elasticities that are practical for patronage analyses in an operating agency. Through statistical analyses of spatial and temporal data aggregation, the methodology explains the differences among elasticity tabulations, and in so doing, provides insights into the variations among elasticity values. The results of this research include (*a*) guidelines for selecting an appropriate value of elasticity among the broad range of values (instead of simply taking the average or midpoint of the range) and (*b*) a method for converting the most commonly available elasticities (which are usually in aggregate form) to a more useful form, such as route-specific and time-of-day elasticities. The results have been demonstrated in a case study of the transit system in York, Pennsylvania.

A common concern expressed by scheduled-transportation operators is how a system can become more attuned to user preferences by innovation in service changes (*1*). Clearly, addressing such concerns requires knowledge of travelers' (shippers') response to changes in such items as user charges, schedules, and route coverage. For example, would a user still patronize the system if the fare is raised or the frequency of service is cut back (*2*)? Users' decisions clearly affect the revenue or well-being of the operator, which in turn leads to either success or failure of the scheduled transportation system.

Increasingly, systems are scrutinized on a route-by-route basis (*3*). Schedules and user charges are becoming more and more differentiated between peak and off-peak periods because user responses are widely different among routes and time of day. For example, peak-hour travel is typically inelastic inasmuch as it is made up largely of work trips on routes from suburb to center city. This contrasts with off-peak travel, which typically consists of discretionary trips. Unless distinctions are made between routes and time of day, fare and schedule changes cannot be planned judiciously to cater to the user's preferences (*4*).

A common way to address this issue is to examine demand elasticities, which by definition measure patronage responses to changes in attributes such as fares and service. Although several tabulations of urban demand elasticities exist (*5*), the use of these tabulations has been limited because of the tremendous variations in reported elasticity values—even for the same city or scenario (*6*). For example, within a given commuter rail authority, the time elasticity can vary from -0.31 to -0.87—a threefold difference. This is mainly due to the different levels of aggregation used in the derivation of these elasticities. Whereas there is tremendous need, operators rarely have usable guidelines available for selecting an appropriate elasticity for day-to-day operations—from route-level to peak versus off-peak analyses. There is no reason to believe that the midpoint (between -0.31 and -0.87) is the number to use (*5*) or that a uniform scale factor can be applied indiscriminately (between -0.31 and -0.87) to arrive at the appropriate value for a specific route or time of day.

Thus there is a knowledge gap to be filled in establishing probable variations in elasticities among routes and peak versus off-peak hours. The provision of these guidelines would allow operators to fine-tune elasticities by route and time of day on the basis of associated activity-system and level-of-service characteristics. Thus user responses to service changes can be accurately estimated for operational planning purposes.

LITERATURE REVIEW

Among the various approaches to estimating demand on a route-specific level, the elasticity method performs extremely well (*7,8*). In fact it can be argued that as a general method, it is most satisfying. Recent work on chain pivot-point analysis is merely a variation of the elasticity method (*9*). The same can be said for areawide-equilibration approaches (*10*).

Various fare-policy evaluation techniques have been built around demand elasticities. Ballou and Mohan (*11*) gave an example of the use of elasticities to analyze differentiated fare change on selected routes. Their analyses indicate the diverse effects of fare changes among routes and according to time of day, trip purpose, trip duration, and schedule frequency. Because demand elasticities are the basis of such fare-policy evaluations, the need for a reliable set of elasticities becomes obvious. To attain this accuracy, we need elasticities disaggregated by route, time of day, service level, and the socioeconomic composition of the potential riders (*12,13*). This in turn requires knowledge about how elasticities are derived, which is the only way to arrive at such a disaggregation procedure.

Calibrated elasticities are of two types, depending on the methodology and data used for their computation. Disaggre-

Department of Operational Sciences, School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, Ohio 45433-6583.

gate elasticities are estimated using detailed information (such as household data), whereas aggregate elasticities are based on coarser data (such as zonal averages). Aggregate elasticities implicitly assume that travel behavior is homogeneous within a geographic subdivision, such as a zone or a route. Use of aggregate elasticities therefore ignores the high levels of variation in travel behavior (particularly modal choice) that may exist among users of a route. The use of such aggregate data fails to extract much of the variational information from the constituent observations used to obtain zonal aggregates. In contrast, disaggregate elasticities, which use households as the unit of analysis, are able to tap an extremely rich source of variation.

In other words, one potential problem of aggregate-level analysis is the risk of ecological fallacy, in which aggregate-level correlations are mistakenly attributed to individuals. Another problem is the loss of variability in the data used for estimation. Because a model's coefficients are determined by explaining variations in observed travel behavior, the less the variation to be explained, the less reliable the model will be. The reduced variability in aggregate data also results in a high level of collinearity between variables at the aggregate level, which does not exist at the disaggregate level. These observation errors tend to impart downward bias to estimated coefficients. For example, the results of a study (*14*) show that by using disaggregate demand models, disaggregate elasticities—for a given level of service—are usually lower than the corresponding aggregate elasticities in algebraic value. Thus, a route-specific fare elasticity may be -0.4, whereas a systemwide figure may be -0.3.

THEORY

To be able to use the tabulated elasticities meaningfully, the relationship between aggregate elasticities and disaggregate elasticities needs to be reviewed as a first step. The most disaggregate level of demand elasticities is at the household level. These elasticities are usually derived from logit models of the following form (*15*):

$$p(m, t) = \frac{\exp(R_m X_{mt})}{\sum_{j=1}^{M} \exp(R_j X_{jt})} \qquad m = 1, 2, \ldots, M \tag{1}$$

where

$p(m, t)$ = probability of Mode m being taken by Household t ($t = 1, 2, \ldots, T$),
X_{mt} = vector of level-of-service and activity-system characteristics of Mode m for Household t, and
R_m = estimated vector of coefficients for the level-of-service and activity-system variables of Mode m.

From this model form, elasticities can be derived as

$$\eta(m, x_m, t) = [1 - p(m, t)]r_x x_{mt} \tag{2}$$

where

$\eta(m, x_m, t)$ = elasticity for Mode m with respect to attribute x_m for Household t given a destination and given that a trip will be made;
r_x = estimated coefficient for level-of-service variable x_m; and
x_{mt} = level-of-service attribute of Mode m, by Household t for a given destination.

There are mathematical relationships between disaggregate and aggregate elasticities. Recall that the elasticity for Household t, $\eta(m, x_m, t)$, is given in Equation 2. The demand elasticity aggregated over T households in a geographic subdivision such as a route is correspondingly

$$\eta(m, x_m) = \frac{\sum_{t=1}^{T} \eta(m, x_m, t)}{\sum_{t=1}^{T} p(m, t)} \tag{3}$$

Equation 3 can also be derived in a different way (*16*). Define $\bar{p}_m$ as the average market share of Mode m for a subdivision with a population of T households:

$$\bar{p}_m = \frac{1}{T} \sum_{t=1}^{T} p(m, t) \tag{4}$$

We proceed to differentiate $\bar{p}_m$ with respect to x_m:

$$\frac{\partial \bar{p}_m}{\partial x_m} = \frac{\partial}{\partial x_m} \left[\sum_{t=1}^{T} p(m, t)/T \right]$$

$$= \frac{1}{T} \sum_{t=1}^{T} p(m, t)[1 - p(m, t)]r_x \tag{5}$$

The aggregate elasticity then becomes

$$\eta(m, x_m) = \frac{\partial \bar{p}_m}{\partial x_m} \cdot \frac{x_m}{\bar{p}_m}$$

$$= \left\{ \frac{1}{T} \sum_{t=1}^{T} p(m, t)[1 - p(m, t)] \right\} r_x \frac{x_m}{\bar{p}_m} \tag{6}$$

By substituting $\bar{p}_m(1 - \bar{p}_m)$ for the terms in brackets { }, the following is obtained:

$$\bar{\eta}(m, x_m) = \frac{\bar{p}_m(1 - \bar{p}_m)}{\bar{p}_m} r_x x_m$$

$$= (1 - \bar{p}_m) r_x x_m \tag{7}$$

This is analogous to the disaggregate elasticity Equation 2.

The substitution above, however, assumes that

$$\bar{p}_m^2 = \frac{1}{T} \sum_{t=1}^{T} p(m, t)^2 \tag{8}$$

This assumption, when interpreted together with the definition of Equation 4, implies that individual mode-choice behavior is homogeneous within the subdivision. Because the individual travel decisions are not likely to be homogeneous in actuality, Equation 7 can only be an approximation of Equation 6.

Systemwide transit elasticities—being another level of aggregation—can also be derived from disaggregate elasticities calibrated for households (*16*). Let us write T as N_k to spe-

cifically denote the number of households on the kth route and to show that there is more than one route systemwide. Now let $\bar{p}(m, k)$ be the average probability of choosing Mode m in the kth route (where $k = 1, 2, 3, \ldots, K$). By definition,

$$N = \sum_{k=1}^{K} N_k \tag{9}$$

For the study area, using the new notation, and with the substitution

$$N_k\bar{p}(m, k)[1 - \bar{p}(m, k)] \leftarrow$$

$$\sum_{t=1}^{N_k} p(m, k, t)[1 - p(m, k, t)] \tag{10}$$

systemwide elasticity can be approximated as

$$\eta(m, x_m, k) = \frac{\partial\bar{p}(m,k)}{\partial x_{mk}} \cdot \frac{x_{mk}}{\bar{p}(m,k)} = \left\{\sum_{k=1}^{K} N_k\bar{p}(m,k)[1 - \bar{p}(m,k)]r_x\right\} \frac{x_{mk}}{\sum_{k=1}^{K} p(m,k)} \tag{11}$$

This relationship is obtained on the basis of the assumption of homogeneous household behavior within a route. The variability in the data is among routes (rather than households) in this case.

IMPLICATIONS

Now that the relationship between household, route, and systemwide elasticities has been derived, their applications in the field can be discussed. Specifically, the numerical values of these elasticities when they are applied in scheduled transportation analysis can be compared. In a study consisting of K routes, for example, the difference between the approximation of Equations 7 and 11 and the theoretical elasticity (Equations 6 and its route equivalent shown by substitution Equation 10) can be computed. The resulting difference between an aggregate and disaggregate elasticity on Route k is

$$\begin{aligned}\eta(m,x_m,k) - \eta(m,x_m,t) &= \frac{r_x x_{mk}}{N\bar{p}_m}\left\{N_k\bar{p}(m,k)[1 - \bar{p}(m,k)] - \sum_{t=1}^{N_k} p(m,k,t)[1 - p(m,k,t)]\right\} \\ &= \frac{r_x x_{mk}}{N\bar{p}_m}\left[\sum_{t=1}^{N_k} p(m,k,t)^2 - N_k\bar{p}(m,k)^2\right] \\ &= \frac{r_x x_{mk}}{N\bar{p}_m} N_k\sigma_k^2 \end{aligned} \tag{12}$$

where σ_k^2 is the variance of $p(m, k, t)$ in Route k. The difference between an aggregate elasticity and a disaggregate elasticity for the kth route is therefore

$$\frac{r_x x_{mk}}{N\bar{p}_m} N_k\sigma_k^2 \tag{13}$$

If $\hat{\sigma}^2$ is defined as the maximum σ_k over Route $k = 1, 2, 3, \ldots, K$ (hence $\sigma_k^2 \leq \hat{\sigma}^2$), the inaccuracy of an elasticity estimated for a route is bounded by the inequality

$$\sum_{k=1}^{K} \frac{N_k}{N\bar{p}_m} \sigma_k^2 r_x x_{mk} \leq \frac{\hat{\sigma}^2}{\bar{p}_m} r_x x_{mk} \tag{14}$$

Because σ^2 is the squared deviation of $p(m, t)$'s from $\bar{p}_m$, in general, and both $p(m, t)$ and $\bar{p}_m$ are between 0 and 1, it appears that σ should be very small. However, in a case study using a 1955 data sample from Chicago, Warner (*16*) found that σ^2 was about 0.04 (or $\sigma = 0.2$).

What can be said about the estimation error associated with an elasticity? Let us compare the estimated value using Equation 7 and the following value, which has been corrected by such error terms as Equations 13 and 14:

$$\tilde{\eta}(m, x_m) = \left[(1 - \bar{p}_m) - \frac{\sigma^2}{\bar{p}_m}\right] r_x x_m \tag{15}$$

By using values of $\bar{p}_m = .824$ and $\sigma^2 = 0.04$ from Warner's experiment in Chicago, the following is obtained from Equation 7:

$$\bar{\eta}(m, x_m) = 0.176 r_x x_m \tag{16}$$

and

$$\tilde{\eta}(m, x_m) = 0.127 r_x x_m \tag{17}$$

according to Equation 15. The ratio between the two values is

$$\frac{\text{direct estimation}}{\text{corrected estimation}} = \frac{0.176 r_x x_m}{0.127 r_x x_m} = 1.386 \tag{18}$$

This indicates that a route elasticity, for example, could be overestimated by as much as 40 percent by an aggregate figure (which assumes homogeneous travel behavior among all users in the route).

Similarly, any individual route elasticity can be overestimated by the ratio

$$\frac{1 - \bar{p}_m - \sigma_k^2/\bar{p}_m}{1 - \bar{p}_m - \hat{\sigma}^2/\bar{p}_m} \geq 1 \tag{19}$$

even after corrections of the estimated values are made according to Equation 15.

SITE TESTING

The theory described in the earlier section was applied to a transit line in a medium-sized city: York, Pennsylvania. The bus line, called the W. Market/E. Market line, runs between York's central city and two suburbs—one to the west and the other to the east. The line covers these zones en route: 15 →

$14 \rightarrow 5 \rightarrow 1 \rightarrow 2 \rightarrow 10 \rightarrow 32$, with 15 denoting the western suburb and 32 the eastern suburb.

We can obtain the travel times for our model from the published time schedule of the transit line. The waiting time (usually approximated by half of the headway) has to be added to the line-haul times to obtain the user's actual travel time. The activity-system and level-of-service data were collected for calibrating a logit model. The model specification includes the following explanatory variables for the zones the transit line goes through:

- Interzonal travel times by automobile at posted speed;
- Transit travel times, including wait time and walk time (at 5 ft/sec from zone centroid to bus station);
- Automobile travel costs expressed in perceived costs (at 5 cents per mile); and
- Transit travel costs at a uniform fare of 65 cents per trip.

The data are summarized in Table 1.

A logit model was calibrated by Chan (*17*) for the city of York, yielding a disutility function (or impedance function) of

$$R_m X_{mt} = -290.2c_{mt} - 70.5\tau_{mt} + 1.51 \tag{20}$$

Only level-of-service variables such as cost (c) and time (τ) are explicitly modeled; activity-system variables were all lumped under the calibration constant of 1.51, and there are only two modes (m = automobile or transit).

From logit Equations 1, a table can be generated consisting of the probability of a household t taking bus transit B, $p(B, t)$, among the seven zones ($i = 1, 2, \ldots, 7$; $j = 1, 2, \ldots, 7$) covered in the transit line. From the 42 (or 7×6) nonzero values in the table, we can obtain the average market share of the bus mode for Route k, $\bar{p}_B$, as

$$\frac{1}{42}\sum_{t=1}^{42} p(B, t) = 0.0064 \tag{21}$$

and

$$\sigma_k^2 = 0.0007 \tag{22}$$

Using Equation 7 for Route k, we have

$$\bar{\eta}(m, x_m) = (1 - .0064)r_x x_m = 0.994 r_x x_m \tag{23}$$

Similarly, through Equation 15, we have

$$\tilde{\eta}(m, x_m) = \left[(1 - 0.0064) - \frac{0.0007}{0.0064}\right] r_x x_m = 0.8842 r_x x_m \tag{24}$$

The ratio of the two values is

$$\frac{\bar{\eta}}{\tilde{\eta}} = 1.124 \tag{25}$$

which means that the route elasticities can be overestimated by as much as 12 percent. Notice that we have not ascertained the other routes, and hence $\hat{\sigma}$, the maximum over σ_k's, is not known. For this reason, it is possible to be off by significantly more than 12 percent in certain routes.

Even though this is a site testing of limited scope, we have obtained a few useful observations. It is clear that care should be exercised in the use of elasticities in route-specific transit analysis in view of the Chicago and York experiences. If any aggregate elasticity is used, adjustments should be made using Equation 15.

Whereas the preceding study illustrates elasticity adjustment for route-specific applications, the same procedure can be applied to time-of-day variations. In the latter application, the subscript k in Equations 12 through 14 simply takes on the meaning "peak period," "off-peak period," and so forth. The rest of the procedure follows in a manner similar to the disaggregation by route, keeping in mind that we are now performing temporal instead of spatial aggregation.

TABLE 1 INTERZONAL TRAVEL TIME AND COST

	To Zone j						
From zone i	1	2	5	10	14	15	32
1	0	2.7/68 (4.1)	3.8/28 (6.0)	6.1/22 (10)	5.6/24 (10.7)	6.1/24 (10.1)	8.3/35 (14.9)
2	27/29 (4.1)	0	5.2/35 (9.2)	4.6/44 (7.5)	7.1/31 (13.8)	7.6/31 (13.3)	6.8/57 (12.3)
5	3.8/31 (6.0)	5.2/99 (9.2)	0	8.6/53 (15.2)	3.7/4 (7.0)	6.2/8 (10.1)	10.8/66 (19.9)
10	6.1/29 (10.1)	4.6/22 (7.5)	8.6/42 (15.2)	0	10.5/38 (19.8)	11.0/38 (19.3)	3.2/28 (5.8)
14	5.6/27 (10.7)	7.1/95 (13.8)	3.7/4 (7.0)	10.5/49 (19.8)	0	4.8/8 (6.9)	13.7/62 (24.5)
15	6.1/27 (10.1)	7.6/95 (13.3)	6.2/12 (10.1)	11.0/49 (19.3)	4.8/8 (6.9)	0	14.2/62 (25.4)
32	8.3/55 (14.9)	6.8/48 (12.3)	10.8/68 (19.9)	3.2/41 (5.8)	13.7/64 (24.5)	14.2/64 (25.4)	0

KEY (for each entry): auto-time-in-min/transit-time-in-min (auto-cost-in-cents)

A FORMULA FOR PRACTITIONERS

The foregoing represents a formal development of the elasticity adjustment procedure. The development is necessarily predicated on the availability of a disaggregate logit model for the study area. In real-world applications, however, a calibrated logit model is seldom available. The only available information is often a set of demand elasticities—if one is fortunate enough to have them. Frequently, elasticity figures have to be borrowed from similar cities in the form of a range of numbers that are disparate in value (*18*). Care should be exercised in the definition of a "similar" city, which should include both size and urban structure to guarantee transferability (*5,10*). The problem now is to choose an approximate value within this range for the application at hand.

Suppose maximum and minimum algebraic values are both available, representing the two extremes of the range. It is possible to work backwards to get an appropriate value through the use of an adjustment factor. From Equations 7 and 15, it can be assumed that

$$\frac{\eta_{\min}(m, x_m)}{\eta_{\max}(m, x_m)} \approx \frac{(1 - \bar{p}_m)}{\left[(1 - \bar{p}_m) - \dfrac{\hat{\sigma}^2}{\bar{p}_m}\right]} \tag{26}$$

inasmuch as the two elasticity estimates come from the two extreme levels of aggregation with the $\eta_{\min}$ having the most aggregation bias and $\eta_{\max}$ the least. Assume further that the average modal split $\bar{p}_m$ can be approximated by empirical data obtainable locally (meaning the logit model replicates observed data reasonably well). The term σ^2, being the only unknown in Equation 26, can now be estimated.

If more than two elasticities are available, we can have even better information on the variability of elasticities among routes or time of day. Let us say we have a third elasticity for the study area. It is clear that a third level of aggregation was used in model calibration, different from the previous two. Again, we take the ratio according to Equation 19:

$$\frac{\eta_k(m, x_m)}{\eta_{\max}(m, x_m)} \approx \frac{\left[(1 - \bar{p}_m) - \dfrac{\sigma_k^2}{\bar{p}_m}\right]}{\left[(1 - \bar{p}_m) - \dfrac{\hat{\sigma}^2}{\bar{p}_m}\right]} \tag{27}$$

which allows σ_k, the only unknown in the equation, to be estimated.

In general, the availability of more than one calibrated elasticity, rather than being confusing, is now an asset. The more elasticity tabulations available, the more we can reconstruct the elasticity variability among routes and time of day. Suppose these σ's are obtained:

$$\sigma_{\min} < \sigma_1 < \cdots < \alpha_{K-2} < \hat{\sigma} \tag{28}$$

We can now match each route k (or time of day k) against one of the σ's. A rule of thumb may be that the shorter the route (or the shorter the time period) the less the σ value. A shorter route has less variability in such explanatory variables as travel time, travel cost, and car ownership, and hence less variability in mode choice:

$$\frac{\sigma_k}{\hat{\sigma}} = \frac{n_k - n_{\min}}{n_{\max} - n_{\min}} = s \tag{29}$$

where n_k is the number of stops on the route under consideration, and $n_{\min}$ and $n_{\max}$ are the number of stops on the shortest and longest routes. Such a linear scale, s, will place the mode-choice variability of route k at

$$\sigma_k = \hat{\sigma} \cdot s \tag{30}$$

Other rules of thumb can be used if additional information on local travel behavior becomes available, thus allowing a better match between routes and σ's in Equation 28.

As an example, suppose the disaggregate fare elasticity values are available (*19*) as a range of values characterized by a mean and a standard deviation: -0.35 ± 0.14. In the absence of the actual constituent figures that make up the range, we take two standard deviations (i.e., 0.28) from the mean of -0.35 as the maximum and minimum elasticity value. For a normal distribution this would cover 95 percent of the 12 data points in the sample: $\eta_{\max} = -0.07$ and $\eta_{\min} = -0.63$.

Suppose further that the observed modal split in York is 88.83 percent automobile trips and 11.17 percent transit trips. From Equation 26, $\hat{\sigma}^2$ can be estimated as 0.0882 (or $\hat{\sigma} = 0.297$), which is a higher variance than the data sample collected in Chicago by Warner.

The W. Market/E. Market line was 1 of 10 bus routes operating in York. With its 17 stops, it was among the longer routes in the York Area Transit Authority. Only two routes were longer, made up of 18 stops, and the remaining seven were 14-stop routes. If route length can be a proxy for mode-choice variability, we can scale the W. Market/E. Market route as somewhere between $\hat{\sigma}$ and $\sigma_{\min}$. If a linear scale s is applied as shown in Equations 29 and 30, the variability associated with the route concerned is $\sigma_k = 0.222$, corresponding to $s = \frac{3}{4}$. According to Equation 27, this translates to a route elasticity of -0.314. Compared with the average, -0.35, a 10 percent difference is observed in this case. Again, other routes can have biases much larger than 10 percent, considering that the maximum variance $\hat{\sigma}$ is close to 0.3.

SUMMARY AND CONCLUSIONS

In assessing user response to cost and service changes, demand elasticities are useful tools. Current compilations of demand elasticities, however, are not helpful to scheduled transportation operators. The range of an elasticity is too wide, and there is no practical guideline for picking an appropriate value in the range. Furthermore, they are often calibrated on the basis of different levels of aggregation, rendering them incompatible with route-by-route or peak versus off-peak analyses, which are critical to current operational concerns.

A methodology providing elasticities that are practical for patronage analyses in an operating agency was presented to address this problem. Through statistical analyses of spatial/temporal data aggregation, the methodology explains the differences among elasticity tabulations, and in so doing, provides insights into the variations that exist among elasticity values.

The results of this research include (*a*) guidelines for scaling an appropriate value of elasticity among its broad range of

values (instead of simply taking the average or midpoint of the range) and (*b*) a method for converting the most commonly available elasticities, which are usually calibrated in different levels of aggregation, to a more useful form, such as route-specific and time-of-day elasticities.

Rigorous yet simple statistical developments were followed in deriving the adjustment procedure. The results were demonstrated in a case study of a transit line in York, Pennsylvania, including a step-by-step calculation procedure for practitioners.

Additional work can obviously be carried out to extend this research. It is recommended that more case studies be performed, particularly studies geared toward time-of-day rather than route-specific applications. One such case study could include before-and-after validation (*20*).

ACKNOWLEDGMENT

The author wishes to thank Morgan Wong of Washington State University for the numerical calculations he performed in York, Pennsylvania. He is also grateful to Fong L. Ou, who stimulated much of his thinking in the formal theoretical development of the aggregation procedure. Obviously, the author alone, not former or present associates, is responsible for the statements made in this paper.

REFERENCES

1. D. J. Moon. *Application of Individual Behavioral Demand Models to Fixed-Route Transit System Planning*. M.S. thesis. Pennsylvania State University, University Park, 1977.
2. D. K. Boyle. Are Transit Riders Becoming Less Sensitive to Fare Increases? In *Transportation Research Record 1039*, TRB, National Research Council, Washington, D.C., 1985.
3. E. J. Miller and D. F. Crowley. Panel Survey Approach to Measuring Transit Route Service Elasticity of Demand. In *Transportation Research Record 1209*, TRB, National Research Council, Washington, D.C., 1989.
4. C. P. Cummings, M. Fairhurst, S. Labelle, and D. Stuart. Market Segmentation of Transit Fare Elasticities. *Transportation Quarterly*, Vol. 43, No. 3, July 1989, pp. 407–420.
5. Y. Chan and F. L. Ou. Tabulating Demand Elasticities for Urban Travel Forecasting. In *Transportation Research Record 673*, TRB, National Research Council, Washington, D.C., 1978, pp. 40–46.
6. T. W. Usowicz. Measured Fare Elasticity: The 1975 BART Fare Change. Draft. San Francisco Bay Area Rapid Transit District, San Francisco, Calif., 1980.
7. Multisystems, Inc. *Route Level Demand Models: A Review*. Report DOT-J-82-6. Urban Mass Transportation Administration, U.S. Department of Transportation, 1982.
8. J. N. Bajpai and R. E. Johnson. DEL: A Fare Revenue Forecasting Model. Presented at the National Conference on Microcomputers in Urban Transportation, San Diego, Calif., 1985.
9. G. Kocur, T. Adler, W. Hyman, and B. Aunet. *Guide to Forecasting Travel Demand with Direct Utility Assessment*. Report UMTA-NH-11-0001-82-1. UMTA, U.S. Department of Transportation, 1982.
10. Y. Chan and F. L. Ou. Equilibration Procedure To Forecast Areawide Travel. *Journal of Transportation Engineering*, Vol. 112, No. 6, 1986, pp. 557–573.
11. D. P. Ballou and L. Mohan. Application of a Fare Policy Evaluation Software Package To Assess Potential Transit Pricing Policies. *Transportation Quarterly*, Vol. 37, 1983, pp. 395–410.
12. D. Krechmer, G. Lantos, and M. Golenberg. *Bus Route Demand Models: Cleveland Prototype Study*. Report DOT-I-83-34. UMTA, U.S. Department of Transportation, 1983.
13. Tri-County Metropolitan Transportation District. *Bus Route Demand Models: Portland Prototype Study*. Report DOT-1-83-37. UMTA, U.S. Department of Transportation, 1983.
14. D. McFadden. The Measurement of Urban Travel Demand. *Journal of Public Economics*, Vol. 3, 1974, pp. 303–328.
15. J. Gerken. Generalized Logit Model. *Transportation Research B*, Vol. 25B, Nos. 2/3, 1991, pp. 75–88.
16. S. L. Warner. *Stochastic Choice of Mode in Urban Travel: A Case Study in Binary Choice*. Northwestern University Press, Evanston, Ill., 1962.
17. Y. Chan. *A Case Study for the Techniques of Transportation Analysis*. Department of Civil Engineering, Washington State University, Pullman, 1985.
18. Y. Chan, F. L. Ou, J. Perl, and E. Regan. *Review and Compilation of Demand Forecasting Experiences: An Aggregation of Estimation Procedures*. Report PTI-7708. Pennsylvania Transportation Institute, Pennsylvania State University, University Park, 1977.
19. *Patronage Impacts of Changes in Transit Fare and Services*. Report 135-1. Ecosometrics, Inc., 1980.
20. D. H. Pickrell. *Urban Rail Transit Projects: Forecast Versus Actual Ridership and Costs*. Report DOT-T-91-04. U.S. Department of Transportation, 1990.

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

Trip Characteristics and Travel Patterns of Suburban Residents

PANOS D. PREVEDOUROS AND JOSEPH L. SCHOFER

Increasing traffic congestion in U.S. suburbs can be explained to a large degree by their rapid growth. Much is still to be learned, however, about the causes of and variations in traffic congestion. The results of investigations of variations in travel behavior across social groups and between locations are presented. The investigations were based on a mid-1989 mail-back survey of individuals residing in selected Chicago suburbs. Four prominent factors associated with traffic congestion are residence location, population aging, working women, and fixed work hours. Residence location in outer-ring, low-density, growing suburbs implies longer trips and more local trips because of low density and more employment opportunities, respectively. The average travel speed by automobile is higher for residents of growing suburbs, but because of longer commutes they still stay in traffic 25 percent longer than residents of stable suburbs. Population aging may offer some relief to suburban traffic congestion, not because older people travel less, but because they make better use of off-peak periods and shorter trips. Hence, their travel behavior may equalize use of the roadway infrastructure over the day. The increasing number of working women and mothers further contributes to congestion because long work trips are added to the large number of household maintenance trips made by women. The morning and evening peak periods remain short in duration. It would make a tremendous difference in peak loads and network performance if the observed 1½-hr peak were spread over 2½ to 3 hr.

Increasing traffic congestion in U.S. suburbs can be explained to a large degree by their rapid growth. Much is still to be learned, however, about the causes of and variations in traffic congestion. The resulting knowledge may be helpful in identifying promising solutions to these problems.

To expand our understanding of contemporary suburban travel, we explored the trip characteristics of individuals residing in selected Chicago suburbs using a mail-back survey conducted in the spring of 1989. The variations of several characteristics of individual travel behavior across social groups and between locations (i.e., outer-ring, low density, growing suburbs and inner-ring, high density, stable suburbs) were investigated.

Initial analysis using aggregate census data enabled us to classify suburbs into growing and stable (*1,2*). From the original sample of 30 suburbs, we selected four, two growing (Naperville and Schaumburg) and two stable (Park Ridge and Wilmette).

We collected a sample of 1,420 responses; the response rate approached 25 percent. One member of the household responded. The responding member was requested to be an adult and preferably a worker. The respondent supplied full demographic and socioeconomic information on the household, work trip destinations, and a list of automobiles available to the household.

Most respondents filled out a 1-weekday travel diary. Thus, trip statistics are based on responses from individuals. An initial plan to request household travel diaries was abandoned because the questionnaire was becoming too long. The final version was 10 pages.

P. D. Prevedouros, Department of Civil Engineering, University of Hawaii at Manoa, Honolulu, Hawaii 96822. J. L. Schofer, Department of Civil Engineering, Northwestern University, Evanston, Ill. 60208.

WORK LOCATIONS

First, some of the most important differences between outer-ring, low-density, growing suburbs and inner-ring, high-density, stable suburbs are reviewed. Table 1 gives basic information for these two types of suburbs. The statistics are from a sample of 30 suburbs, 13 stable and 17 growing. The data were taken from census reports (*1,2*).

Outer-ring, low-density, growing suburbs are different in important ways from stable suburbs. Growing suburbs are far from Chicago's central business district (CBD) and have both a low population density and a high population growth rate. Furthermore, they are populated by larger and younger households, which have a higher automobile ownership compared with households of stable suburbs.

Table 2 presents the work locations of employed residents in the four suburbs examined, broken down by employment status.

The overwhelming majority (78 percent) of full-time-employed residents of growing suburbs are employed in the suburb where they live or in another suburb, and less than 20 percent are employed in the central city (Chicago). In contrast, 45 percent of residents of stable suburbs work in the central city. The statistics clearly support the hypothesis that part-time workers tend to work closer to home than full-time workers (68 and 44 percent work in the suburb of residence for growing and stable suburbs, respectively).

Only 16 percent of full-time-employed residents of stable suburbs work in the suburb of residence. In contrast, about 29 percent of full-time-employed residents of growing suburbs are employed in their home suburb. This difference may contribute to local traffic congestion in growing suburbs (i.e., more commuting on arterials and local streets than on expressways).

This pattern may be attributed to the fact that suburbs experiencing recent growth spurts have been willing and able to accommodate employment as well as residential development. On the other hand, older, inner-ring communities were

TABLE 1 CHARACTERISTICS OF GROWING AND STABLE SUBURBS

COMMUNITY CHARACTERISTICS	GROWING	STABLE
DISTANCE FROM CBD (mi)	27	15
POPULATION DENSITY (residents/square mile)	2900	5200
POP. GROWTH (1970-80)	142%	-2%
AVG. HOUSEHOLD SIZE	2.89	2.81
AVG. POPULATION AGE	31.7	37.8
AVG. AUTOS/HOUSEHOLD	2.07	1.96

All differences significant at 95% [3]

TABLE 2 WORK DESTINATION BY TYPE OF SUBURB

	RESIDENCE LOCATION AND EMPLOYMENT STATUS			
	GROWING SUBURBS		STABLE SUBURBS	
DESTINATION	FULL-TIME	PART-TIME	FULL-TIME	PART-TIME
SAME SUBURB	29.3	68.1	16.4	44.0
OTHER SUBURB	48.6	28.4	37.9	33.3
suburb total	77.9	96.5	54.3	77.3
CENTRAL CITY	19.3	2.8	44.6	18.4
MULTI-PLACE	2.8	0.7	1.1	4.3
	100%	100%	100%	100%

designed primarily as residential communities. Therefore, there are fewer employment opportunities for their residents. The employment-to-residents ratio is higher in the growing suburbs of our sample: Schaumburg's is 1.31; Naperville's is 0.93. Both are likely to be much higher now, because the pace of development in these areas between 1980 and 1990 was higher than ever before, whereas it is lower in the stable suburbs—Evanston (0.99) and Park Ridge (0.79). Evanston is atypical in this respect because it has a CBD of considerable size and a major university (Northwestern University). Aggregate census data indicate that stable suburbs tend to be consistent with respect to the employment-to-residents ratio. Most average about 0.85 with a standard deviation of 0.30. On the other hand, growing suburbs appear to form two extreme clusters: one with strong employment orientation (average ratio 1.30) and one with strong residential orientation (average ratio 0.50). The overall average for growing suburbs is 0.91 and the standard deviation is 0.60. These statistics are based on the 17 growing and 13 stable suburbs.

TRIP STATISTICS

This section presents an overview of the trip characteristics of respondents for various combinations of locations and sociodemographic groups. Analysis of the factors affecting trip characteristics is presented later. All trip statistics are from weekday travel diaries of individual respondents (1-day travel activity; the day was chosen by the respondent).

Table 3 presents mode shares for the primary work trip (i.e., the line-haul trip). Statistics do not include the characteristics of access trips at the ends of the primary trip; this explains why the share of bus, walk, and bicycle modes are not included. None of these modes was used for the primary trip by the surveyed respondents.

The automobile mode dominates by far. However, the share varies substantially: 81.0 and 87.6 percent for stable and growing suburbs, respectively. [In the growing suburb of Schaumburg the automobile share is as high as 95.0 percent because of the lesser public transportation service and the fewer workers who are employed in the central city (*3*).] The rest of the market share is picked up by public transportation, mostly commuter rail (Metra). Most residents of Naperville (one of the two growing communities surveyed) who work in Chicago commute by Metra (17.8 percent of full-time workers work in the central city and 17.4 percent of all workers commute by rail to Chicago's CBD). Rapid transit is not available to any outer-ring suburb, including the two growing suburbs in the sample; it is available to residents of both stable suburbs (Park Ridge and Wilmette).

Table 4 presents average distances and speeds for commuters with respect to the geography of their work trips. The presentation is separate for automobile and transit.

TABLE 3 MODE SHARES FOR PRIMARY TRIP TO WORK (PERCENT)

MODE	GROWING SUBURBS (n=776)	STABLE SUBURBS (n=644)
DRIVE ALONE	82.7	77.0
DRIVER AND PASSENGER(S)	4.3	3.6
PASSENGER IN CAR	0.6	0.4
TOTAL AUTO	87.6%	81.0%
COMMUTER RAIL (METRA)	12.4	11.7
RAPID TRANSIT (CTA 'EL')	0.0	7.3
TOTAL TRANSIT	12.4%	19.0%

TABLE 4 PRIMARY WORK TRIP MODE STATISTICS

Part (a): auto mode

FROM ↓ / TO →		SAME SUBURB	OTHER SUBURB	CENTRAL CITY
GROWING SUBURB	avg.distance	6.3 *	15.5 *	28.4 *
	avg.speed	19.5 N	26.0 N	31.5 *
	cases	152	314	69
STABLE SUBURB	avg.distance	4.5	10.9	12.9
	avg.speed	20.3	25.1	22.3
	cases	56	198	56

Part (b): public transit

FROM ↓ / TO →		SAME SUBURB	OTHER SUBURB	CENTRAL CITY
GROWING SUBURB	avg.distance	no observations due to minimal or non-existent public transit		31.2
	avg.speed			49.6
	cases			62
STABLE SUBURB	avg.distance			17.0
	avg.speed			28.7
	cases			73

NOTES: 1) distance in miles; speed in miles/hour is derived from respondents' reports of time and distance
2) * = t-test between growing and stable: significant at 95%
N = t-test between growing and stable: not significant

Average automobile speeds for trips within the same suburb and to other suburbs are similar for growing and stable suburbs; they are also remarkably low (i.e., approximately half the typical 35- to 45-mph speed limit on suburban arterials). However, the data do not support the hypothesis that congestion in growing suburbs is worse (average speed is higher for residents of growing suburbs). Furthermore, things look better for growing-suburb residents who commute to Chicago; these are mostly trips on expressways.

On the other hand, growing-suburb residents who work in the city are exposed to traffic and congestion for longer times. Residents of growing suburbs sit in their cars for 54 min for a typical commute to Chicago's CBD during the rush period, whereas residents of stable suburbs do the same for 35 min. In addition, because of low densities, commutes of growing-suburb residents to the same or another suburb are longer (see Table 4); therefore they are again exposed to more traffic and congestion than residents of stable suburbs. For trips to

the CBD, commutes by public transportation are substantially faster than by automobile. That is because most CBD trips are made on the commuter rail system.

Conceivably, the uproar about traffic congestion in growing suburbs may be because growing-suburb residents are exposed to traffic over longer times. As a result, they are more inconvenienced and tend to be more critical of the performance of the roadway system. Another reason may be that 10 years ago the traffic conditions in most outer-ring suburbs were acceptable if not really good; these conditions have worsened dramatically over the past few years, which may have created the impression of a crisis to residents. In contrast, worsening of traffic in inner-ring suburbs came much more gradually, so people had more time to adapt to and accept them.

Analysis of variance indicated that residence location has an insignificant effect on the daily number of trips made by individuals. It has, however, a significant effect on the distance traveled, as will be discussed later. Sex and employment status have a significant effect on the daily number of trips of suburban residents (*3*); their effect is explored in Table 5, which presents a breakdown of trip statistics by purpose for each sex and employment status combination.

The trip rates in Table 5 indicate the following:

- Females consistently make more trips than males in each employment status category, whereas not-employed people make roughly 0.5 more trips in a day than individuals employed full time.
- Females employed part time indicate a remarkably high trip activity. This may be partly because they have the burden of both household maintenance trips and work-related trips.
- Not-employed people make up for their minimal trips to and from work by making more trips for errands, groceries, personal business, and recreation.
- Employed females make nearly twice as many trips as males for errands, groceries, and shopping. Also, females make more trips than males to serve passengers (i.e., drive children, day-care person, husband to station, etc.). Full-time-employed males make slightly more work-related trips (i.e., more men are in travel-intensive jobs) than full-time-employed females. (Of all the occupations listed by our respondents, we considered as travel-intensive the following: managerial/business owner, sales, and professional/technical. Eighty percent of male respondents have a travel-intensive occupation; the corresponding number for females is 58 percent.)

These trip rates represent the trip activity of adult household members, most of whom are employed. This may explain the seemingly high numbers. The total household trip activity per person (number of trips per person) is expected to be lower because some household members cannot travel on their own or may not need to travel regularly.

Analysis of trip purposes separately for each weekday indicated that Thursday is the most representative day because it closely matches the average distribution of trips by purpose for the five weekdays (*3*). On the basis of the data, Thursday may be the best day for representative traffic measurements.

TIME-OF-DAY TRIP PROFILES

Figure 1 shows time-of-day trip profiles for selected groups of respondents. The plots show the distribution of trips in each hourly period for each population group analyzed (for example, the portion of trips between 8 and 9 a.m. is the ratio of the number of trips between 8 and 9 a.m. to the total number of trips for each population group between 6 a.m. and 9 p.m.). Figure 1a shows an interesting difference between workers residing in growing and stable suburbs: full-time-employed residents of growing suburbs depart from home earlier in the morning (and from work in the evening), presumably to compensate for their longer commutes.

The lowest line represents the difference between growing and stable suburbs in the portion of trips per hour of full-time workers. The difference again indicates that growing-suburb workers make a larger portion of their daily trips

TABLE 5 TRIPS PER DAY BY PURPOSE, EMPLOYMENT STATUS, AND SEX

TRIP PURPOSE	ALL (n=1420)	FULL-TIME MALE (781)	FULL-TIME FEMALE (301)	PART-TIME MALE (47)	PART-TIME FEMALE (91)	NOT-EMPL'D MALE (93)	NOT-EMPL'D FEMALE (107)
WORK	1.53	1.91	1.74	1.18	1.02	0.10	0.05
RETURN HOME	1.76	1.65	1.68	1.68	2.42	2.00	2.05
ERRANDS/GROCERIES	0.54	0.30	0.55	0.55	1.08	1.12	1.31
SHOPPING	0.09	0.04	0.09	0.02	0.13	0.23	0.25
PERSONAL	0.17	0.09	0.14	0.20	0.27	0.47	0.52
SERVE PASSENGERS	0.34	0.22	0.31	0.36	0.90	0.29	0.81
EXERCISE/SPORTS	0.12	0.10	0.11	0.09	0.21	0.14	0.17
RECREATION/SOCIAL	0.35	0.29	0.33	0.27	0.42	0.74	0.48
TOTAL TRIPS PER DAY	4.90	4.60	4.95 *	4.35	6.45 *	5.09	5.64 N

NOTE: (*) = difference significant at 95%; (N) = not significant

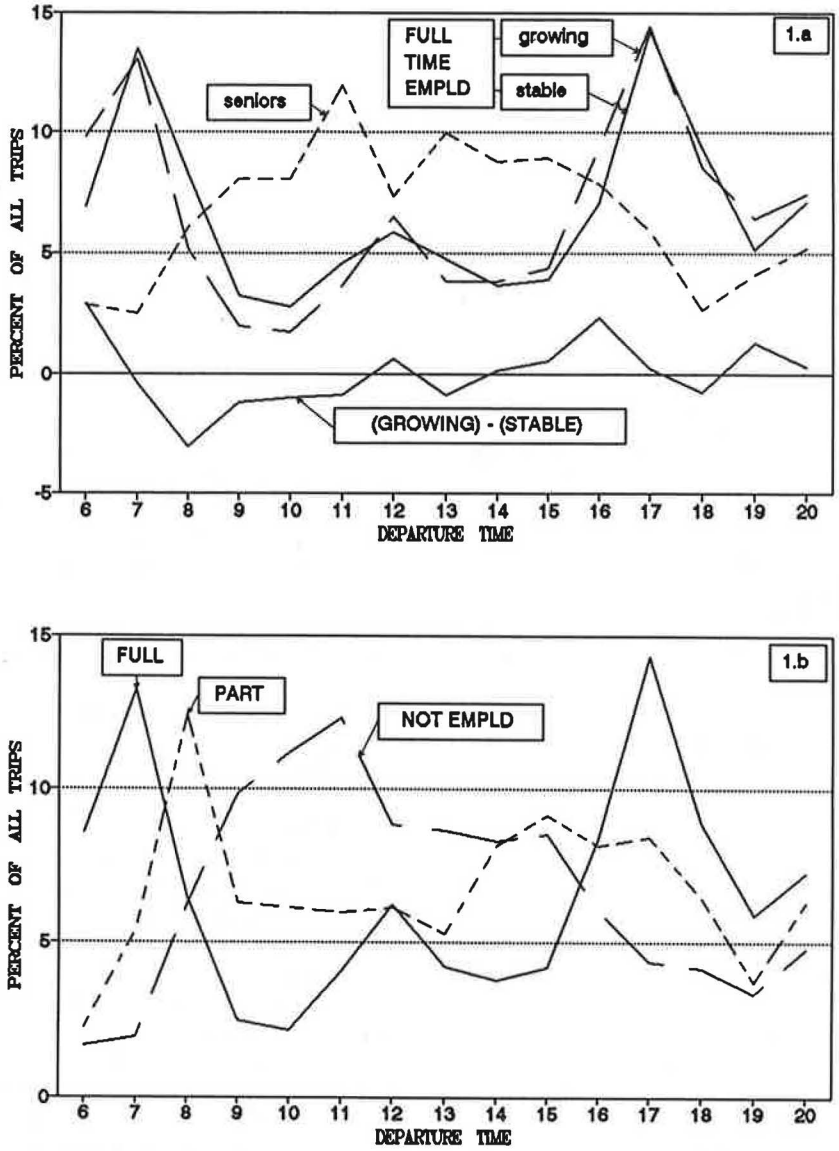

FIGURE 1 Trip profiles by time of day for selected population groups.

during the traditional three peaks of a weekday (morning, noon, and evening) than stable-suburb residents. The dense dashed line in this graph represents the trip profile of respondents from all the sample's senior households (single-person or couple households without children; at least one of the members age 65 or older). It can be inferred that they tend to avoid the morning, noon, and evening rush periods. These findings provide indications and inferences only. The averages cannot, at this point, be subjected to meaningful statistical significance testing because they are the results of trip aggregation per time slot, not per individual.

Full-time-employed people make the bulk of their trips during the morning and evening rush hours, whereas not-employed people travel mostly during the valleys in traffic demand (Figure 1b). The bulk of their trips take place after the morning and before the evening rush periods. Figure 1b also shows that part-time-employed people make the bulk of their morning trips 1 hr after the full-time workers, and they tend to spread the rest of their traveling uniformly over the hours between 9 a.m. and 9 p.m.

The resulting figures confirm the speculation that substantial traffic is observed on suburban road networks during off-peak periods. For example, excluding senior and not-employed people, the portion of trips made at 7 or 8 p.m. is higher than the portion made at 1 or 2 p.m. for all the other population groups examined. This finding applies to growing as well as to stable suburbs.

Figure 2 categorizes rush-hour automobile trips by purpose. No distinction is made between growing and stable suburbs because the differences in the profiles are not statistically significant. The figure shows that in the morning peak the destination of 77.5 percent of trips is work, whereas the destination of 62.6 percent of trips in the evening peak is home. The share of secondary purposes (other than work and return home) increases substantially in the evening peak, with the exception of trips to serve passengers, which is higher in the

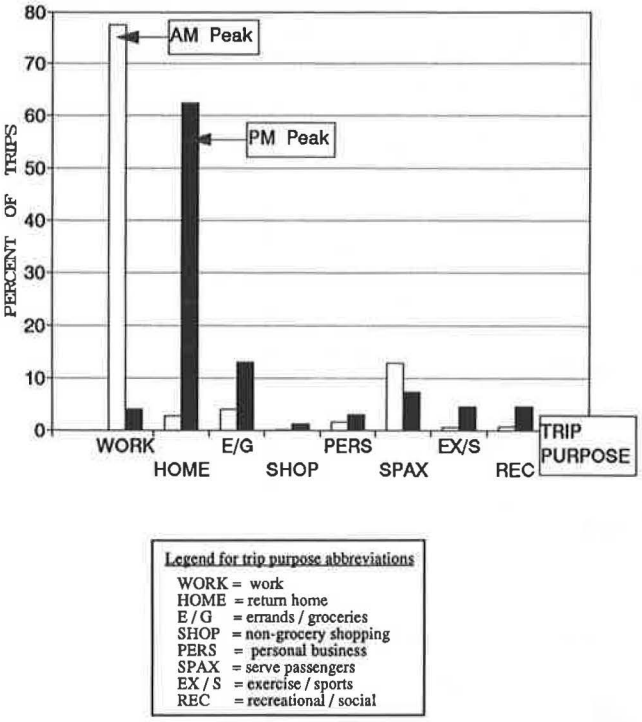

FIGURE 2 Breakdown of rush-hour trips by purpose (automobile mode only).

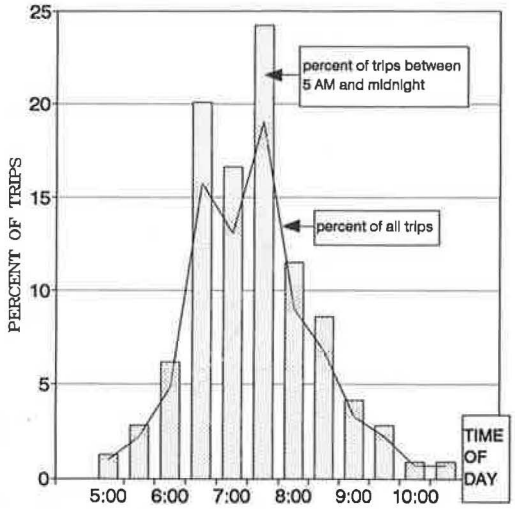

FIGURE 3 Pattern of morning departures from home for trip to work (all modes).

morning peak. It is important to realize that in the evening peak nearly 35 percent of trips are destined to places other than home or work. Part of this phenomenon is due to trip chaining (some people stop at intermediate destinations before returning home). A substantial proportion of people (12.0 percent) return home late in the evening (after 8 p.m.).

The fact that more evening rush-hour trips are for purposes other than return to home may be partly because of additional congestion—people stay in the system longer in terms of time and distance—and many persons drive to multiple destinations that may not be on the route from work to home.

Figure 3 shows the temporal pattern of the primary trip to work (the primary trip is the line-haul trip without the access trips at the ends, if such trips exist). The bulk of departures are observed between 6:30 and 7:30 a.m., so if commuting time range between 30 and 60 min, the road network is expected to carry peak loads between 6:30 and 8:30 a.m. These results prompt us to suggest that the time interval for morning traffic counts in suburban residential areas should be between 6:30 and 8:30 a.m., and for employment centers the interval for morning traffic counts should be from 7:00 to 9:00 a.m.

The figures from the morning peak period indicate little evidence of substantial peak spreading, partly because staggered and flextime work schedules do not appear to be popular among employers in the Chicago metropolitan area (*4*).

Similar analysis for the late afternoon and evening hours indicates that the bulk of departures occur between 4:30 and 6:00 p.m., whereas the bulk of arrivals—at all destinations—occur between 5:00 and 6:30 p.m. Thus, the transportation network carries peak loads between 4:30 and 6:30 p.m., which is the recommended interval for evening peak-period traffic counts.

FACTORS AFFECTING TRAVEL BEHAVIOR

This section presents results of analyses of factors that affect or explain travel behavior characteristics of individual respondents. The characteristics analyzed are number of trips (all trips, work trips, and nonwork trips), automobile share for the total number of trips reported by the individual respondent, and distance traveled by automobile.

Several potentially explanatory variables were tried using an analysis of variance procedure. Variables that have a significant effect on at least one of the trip characteristics examined are presented in Table 6, which gives the percentage of the contribution of each independent variable to the explained variance of the trip characteristics examined.

The independent variables listed on the left-hand side of Table 6 are defined as follows: EMPL.STATUS is the employment status of the respondent (full-time, part-time, or not employed). RES.LOCAT. is the residence location (growing or stable suburb). CAR AVAIL. is the availability of automobiles to each eligible-to-drive member of the household, specified for three categories: less than 0.5, 0.5 to 0.9, and more than 0.9 automobiles per driver. TRANSIT is the utilization (and familiarity) with public transportation by the respondent (1 if the respondent listed at least one trip by public transportation in his or her travel diary, 0 otherwise). AGE includes four groups for the age of the respondent: 34 or younger, 35 to 49, 50 to 64, and 65 or older.

A substantial portion of variance is explained by the variables available for the number of trips for work purposes and the distance traveled by automobile for all purposes. Observations from Table 6 are as follows.

Employment status explains much of the variation in the number of trips (particularly work trips) a respondent makes each day (see also Table 5). Part-time-employed females make more but shorter trips (6.45 trips with average distance equal to 5.6 mi) than their male counterparts (4.35 trips with average distance equal to 7.8 mi).

Sex plays an important role for the trip purposes of errands, groceries, and shopping, which are dominated by females.

TABLE 6 CONTRIBUTION OF EACH INDEPENDENT VARIABLE TO THE EXPLAINED VARIANCE OF TRIP CHARACTERISTICS (PERCENT)

Dependent->	# of TRIPS		% AUTO	TOTAL DIST.
Independent	ALL	WORK	SHARE	by AUTO
EMPL.STATUS	36.8*	32.6*	11.9@	1.4N
GENDER	0.9N	0.5N	10.2*	0.4N
RES.LOCAT.	0.2N	0.0N	15.5N	3.4@
CAR AVAIL.	0.7N	0.1N	5.7@	1.2N
TRANSIT	21.3*	22.9*	- -	78.7*
AGE	40.1*	1.0*	7.6N	6.0@
Interactns	0.0	42.9	49.1	8.9
	100%	100%	100%	100%
% Explained	7.7	49.6	14.5	46.7

NOTE: (*) = significant at 95% or higher
(@) = significant at 85%
(N) = not significant

The difference in automobile share between males and females is significant. Suburban females exhibit a behavior opposite to that of central city females—suburban females use automobiles more than do males (93.7 versus 88.7 percent). In contrast, old and recent studies (*5*) both suggest that central city females use public transportation more than do males.

This may partly be an outcome of work destinations. A larger proportion of suburban females work in the suburbs than males (78 versus 67 percent). People traveling to the central city have the option of using public transportation. This option hardly exists for intra- or intersuburban travel. As expected, more males and females of growing suburbs work in the suburbs: growing = 73 percent of males and 85 percent of females; stable = 54 percent of males and 67 percent of females.

The contribution of the residence location suburb in the portion of the variance explained for the total distance traveled by automobile in 1 day is marginally significant in the particular specification tested in Table 6. However, a specification without the TRANSIT variable as well as a *t*-test between the means indicated that the contribution and the difference, respectively, are statistically significant at the 95 percent level (i.e., the total distance traveled by automobile for growing-suburb residents is significantly higher than for stable-suburb residents). Nonwork distance is significantly different for residents of growing and stable suburbs as well.

Transit usage reduces the number of trips and increases the number of work trips. It greatly reduces total distance by automobile. These results are logical, because most transit users in the sample are daily work trip commuters going to the Chicago CBD by commuter rail.

Poor results were obtained from analysis of variance specific to each trip purpose (*3*). This may be an outcome of a lack of comprehensive information on a complex process, the intrahousehold trip-trading process (i.e., we have information from one household member only). This is also true for the total number of trips, given in Table 6. Trip trading among adult household members is largely defined by their role in the household and by their personal needs and constraints.

Trip trading is an underlying process of great importance. Ignoring this process by considering individual members only is bound to result in limited understanding of the trip-making behavior of individuals. Household travel diaries must be gathered to assess this household process.

We employed linear regression modeling to assess the effect of each contributing (or causal) factor on the travel behavior of individuals. Table 7 gives four models of trip characteristics.

The independent variables are SEX (0 = female, 1 = male), AGE (the exact age of the respondent), FULL TIME (1 if employed full-time, 0 otherwise), NOT EMPL (1 if not employed, 0 otherwise), RES.LOCAT. (1 if residence in growing suburb, 0 if residence in stable suburb), and TRANSIT (1 if at least one trip in diary made by public transportation, 0 otherwise).

The overall fit of the number of total trips model is poor, but all parameters are significant and correct in sign. Much better results were obtained with separate estimates for work and nonwork trips.

The number of work trips appears to be affected to a large extent by the employment status of the respondent. Males tend to make fewer nonwork trips compared with the sample average.

The role of a person's age in travel behavior in terms of total number of trips, work trips, and nonwork trips made in a day is fascinating, particularly for nonwork trips. The relationship between the number of trips and age is known to be nonlinear. The nonlinear effect plotted in Figure 4 is the combined effect of the AGE + AGE^2 specification, which resulted in better model performance than specifications with the AGE variable alone.

The age of a person, up to 80, is positively correlated with the number of trips made in 1 day. The number of trips contributed by the person's age to the total number of trips is highest at 40 [i.e., most people at age 40 are at the stage

TABLE 7 MODELS OF TRIP CHARACTERISTICS

Dependent	TOTAL TRIPS	NONWORK TRIPS	WORK TRIPS	DISTANCE by AUTO
R-squared	0.08	0.28	0.33	0.14
Cases	1296	1303	989	1311
Constant	3.52	3.44	0.11@	37.60
SEX (male)	-0.30	-0.44	0.14	-
AGE	0.15	0.12	0.04	-
AGE-sq'd	-0.002	-0.0015	-0.0005	
FULL TIME	-1.59	-3.41	1.82	-
NOT EMPL.	-	-	-	-13.70
RES.LOCAT.	-	-	-	6.30
TRANSIT	-	-	-	-33.90

NOTE: all parameters significant at 95%, except
(@) = parameters not significant

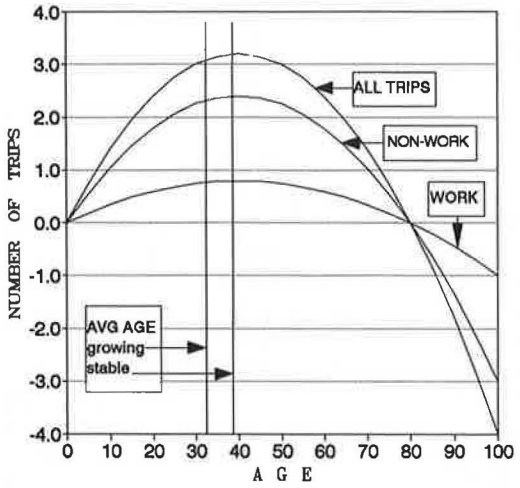

FIGURE 4 Contribution of age to travel behavior (number of trips by purpose).

of the life cycle with young, dependent children, whose needs induce a substantial number of trips among the adults in the household (*6*)]. Clearly, age has a much larger effect on non-work trips. This is intuitive, because the number of work-related trips is largely affected by the type and hours of work and transportation alternatives and costs.

The underlying age distributions are different in growing and stable suburbs. The average population age is 31.7 and 37.8 years in growing and stable suburbs, respectively, which suggests a higher trip activity in stable than in growing suburbs (Figure 4). This is an aggregate estimate that tends to ignore the real age distribution in our sample. Disaggregate estimation suggests that the correlation between the number of trips and age alone results in 449 and 534 trips in growing and stable suburbs, respectively, for every 100 residents in each type of community.

Given the U.S. population aging projections (*7*), it is likely that overall stable suburbs will progress toward higher age cohorts (i.e., beyond 40), which may result in reduced trip activity. Growing suburbs will progress closer to 40 years of average age, which will put them at the highest trip activity age cohort. Precise estimation of the effect of aging cannot be obtained with our data because important information on in- and out-migration rates is not available. Projections rely on the assumption that behaviors do not change over time. We may wonder, though, whether it is safe to assume that the baby-boom generation (the old generation of tomorrow) will have a travel behavior similar to that of today's old generation.

The distance traveled in 1 day by automobile is greatly affected by the use of public transportation. Commuting to work by public transportation saves the average respondent 34 mi of driving. This result is largely due to suburban residents employed in the Chicago CBD who commute by public transportation (commuter rail or rail rapid transit). Most of these people make few other trips by automobile during weekdays (i.e., they have little time left for such travel), and most of their daily mileage is picked up by public transportation.

Residence in a growing suburb adds, on the average, 6 mi to the daily distance traveled by automobile. This estimate translates into roughly 15 more min of exposure to traffic, plus added fuel consumption and pollution. [For each automobile user, 6 more mi a day translates into 80 gal of fuel a year additional (assuming 250 workdays and 20 mi/gal average fuel efficiency). This estimate does not include weekend travel; conceivably, weekend travel distances are also longer for growing-suburb residents.] Not-employed people travel on the average 14 mi less by automobile compared with the sample average.

CONCLUSIONS AND IMPLICATIONS

Findings from an analysis of contemporary transport behavior survey data collected from selected Chicago suburbs were presented. Two classes of suburbs were used in the analysis: inner-ring, high-density, stable suburbs and outer-ring, low-density, growing suburbs. The presentation focused on empirical findings of trip characteristics of individual respondents and on the factors affecting or explaining them.

Key findings are summarized as follows:

- More than 40 percent of stable-suburb full-time workers are employed in the central city, but less than 20 percent of

growing-suburb full-time workers are employed in the central city. The majority of part-time workers work in the suburb where they reside.

- Average speeds for the trip to work by automobile are similar for growing- and stable-suburb residents, but growing-suburb residents stay in traffic at least 25 percent longer because of their longer commuting distances.
- Full-time-employed people make the bulk of their trips during the morning, noon, and evening rush periods, part-time-employed people spread their trip activity almost uniformly, and not-employed people use mostly off-peak time periods.
- A substantial number of trips are made during off-peak times, including late-evening hours, which supports the impression of around-the-clock traffic (and congestion).
- The number of trips made in 1 day by individual respondents is significantly different across employment status categories: full-time-employed people make 4.7 trips, part-time-employed people make 5.9 trips, and not-employed people make 5.3 trips.
- Males make more work-related trips, but females make more trips to run errands, buy groceries, and shop.
- Suburban females depend more on automobiles for their transportation (93.7 percent automobile share) than do males (88.7 percent automobile share).

Most of the findings have clear associations with and implications for traffic growth and congestion in the suburbs. The four most prominent factors are residence location, population aging, working women, and fixed work hours.

Residence location in growing suburbs implies longer trips and more local trips because of low density and more local employment opportunities, respectively. Specifically, growing-suburb residents travel a 40 percent longer total daily distance compared with stable-suburb residents. The average travel speed by automobile is higher for growing-suburb residents, but they stay in traffic 25 percent longer than stable-suburb residents because of longer commutes. The longer exposure to traffic (and the consequent waste of time) may have caused the uproar against congestion in growing suburbs. Despite the fact that the number of daily trips does not differ significantly between residents of growing and stable suburbs, growing-suburb residents make 80 percent of their trips in the suburbs, whereas stable-suburb residents make less than 60 percent of their trips in the suburbs. The result is more traffic (and congestion) in growing suburbs.

The U.S. population will be increasingly moving to higher age cohorts until about 2020 (*7*). This natural phenomenon may offer some relief of traffic congestion, not because older people travel less, but because they use off-peak periods and make shorter trips. Hence, their travel behavior may equalize use of the roadway infrastructure over the day and thus may be less wasteful and polluting. The data suggest that older people now reside in stable suburbs to a greater extent than in growing suburbs; thus, in the short term (i.e., the next 5 to 10 years) some relief may come to places where public discomfort due to congestion is less pronounced. Worse, natural aging alone can be expected to contribute to increasing trip rates in growing suburbs, at least for the next decade.

The increasing number of workers, working women and mothers in particular, further contributes to congestion because long work trips are added to the traditionally large number of household maintenance trips by women. Equal employment rights and the independence gained from earning an income are incentives for women to work. In addition, the increasing cost of living and opportunities for consumption may be forcing households to have multiple workers, first to make ends meet, and then to improve their standard of living by spending on education, entertainment, fitness, and possessions.

Finally, there is room for spreading peak-period demand over longer periods. It is disappointing that flextime or staggered work hours has not appealed to company executives. Recall that the overwhelming majority of the workers in our sample start their trips between 6:30 and 8:00 a.m. (Figure 3). It is clear that it would make a tremendous difference in peak loads and network performance if this 1½-hr peak were spread over 2½ to 3 hr.

ACKNOWLEDGMENT

This work was supported in part by a contract from the Illinois Department of Transportation. The authors gratefully acknowledge this support and the helpful suggestions made by Ronald Eash and others at the Chicago Area Transportation Study.

REFERENCES

1. P. Prevedouros and J. Schofer. Suburban Transport Behavior as a Factor in Congestion. In *Transportation Research Record 1237*, TRB, National Research Council, Washington, D.C., 1990, pp. 47–58.
2. P. Prevedouros and J. Schofer. Social and Cultural Factors Affecting Transportation Behavior in Rapidly Developing Suburbs. In *1989 WCTR Proceedings*, Yokohama, Japan, 1989.
3. P. Prevedouros and J. Schofer. *Demographic, Social and Cultural Factors Affecting Suburban Congestion.* Illinois Department of Transportation, 1990.
4. B. Kamin and D. Ibata. Choking on Growth. *The Chicago Tribune*, Feb. 18–22, 1990 (series of five articles).
5. Midwest System Sciences, Inc., and Market Opinion Research, Inc. *Market Analysis of the Chicago Travel to Central Business District.* Chicago Transit Authority, Chicago, Ill., 1989.
6. P. Allaman and T. Tardiff. Usefulness of Indicators of Household Structure and Characteristics of Residential Zones in Trip Generation Models. Presented at 61st Annual Meeting of the Transportation Research Board, Washington, D.C., 1982.
7. *Moving America: New Directions, New Opportunities, Volume I: Building the National Transportation Policy*. U.S. Department of Transportation, 1989.

The findings of this research are the responsibility of the authors, and do not necessarily reflect the views of the Illinois Department of Transportation or the Chicago Area Transportation Study.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Socioeconomics of the Individual and the Costs of Driving: New Evidence for Travel Demand Modeler

A. P. Talvitie and Maria Koskenoja

Automobile travel costs have long been neglected in travel behavior analyses and travel demand models. The socioeconomic determinants of automobile travel cost choices are explored as part of a study into the effects of quality of work on work-related travel conducted in the Road and Traffic Laboratory, Technical Research Center of Finland. The study was motivated by the need to assess driving cost estimates for the mode choice model. It was approached through separate functions for fixed, variable, and total driving costs of the respondent. Within-household effects were examined through driving cost regressions of the household members of the respondent. The within-household regressions indicate that there are within-household work life effects on driving costs. The effects take place on a detailed level, which is not captured in a variable describing the household member's employment status. Rather, it appears that the quality of occupations of the household members has more effect on driving costs than the number or share of employed persons in the household. Automobile operating costs are not equal for everybody. They are an outcome of a choice and are influenced by the characteristics of the individual and his activities and the other household members and their activities. The same applies for automobile capital costs. The operating costs of automobiles had a low statistical significance on mode choice, and high capital costs were associated with automobile choice. These findings are not surprising but ones not believed nor a part of present travel demand model systems. But it is believable that automobile operating costs do not influence mode choice, or, that once automobile is chosen, it is an expensive one—comfort costs—by choice. The driving cost models indicate that travel choices are a part of complex behavioral interplay within a household in which the costs of transport have a major role.

Analyses of automobile travel costs have long been neglected in travel behavior studies and in developing travel demand models, with some exceptions (*1–4*). Models for automobile travel costs are developed and the socioeconomic determinants of automobile travel cost choices are explored as part of a study into the effects of type of work on work-related travel conducted in the Road and Traffic Laboratory, Technical Research Center of Finland.

In the present study rigid ex ante hypotheses are avoided. Rather, it is generally hypothesized that work arrangements in their entirety influence work-related travel choices and that these decisions interact with non-work-related transport decisions (*5*). In a similar vein, the models developed may not satisfy all the technical, statistical assumptions. The emphasis has been on exploring behavior, not mathematics.

A. P. Talvitie, Viatek Consulting Engineers and Architects, Pohjantie 3, Espoo, 02100 Finland. M. Koskenoja, University of California, Irvine, Calif.

The data for the study come from Finland, where both men and women work. In 1984, 49 percent of all employed people in Finland were women, of whom 88 percent work more than 30 hr per week; for men the figure is 98 percent. The percentage of the work force at the working ages is 79 for men and 72 for women. Work histories of employed men are only 2 years longer than those of employed women for persons aged 35 to 45. The difference gradually increases, but only to 7 years for persons aged 55 to 64. Thus, work and the working environment are influential factors for both sexes in the formation of preferences for everyday choices (*6–8*).

The contents, tools, and procedures of work are also changing because of changes in market structure and technology. Besides, work conditions are an object of planning by management and institutions. If there is a link between work arrangements and travel behavior, there is a need to understand the effects of such arrangements in order to respond to the changing transportation needs of different groups of working individuals.

DATA

Sampling

The sample was drawn by systematic sampling using the Population Register of Finland. The population was divided into two classes by age: adults (over 15 years) and children (7 to 14 years). Of the adults, persons were selected who are in the active work force and reported the travel diary day to be a normal workday. There were 1,518 such adult, working persons in the survey, which the diary cleaning reduced to the final sample size of 1,333 persons.

Data Collection

The transportation data were collected by the Central Statistical Office of Finland between September 1985 and February 1986. The collection procedure was as follows:

1. Persons in the sample received a letter from the Statistical Center explaining the purpose of the survey and the expectations concerning the respondent.
2. Three questionnaires to be filled out by the respondent were attached to the letter, including a note pad for the daily travel pattern, called the travel diary.

3. The interviewer contacted the respondent before the intended day of the study and made an interview appointment.
4. The data were gathered in the interview. The travel diary was translated from the note pad to the questionnaire form with help from the interviewer to avoid misperceptions.

The collected data are representative of the population, but the process of selecting diaries may bias accepted travel diaries toward young, male, transportationally active people. This was just a feeling that came when choosing the usable data diaries from the original sample.

Occupational Characteristics

Occupational characteristics come from a study conducted by the Central Statistical Office of Finland (*7*), in which 4,502 wage earners were interviewed about their conditions of work. The interviews took place in 1984, just 1 year before the transportation data collection, ensuring the compatibility of the two data sets.

In the occupational data survey the respondents gave subjective ratings about different dimensions of occupational requirements. The occupations were classified into 34 groups on the basis of similarities in activities; this, however, is not the official occupational grouping used in government statistics. The grouping of occupations causes some aggregation error, but the grouped data still include more of the variance than the normally used, single-dimension, class description of organizational status classes (blue collar, white collar, etc.).

Occupational groups were described by detailed questions. For example, the physical demands of an occupation were scaled through deviations from normal temperature, draft, dustiness of air, heavy lifts, extreme stretching, and so forth; from these an index of physical demands of work was formed. The chosen occupational dimensions are mental demands, physical demands, monotonousness, and the power to influence work rhythm.

The subjective nature of occupational ratings, together with technological and market structure changes, could soon cause the description of occupations to be outmoded if the characteristics of occupations are not updated. Fortunately, it is in the interest of labor unions and employer associations as well as university and statistical bureaucracies to produce these kinds of data. Therefore, the usage of occupational characteristics offers a possibility to use more disaggregated data, which are updated by existing institutions.

The occupational data were connected with the transportation questionnaire data by using the occupation reported in the questionnaire.

In the transportation questionnaire the respondents were also asked about weekly working hours, whether they had a shift schedule or irregular working hours, whether flextime was used, and whether they worked at home or at several locations. These variables are truly disaggregate in the combined data base.

Household Structure Variables

The age, gender, and employment status of the household members are known. Also, the personal income class and the income class of the household were asked. These data were used to form the life cycle variable and an indicator of whether the person was self-sustaining, the spouse of the primary earner, or a dependent. Intrafamily effects are not well covered, because there are virtually no transportation data from other household members. The relationships of daily tour combinations of household members remains to be investigated.

There was no clear indicator of the main wage earner of the household. This would have been desirable for determining the economic power positions of the respondents inside the household and for deducing the life cycle of that household. The decision rule used for selecting the main wage earner was the following: the respondent was classified as the main earner if she or he earned at least half of the total income of the household. If that was not the case, the presence of other working household members between ages 19 and 64 was determined. If they existed, the main earner was deemed to be the other in-law, mother, father, or a spouse, in that order. For example, if there were both a father and a spouse present, the spouse was classified as the main earner. This information was condensed to identify the respondent's position in the household: self-sustaining, spouse, or dependent.

The household life cycle has five classes. The first life cycle class contains childless households in which the main earner is younger than 35. Households are classified in the second class if the youngest child is less than 7 years old, irrespective of the age of the main earner. The third life cycle class has a youngest child between 7 and 16 (the school age), again regardless of the age of the main earner. A household without children under 17 is classified in the fourth class if the main earner is between 35 and 54. The fifth class contains households in which the main earner is 55 or older and there are no children under 17.

The classification of life cycles based on the age of the youngest child was selected because it was thought that the youngest person restricts mobility the most. This assumption is commonly made when forming life cycle variables. Zimmerman (*9*) discusses alternative ways to select life cycle variables.

Diary Data Cleaning

The following information concerning starting place and destination was collected in the diary: addresses, place codes, and times of day. The purpose of the trip, travel mode, and waiting time for public transportation were also asked.

The checking of the data had to be done manually, case by case, to ensure the right corrections. It took almost 1 year to clean the data. Distances were checked for illogical or not reported entries and, when necessary, measured from maps. In addition, starting places were corrected to be the same as previous destinations. Diaries that contained intractable obscurities in the paths were deleted. Diaries were also deleted if any work trip was done on an unusual mode (plane, boat, or "other"). The modes included in the analyses were automobile driver, automobile passenger, bus, tram, train, bicycle, and walk.

The data contained diaries that were otherwise tractable but in which the distance of a home-based walk round-trip (with purpose sport or unspecified) was unreported. These

diaries were accepted with a change of distance of 1.0 km for the walking trip. Most of these walks were the last trips in the diary, so they are "evening walks" (taking the dog out, window shopping). Finally, a diary was deleted from the data if the respondent's travel day started from an implausible place, such as school, shop, other place of errands, or a day-care center.

Even though numerous judgments had to be made in preparing the data for model estimation, they are believed to be no worse or less accurate than data normally used in transportation models. That references to data-cleaning work are rarely made in the literature or, worse, that the analyst is not aware of how data were prepared, does not improve data quality. The view taken is that a thorough knowledge of data improves the appraisal of model results.

DRIVING COSTS

In developing a mode choice model, the time and money costs of driving must be approximated for the sampled individuals. The driving costs of cars are normally obtained first, then the driving costs of trips are estimated using the information about the cars available to the person in question. This figure is an estimate of the realized driving costs for persons who had a car available that day.

The method just described can also be used to approximate the driving costs for a person who at the moment did not have a car. The question to be answered in that case is, If the person had a car, what kind of driving costs would he or she have incurred per driven kilometer? This approximation to driving cost is different from the one customarily used in travel model studies, where driving costs are assumed to be constant regardless of the car driven and, indeed, regardless of whether or not a car is available.

Driving Costs per Car—Assumptions and Initial Values

The cost approximations that appear in this section are based on research conducted in the Technical Research Center of Finland about car prices and factors affecting driving costs. To calculate driving costs, cars are divided into five size groups and three age groups:

Size Class	*Size (cc)*	*Purchase Price, New (FIM)*	*Age Class (years)*
I	$x < 1300$	40,000	0–3
II	$1300 < x < 1600$	55,000	4–9
III	$1600 < x < 2000$	75,000	>10
IV	$2000 < x < 2500$	120,000	
V	$2500 < x$	150,000	

For simplicity it is assumed that all cars are driven 17 000 km/year. This was the average annual mileage in Finland in 1985. It is casually believed that most kilometers are driven with new cars. If this is true, their driving costs will be estimated high and the driving cost for old cars low. Unfortunately, the yearly mileage figures were not available. This lack of better data is regrettable but not unusual in transportation studies as the analogous example by Train (*10*), reported later, illustrates.

Fixed costs include depreciation, interest, mandatory traffic and car insurance, anticorrosion treatment, storage, and parking at home. Variable costs include tires, repair and maintenance, and fuel.

Depreciation of capital during the first 3 years is set at 15 percent of the remaining capital. The rest is depreciated over the following 6 years. When the car reaches the age of 10 years, all the original capital is depreciated. Whatever value the car may have thereafter is considered to be because of spare parts and repair and maintenance put into it. The interest rate used is 6 percent per the average capital of the year in question. The rate is lower than the market rate because interest costs are partly tax deductible; 6 percent is estimated to be the part of interest the owner really has to pay.

Anticorrosion treatment is calculated to be 450 FIM/year regardless of car size or age. The costs of tires and insurance rise with car size class as follows:

Size Class	*Tires (FIM)*	*Insurance (FIM)*
I	650	2,014
II	700	2,362
III	800	2,709
IV	900	3,057
V	900	3,751

Storage and parking costs depend on the respondent's living environment. In the central city they are 150 FIM/month, in the suburbs they are 75 FIM/month, and in towns and rural areas storage and parking costs are assumed to be nil.

Repair and maintenance costs are calculated as a percentage of the purchase price of the car. The percentage is lower for new cars, reflecting their better condition, and also lower for bigger cars due to their better quality. The following percentages apply in each age and size group; fuel consumption is also given:

Size Class	*New Cars (Age 0–3 years) (percent)*	*4–9 year-old-cars (percent)*	*Fuel consumption (L/100 km)*
I	2	4	8
II	2	4	9
III	1.75	3.5	10
IV	1.5	3	12
V	1.5	3	14

For cars more than 10 years old, the repair percentage is a flat 10 percent in every size class. The fuel price is 3.54 FIM/L, which was the price of regular gasoline in October 1985.

Fixed and variable costs are summed and divided by 17,000. The results are shown in Table 1. The three different costs for each car size class are due to the different storage and parking costs in city centers (c), suburbs (s), and villages or rural areas (r).

After the driving costs per car are approximated, the cost per driven kilometer for different persons can be estimated.

Driving Costs of Respondents with Cars at Their Disposal

The data indicate the number and makes of vehicles in the household and their primary users. One of the many alter-

TABLE 1 VARIABLE AND FIXED COSTS OF DRIVING IN FINLAND (FIN/km) ($1 U.S. = 4.0 FIM)

Size Class		I	II	III	IV	V
* CAR AGE 0-3 years						
fixed	c:	0.64	0.80	1.02	0.47	1.80
	s:	0.58	0.75	0.96	1.42	1.75
	r:	0.53	0.70	0.91	1.36	1.69
variable	:	0.36	0.42	0.48	0.58	0.68
tot.cost	c:	1.00	1.22	1.50	2.05	2.48
	s:	0.94	1.17	1.44	2.00	2.44
	r:	0.89	1.12	1.37	1.94	2.37
* CAR AGE 4-9 years						
fixed	c:	0.52	0.65	0.80	1.12	1.38
	s:	0.47	0.59	0.75	1.07	1.33
	r:	0.42	0.54	0.70	1.01	1.27
variable	:	0.41	0.49	0.56	0.69	0.81
tot.cost	c:	0.93	1.14	1.36	1.81	2.19
	s:	0.88	1.08	1.31	1.76	2.14
	r:	0.83	1.03	1.26	1.70	2.08
* Car AGE 10+ years						
fixed	c:	0.25	0.27	0.29	0.31	0.35
	s:	0.20	0.22	0.24	0.26	0.30
	r:	0.14	0.17	0.19	0.21	0.25
variable	:	0.58	0.68	0.84	1.18	1.43
tot.cost	c:	0.83	0.95	1.13	1.49	1.79
	s:	0.78	0.90	1.08	1.44	1.73
	r:	0.72	0.85	1.03	1.39	1.68

natives was that use of the car was shared. A small proportion of the respondents indicated this to be the case. The most common case was that the car belonged to a certain person in the household. Often one person was the main user of several cars.

The cost allocation problem arises because the diary data do not indicate which car the person was driving or got a lift on. The following rules are used to approximate the personal driving cost of the respondent:

1. If the person reports one or more cars to be at his personal disposal and there are no cars in common use, the fixed costs are the sum of the cars that the person has.
2. If no one else can use the cars, the fixed costs are all summed, because the owner can drive only one car at a time, but the time runs on all of the cars. The variable costs are estimated to be a mean of the variable costs of those cars.
3. If the household has one or more cars in shared use, the fixed cost is the sum of all those cars divided by the number of driving licenses in the household. The variable costs in shared use are again a mean of the variable costs of the cars.
4. If the person has some cars at his or her personal disposal and some in shared use, only the cars at personal disposal are counted in estimating costs. This is a simplifying assumption, but in most cases the vehicle in shared use was a van and the vehicles in personal use were cars. Thus the vans drop off when the mode of travel is driving a car, because in the questionnaire there is another mode for driving a truck or van.
5. Total driving costs are the sum of fixed and variable costs.

Driving Costs for Respondent's Household Members with Access to a Car

Because cars are allocated to certain household members, it is possible to formulate models describing the effect of the respondent's characteristics on the chosen driving cost level of some other household member, provided he or she has access to a car. To gain this understanding of the effects of a person on another person's driving cost choices, the data were used in reverse order: the cost choices of the respondent's household members were regressed on the characteristics of the respondent.

Of interest here are the respondent's work type or socioeconomic characteristics that may cause changes in the household members' driving costs. For example, the young age of a respondent could induce someone else in the household to change driving costs. For example, a British study (*11,12*) reports higher values of time when there are small children in the car. It is plausible that this affects the driver's choice of driving costs, that is, the choice of car type.

In the data there is much information about the personal characteristics of the respondent. However, there is not much information about the other household members. In the linear regression model all variables are introduced to account for both the effect of those household characteristics that are common to all of the household members (e.g., location of the household) and those that vary among individuals (an example of this type of variable is "main earner of the household"). As in the previous section, three cost variables were created: others' fixed costs (OFC), others' variable costs (OVC), and others' total costs (OTC). They refer not to the respondent, but to another household member with access to cars. The following rules were adopted for calculating OFC, OVC, and OTC: OFC is the mean of the fixed costs of the cars belonging to the other household members. OVC is the mean of the variable costs of these cars. OTC is simply OFC plus OVC.

Correction for Self-Selectivity Bias in Modeling Driving Costs for Persons Who Do Not Have a Car

Marketing science suggests that different persons drive different kinds of cars. There may be combinations of driving costs a particular person would not even consider as an alternative.

In estimating potential driving costs for persons who at present are nondrivers, it is important to know what determines the chosen costs of driving for those who are drivers. This information can then be applied to the nondrivers to produce a realistic driving cost alternative for them.

If this calculation is done simply on the basis of the observed characteristics of the drivers, it will yield biased estimates. This is because the variables that affect the chosen driving cost level also affect the choice to be or not to be a driver. This bias is called the self-selectivity bias. The correction for this bias is presented next.

The decision to have a car and the decision of driving cost level are made jointly. According to Train (*10*), models of car ownership and driving quantity should be linked to avoid the self-selectivity bias. He proposes an instrumental variable

estimation, where the probability of car ownership is estimated first and this estimate is used to form the correction term in the driving quantity regression. In Train's model the driving cost was fixed and the amount of driving was estimated. In the model here, the amount of driving is fixed and the cost of driving varies. The logic of the correction term is not violated by this change.

Following Train, if the choice probabilities are logit and the error term of the driving cost function is normal, the correction term is

$$E(e_c) = -(\sqrt{6}\sigma/\pi) * p_c[P_q \ln P_q/(1 - P_q) + \ln P_c]$$

where

e_c = error term in the driving cost equation;
$E(e_c)$ = $e_c - \tau$, selectivity correction term, which equals the error term in the driving cost function—normally distributed;
P_q = probability of owning one or more cars;
P_c = probability of owning no cars;
σ = standard deviation of e in the entire population (not conditional on the choice of the number of cars); and
p_c = correlation of e with the unobserved utility associated with owning no cars.

When this correction term is added to the driving cost regression, the rest of the parameters are unbiased:

$$vc = \boldsymbol{\beta}\mathbf{s} + \Gamma C_c + \tau$$

$$fc = \boldsymbol{\alpha}\mathbf{s} + \Gamma C_c + \tau$$

where

vc = variable cost conditional upon being a driver,
$\boldsymbol{\beta}$ = vector of parameters to be estimated,
$\mathbf{s}$ = vector of characteristics of the person and other explanatory variables,
τ = a normally distributed error term,
C_c = $[P_q \ln P_q/(1 - P_q) + \ln P_c]$, and
Γ = $-(\sqrt{6}\sigma/\pi) * p_c$.

Because the value of C_c can be calculated, the value of p_c comes from the estimation.

Train also proposes that the utility of choosing the car (make or model and vintage) from the varying number of alternatives inside each class should be accounted for by a correction term. This reflects the assumption that if a car is selected from a class containing several alternatives, the choice will be closer to the optimum than when there are only a few makes or models to choose from. Because this kind of correction is not done here, the estimates may be biased. Because each class is likely to contain a large number of alternatives, the bias is small.

Train estimated the models for the probability of having no car, one car, and two cars and used these estimates in the correction term of the equation for operating costs. In his model the exogenous variables for predicting the operating costs of households' cars were gas price in the area of residence, household income, household size, type of housing unit, number of adults and adolescents, number of workers, age of household head, education level of household head, sex of household head, distance to work, population of household's area of residence, and number of transit trips in the area of residence.

The variables in this study are expected to be different because the equations are estimated not for a household but for an individual in the household. An important variable, distance to work, does not exist in the present data.

Logit Model for Respondent's Car Availability

To obtain an estimate for the correction term to avoid the self-selectivity bias, a binomial logit model was estimated. The alternatives were no cars available (nondriver) and one or more cars available (driver). If the person shared the use of at least one car, he or she was classified as a driver. The set of variables that were considered important is given in Table 2, and a "best" model estimated using these variables is given in Table 3.

The model predicts the probability that the respondent is a driver, that is to say, is the primary user or shares the use of at least one car. The automobile availability model is needed

TABLE 2 EXPLANATORY VARIABLES IN CAR AVAILABILITY MODELS

Biological Variables:	
family life cycle	age of the youngest hh member
year of birth of the respondent	sex (female=1, male=0)
Time connected Variables:	
works outside home	works flexible hours
several places of employment	hours worked in a week
works regular hours	works in shift
works irregular hours (=deadline)	
Organizational Variables:	
farmer	self-employed
upper white collar	lower white collar
blue collar worker	
Industry Variables:	
primary production	manufacturing
energy and utilities	construction
commerce, restaurant and hotel	finance and insurance
transportation, and communications	services
Occupational Variables:	
mentally demanding	physically demanding
much/little impact on the work rhythm	monotonous
Relative Economic Power of the Household Members:	
monthly personal income	monthly household income
main wage earner	spouse of the main earner
dependent (=Neither of the above)	
Household Structure:	
number of persons in the household	number of working persons
number of persons not active in work life	
Distances:	
distance to the nearest bus stop	distance to the food store
distance to a bank	distance to a post office
Locational variables:	
traffic volume of home street	home in the central city
home in a suburb	home in rural area
home in a small town or village	
House type:	
single family or a detached house	townhouse apartment
owns a summer cottage	can use somebody's summer cottage
Driving variables:	
r(espondent) has a car available	r's household member has a car available r has a
drivers licence	no. of driver licenses in the household
others' driving costs	

TABLE 3 LOGIT MODEL FOR RESPONDENT'S CAR AVAILABILITY (DEPENDENT VARIABLE: RESPONDENT HAS A CAR AVAILABLE = 1)

Independent Variable	Estimated Coefficient	t-Statistic
constant	-13.997	-6.907
female	-1.895	-11.076
self-employed	1.659	3.749
works outside home	0.619	1.870
several employments	0.499	1.882
hours worked	0.000	2.300
log of pers income	0.883	3.020
log of hh income	0.621	2.289
dependent	-0.785	-2.321
others drive	-2.743	-11.194
# of other licenses	0.939	5.831
rural area	0.963	3.722
one family house	0.930	4.177
row house	0.817	2.986

auxiliary statistics	at convergence	initial
log likelihood	-461.674	-779.097
number of observations	1124	
percent correctly predicted	82.2	

in estimating car driving costs. A few comments on the model follow.

Females have a smaller probability of being drivers than do males. Self-employment, workplace outside home, several employers, and the hours worked during a week all increase the probability of being a driver. The work location variables are statistically significant at one-sided 0.05 level, and they were included in the model because of their expected sign and because the effect of work type on travel behavior is the subject of this study.

The occupational variables (mentally demanding, physically demanding, monotonous, and possibility of influencing work rhythm) all lost their significance when other variables were introduced. Both physically demanding and monotonous occupations are correlated with not being a driver. Contrary to expectations, flexible working hours, shift, or variable (deadline) working hours did not have a statistically significant impact on car availability.

Both the logarithm of personal income and the logarithm of household income are positive and statistically significant. Personal income would have been statistically significant also without the logarithmic transformation, but the logarithmic transformation was preferred, to reflect the declining marginal utility of income.

The variables describing the person's relative economic strength in the household—main wage earner, spouse, and dependent—are mutually exclusive dummy variables. "Dependent" receives a value of 1 when the person in question is not the head of the household (is not the main wage earner) and is not the spouse of the head. "Spouse" marks all persons whose spouse is the main wage earner.

"Dependent" receives a statistically significant negative value indicating that, in spite of income, this social standing works toward not having a car at one's disposal. If "dependent" and "spouse" are both included in the model, they both get a negative coefficient, but for "spouse" it is smaller and not statistically significant ($t = 1.3$).

Car availability in the household by persons other than the respondent has a strong negative effect on the respondent's car availability and suggests transferability of driving tasks. Conversely, when the number of driving licenses in the household increases, so does the respondent's probability of acquiring a car. The explanation for this could be a possibility of more efficient car usage, learning the way of life from other household members, or both.

Finally, living in a rural area and in a one-family detached house or a town house is positively correlated with the respondent's car availability.

Logit Model for Car Availability of the Household Members of the Respondent

The estimated logit model for the respondent's household members to have a car available is given in Table 4. The variables, which all describe the respondent, were chosen with three primary concerns in mind: they should be statistically significant, interpretable, and preferably the same variables as in the respondent's car availability model, to see their effect inside the household and in relation to the outside of the household.

The other (nonrespondent) members of the household tended to have a car available if the respondent was female. The nonavailability of the respondent for household tasks (much employment and large number of hours worked weekly) is one inducement for the other members of the household to acquire a car. Self-employment of the respondent works in the same direction. The explanation could be lack of time for household tasks or, just as well, a learned way of life. Again, flexible working hours, shift, or variable (deadline) working hours did not have any impact.

The probability that a household member has a car available increases with household income, as expected. However, the probability diminishes with increasing income of the respondent. This is not totally explained by the income differentials, because those are captured in the dummy variables "main wage earner" and "dependent." This characteristic is statis-

TABLE 4 LOGIT MODEL FOR CAR AVAILABILITY OF HOUSEHOLD MEMBERS OF RESPONDENT (DEPENDENT VARIABLE: RESPONDENT'S HOUSEHOLD MEMBER HAS A CAR AVAILABLE = 1)

Independent Variable	Estimated Coefficient	t - Statistic
constant	-11.485	-5.540
female	0.841	4.116
sevrl employments	0.365	1.446
hours worked	0.000	2.521
self-employed	0.657	2.015
log of hh income	3.110	9.163
log of pers inc	-2.216	-6.241
main earner	0.603	1.939
dependent	0.793	2.046
traffic volume	-0.173	-2.683
rural area	0.173	0.739
summer cottage	0.827	4.262
one family house	1.493	6.753
townhouse	1.057	4.009
r has a car avail	-2.611	-10.334
has a licence	0.874	3.544

auxiliary statistics	at convergence	initial
log likelihood	-483.621	-765.234
number of observations	1104	
percent correctly predicted	79.9	

tically significant and does not change when other variables are changed. It needs further investigation.

Car availability is negatively correlated with the traffic volume on the street of residence, perhaps reflecting the level of service of public transportation. Curiously enough, the distance to the nearest bus stop, which was thought to indicate the level of service, did not have any significance. The same locational variables that were significant in driver model estimation are significant here: living in a rural area and in a one-family house or a townhouse are all associated with an increased probability that the household member has a car available.

Another similarity to the driver model is that the driver status of the respondent reduces the probability that other household members are drivers, but the driving license of the respondent increases it. This again could be interpreted as indicating the transferring of travel-related tasks inside the household.

Driving Cost Models for Respondents

The driving cost models were estimated by standard least squares estimation with the previously discussed correction term included in the model. The correction term's logic is briefly presented here; see Figure 1. The term corrects the regression for "unobserved observations," that is, for the would-be-chosen cost levels that are not possible because of technical or market reasons such as one shown in Figure 1.

If income were the only factor influencing the chosen driving cost level, and the threshold amount of income were the one marked with the horizontal line, then the observed driving cost level and income pairs would be the ones above the line. If a regression were run on these pairs only, the estimated line would not be steep enough and the intercept would be higher than in the true regression. The reason for this is the double influence of income: it affects the decision to have or not have a car and the level-of-cost decision. If totally different variables affected the automobile and the cost-level choice, the correction term would be unnecessary.

The cost regressions were estimated separately for fixed costs, variable costs, and total costs. The division between these costs is interesting because decisions about these costs are supposedly made on different time horizons—fixed costs for longer horizons and variable costs for shorter. This assumption is supported in the empirical findings from longitudinal studies that the transfer cost of selling the old car and buying a new one is considered high (*10*). This means that people do not buy and sell their cars to adjust to the optimal combination of costs, but rather decide the fixed-cost level and let the variable costs have their impact on the amount of driving.

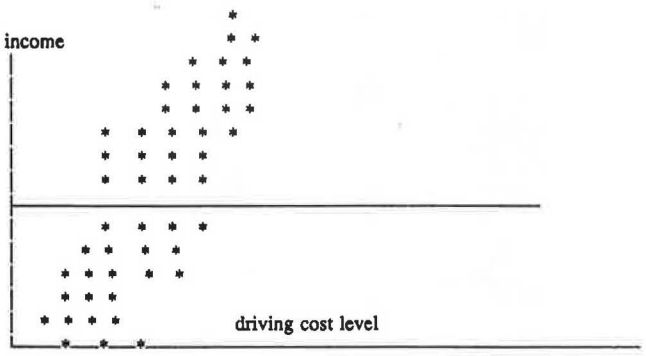

FIGURE 1 Rationale for correction term due to "unobserved" observations in driving cost regression.

Model for Fixed Costs of Driving a Car

The fixed-cost model is discussed first (see Table 5). The correction term, which carries the effect of presently having/not having access to a car, has the expected sign; the probability of having a car available increases the fixed costs. The coefficient has a low significance, but the term was left in the regression on theoretical grounds.

The life cycle dummy (cycle2) marks households in which the youngest child is under school age. People living in households with small children are driving cars with lower fixed costs. (Fixed costs are lower for smaller and older cars.) Female drivers have lower fixed cost.

The following occupational attributes are statistically significant: irregular working hours (deadline), self-employed, working in the energy and utility industry, and mentally demanding occupation. These attributes increased the chosen fixed driving costs.

Working irregular hours increases the person's scope for planning the day's activities and the number of alternatives from which to choose. It is consistent with the theory that if the person can choose an alternative closer to the optimum, the value of time is higher. This also implies that the chosen monetary cost is higher. The MVA Consultancy (*11*) found that the value of time was higher for people working irregular hours.

The occupational variable concerning impact on work rhythm did not have significance. The effect of self-employment may explain that variation. It is usual for entrepreneurs to describe the positive sides of their work as independence and flexibility in the sense that the person can agree about the deadlines.

TABLE 5 MODEL FOR FIXED COSTS OF DRIVING A CAR (DEPENDENT VARIABLE: FIXED DRIVING COSTS OF RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.023	1.016
constant	0.358	7.607
cycle2	-0.072	-2.640
female	-0.101	-2.459
deadline	0.113	3.290
self-employed	0.166	3.803
energy industry	0.131	1.961
mentally demanding	0.002	2.685
personal income	0.000	2.731
summer cottage	0.071	2.480

Number of Observations	604
R-squared	0.16
Sum of Squared Residuals	50.570
Standard Error of the Regression	0.292
Mean of Dependent Variable	0.566

Working in the energy and utility industry may increase fixed costs simply because the locations visited during the day may be hard to reach by other modes; for example, distant farmhouses, voltage leveling stations, or power lines are rarely served by public transit. Mental and physical demands of occupations have a strong negative correlation with each other. Of the two, mental demands had a clearer effect on driving cost.

Higher personal income increased the chosen fixed-cost level, as expected. Income segmentation was also tried by dividing the sample into three income classes and estimating separate coefficients for all of them. The nonsegmented and segmented incomes were also transformed to see which would best reflect the true effect on income. None of the combinations was significantly better. The same procedure was executed on household income. No form was statistically significant. Household income was left out of this regression, and personal income was used unsegmented and untransformed.

Owning a summer cottage increased the chosen fixed-cost level. The possibility of using a summer cottage of a relative or a friend did not have any effect on the chosen fixed-cost level.

Model for Variable Costs of Driving a Car

The model is given in Table 6. Even though the R^2 is very low, the model has statistical significance. The variable costs are nearly the same, captured by the constant term of the regression; however, many other factors influence the variable costs in a statistically significant way.

The correction term is right-signed and significant; owning a car increases variable costs. Cycle1, marking childless households with a young head; working in the transportation, storage, or communications industry; and the number of persons in the household are all associated with higher variable costs. The only variable associated with lower than "constant" variable costs is living in a town house.

Because maintenance costs are included in variable costs, the accepted level of variable costs could be connected with the accepted level of risking the car trip. Thus, young and childless adults may find it acceptable to change mode and timetables; working in the transportation or communication industry may be connected with the ability to perform unexpected repairs, and so forth.

The number of persons in the household may work its way through the effect that accepting higher variable costs may be the only way to acquire a big enough car. On the other hand, households with small children had lower-than-average fixed driving costs. It may be that the increase in family size causes a change to bigger but older cars. Living in a town house may be a proxy for a certain way of life, or for traffic environment; at this point no other possible explanation was found.

TABLE 6 MODEL FOR VARIABLE COSTS OF DRIVING A CAR (DEPENDENT VARIABLE: VARIABLE DRIVING COSTS OF RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.018	2.064
constant	0.509	22.605
cycle1	0.051	2.553
transp industry	0.048	2.214
# of persons in hh	0.011	2.008
townhouse	-0.047	-2.613
Number of Observations	614	
R-squared	0.04	
Sum of Squared Residuals	14.946	
Standard Error of the Regression	0.157	
Mean of Dependent Variable	0.536	

Model for Total Costs of Driving a Car

Total cost, the sum of fixed and variable costs, is given in Tables 7 and 8. For easier interpretation, the union of variables in fixed-cost and variable-cost regressions was first introduced for regression. After dropping the statistically nonsignificant variables the model in Table 7 was obtained.

To determine whether there are some different factors that influence the choice of total driving cost level, all the available variables were introduced to the model and then dropped if proven insignificant. The resulting model is given in Table 8.

TABLE 7 MODEL FOR TOTAL COSTS OF DRIVING A CAR, RESTRICTED SET OF VARIABLES (DEPENDENT VARIABLE: TOTAL DRIVING COSTS OF RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.019	0.822
constant	0.954	21.047
female	-0.132	-3.142
deadline	0.092	2.520
self-employed	0.213	4.694
personal inc	0.000	3.747
summer cottage	0.079	2.645
Number of Observations	611	
R-squared	0.15	
Sum of Squared Residuals	57.693	
Standard Error of the Regression	0.309	
Mean of Dependent Variable	1.103	

TABLE 8 MODEL FOR TOTAL COSTS OF DRIVING A CAR, ALL VARIABLES POSSIBLE (DEPENDENT VARIABLE: TOTAL DRIVING COSTS OF RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.021	0.883
constant	1.076	18.116
female	-0.150	-3.529
self-employed	0.229	5.123
manufact industry	-0.071	-2.379
physic demanding	-0.001	-2.172
personal income	0.000	2.287
summer cottage	0.072	2.396
Number of Observations	604	
R-squared	0.15	
Sum of Squared Residuals	56.055	
Standard Error of the Regression	0.307	
Mean of Dependent Variable	1.103	

The variable "deadline" drops off and the new variables "working in manufacturing industry" and "physically demanding occupation" appear as significant. The coefficients of the two models are similar, which may indicate that the mentioned variables measure the same variance. Because the purpose of these regressions is to assign estimated driving costs to persons who are not drivers currently, the model with fewer variables is preferred. The first model is selected.

These regressions will be used for assessing the driving costs for persons who do not have cars available but who in principle could have one and, thus, choose a car alternative in a mode choice situation (model).

Driving Cost Models for Household Members

This section discusses relationships inside the household that may affect travel behavior. The internal dependency is analyzed through the variables OFC, OVC, and OTC. The regressions are run on the characteristics of the respondent to determine their connections with driving cost levels of the other household members.

OFC of Driving a Car

The model is presented in Table 9. It was created by the same procedure as the previous models. All the variables were introduced and their significance investigated.

The sample was segmented by the person's own income and the household's income. These segmentations did not prove to be worthwhile, and it was deemed best to keep the income variable as simple as possible. The logarithmic transformation of income did not improve the regression.

Among the occupational characteristics, monotonousness, physical demand level, and "little possibility to influence work rhythm" first showed strong negative impact in the regression. When the organizational and location variables were later introduced, the occupational variables lost their statistical significance. In the regression the organizational "blue/white collar" dummy variable effectively captured the variation, even though occupation and organization hardly measure the same things. It may be noted here that Cubukgil and Miller (*13*) found, in a Toronto study, that persons from different organizational status groups have different journey patterns.

As an aside, Cubukgil and Miller divided the blue collar worker group into two according to skill level and found the groups to be located and to behave differently. The location variables may, thus, "eat up" the effects of occupation. It is an interesting chicken-and-egg proposition to ask whether people live where their occupation leads them or choose the kind of occupation that is common in the neighborhood.

The other household members tended to have higher fixed driving costs if the respondent was a woman. This is true even when the correction for the probability of having or not having a car is taken into account. Household income has the expected sign, and the respondent's income is not significant in the model.

The only locational variable that kept its significance was the distance between home and the nearest grocery store. The negative sign of it is unexpected, implying that the further away the nearest store, the less people spend on the fixed cost of driving. The only interpretation at this moment is that the correction term really works here as it should work: it accounts for the probability of having a car (which is presumably higher in a place far away from shops). When this effect is accounted for, the net effect of a long distance to shop may indeed be connected with a lower fixed driving cost.

OVC of Driving a Car

Again, as with variable costs regression for the respondent, the R^2 is very low. The correction term is kept in the regression for theoretical reasons (see Table 10).

The significant variables are "respondent works in manufacturing industry," "house in area of high traffic volume," and "the respondent has high variable driving costs." The last variable correlates with the same phenomenon as "living in a one-family house." Either of these variables gets a statistically significant coefficient. Living in a one-family house enables maintenance work on and storing of cars that in other living arrangements would simply be sold. That the respondent has higher variable costs implies that mechanics may be a hobby or at least that there is some car repair know-how

TABLE 9 MODEL FOR OFC OF DRIVING A CAR (DEPENDENT VARIABLE: FIXED DRIVING COSTS OF THE HOUSEHOLD MEMBERS OF THE RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.021	1.396
constant	0.423	7.073
female	0.103	3.727
shift	-0.061	-1.927
blue-collar	-0.074	-3.074
hh income	0.000	2.675
distance to grocery	-0.000	-2.485
Number of Observations		473
R-squared		0.13
Sum of Squared Residual		26.843
Standard Error of the Regression		0.240
Mean of Dependent Variable		0.494

TABLE 10 MODEL FOR OVC OF DRIVING A CAR (DEPENDENT VARIABLE: VARIABLE DRIVING COSTS OF THE HOUSEHOLD MEMBERS OF THE RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.004	0.430
one	0.477	25.720
manufac ind	0.040	2.172
traffic volume	0.017	2.884
variabl driv costs	0.071	2.340
Number of Observations		474
R-squared		0.04
Sum of Squared Residuals		11.201
Standard Error of the Regression		0.155
Mean of Dependent Variable		0.539

in the household, which can be relied on in case of need. These two states appear to coincide.

OTC of Driving a Car

Finally, the total driving cost regression of the respondent's household members is presented.

In the model (Table 11), household characteristics, inner city location, and household income increase the total driving costs. This is expected, because parking and storage costs are higher in a city, and household income is the classical explanation for owning expensive cars.

Physically demanding occupation decreases the respondent's household members' chosen level of driving costs. The physically demanding occupation was also significant in the "total driving costs of the respondent" regression. This characteristic "spills over" its influence to the other members of the household, too.

The other occupational characteristic, little influence on work rhythm, was also significant in the fixed driving costs of the household members' regression. There it covaried with the shift work dummy and was replaced by it. In the total driving cost regression, however, the occupational characteristic is stronger.

CONCLUSIONS

This study started from the need to assess driving cost estimates for the mode choice model. It was approached through separate functions for fixed, variable, and total driving costs of the respondent. More insight was gained by examining within-household effects, which was done by estimating the driving cost regressions of the household members of the respondent.

It is not known which of the costs should be entered in the mode choice logit model. Should one use only the fixed, the variable, the total, or both fixed and variable costs, or perhaps a combination of total and variable costs? Every alternative has a plausible explanation for having a best fit and a reason to it.

TABLE 11 MODEL FOR OTC OF DRIVING A CAR (DEPENDENT VARIABLE: TOTAL DRIVING COSTS OF THE HOUSEHOLD MEMBERS OF THE RESPONDENT)

Independent Variable	Estimated Coefficient	t-Statistic
correction term	0.013	0.808
constant	0.974	15.185
female	0.095	3.469
phys dem work	-0.001	-2.501
little influence	-0.004	-2.462
hh income	0.000	2.904
home in city	0.085	2.379
Number of Observations		470
R-squared		0.12
Sum of Squared Residuals		26.846
Standard Error of the Regression		0.240
Mean of Dependent Variable		1.033

The within-household regressions indicate that there are within-household work life effects on driving costs. These effects take place on a detailed level, which is not captured in a variable describing the household member's employment status. On the basis of these regressions, it appears that the types of occupations of the household members have more effect on the driving costs than the number or share of the employed persons in the household as such.

It remains to be seen whether this information can be used to improve the mode choice, automobile ownership and availability, and daily tour combination models. The estimated cost models provide, however, an explanation for the poor performance of cost variables in situations where the cost is a priori regarded as "the" explanation but approximated to be the same for all drivers. The operating costs are not a flat figure, equal for everybody. They are an outcome of a choice, influenced by the characteristics and activities of the individual and those of other household members. The same applies to capital costs. Capital costs are not equal for everyone, for drivers and nondrivers, or for persons in different occupations, household situations, or stages of life.

It may be mentioned as an aside that the mode choice model developed using the concepts and models presented here provided some surprises. The coefficient of the variable (operating) costs of automobiles was near zero and had a very low statistical significance on mode choice; high capital costs were associated with automobile choice. These findings are not really surprising, but ones not believed nor a part of present travel demand model systems. But it is believable that automobile variable costs do not influence mode choice, or, that once automobile is chosen it is an expensive one—comfort costs—by choice.

Other demand models of the nested system, automobile driver status and travel diary (pattern) choice, are not yet ready. It is known from the mode choice model, however, that the daily travel pattern, travel diary, affects mode choice. That is, mode choices of daily trips of travelers are interdependent.

The main determinants of the other two travel choices in the model system are still unknown. The driving cost models indicate that all travel choices, including the cost of travel, are a part of complex behavioral decisions within a household. Because the market offers a wide variety of transportation choices in terms of costs, there is no surprise involved in finding that the chosen cost level is a part of the mode choice, that is, it belongs to the left-hand side of the equation, and, therefore, is not significant in the right-hand side of a marginal mode choice model. And conversely, assigning equal—but fundamentally arbitrary—automobile costs to all users captures, with that variable and its parameter, effects that, in fact, are behavioral functions and not parameters.

ACKNOWLEDGMENT

Research reported in this paper was conducted when the second author was employed by the Technical Research Center of Finland and received support from the center, the Helsinki Regional Council, and the Finland Highway Administration. The first author was supported, in part, by Yrjö Jahnsson Foundation.

REFERENCES

1. D. McFadden, A. Talvitie, et al. *Demand Model Estimation and Validation. Final Report Series, Vol. V.* Institute of Transportation Studies, University of California, Berkeley, 1977.
2. A. M. Malecki. Perceived and Actual Costs of Operating Cars. *Transportation*, Vol. 7, 1978, pp. 403–415.
3. A. Talvitie and Y. Dehghani. Models for Transportation Level of Service. *Transportation Research B*, Vol. 14B, 1980, pp. 87–99.
4. C. Winston. Conceptual Developments in the Economics of Transportation: An Interpretive Survey. *Journal of Economic Literature*, Vol. 22, March 1985, pp. 57–94.
5. R. Kitamura. A Panel Analysis of Household Car Ownership and Mobility. *Proc., Japan Society of Civil Engineers*, JSCE No. 383/IV-7, July 1987, pp. 13–27.
6. A.-M. Lehto. Naisten ja Miesten Työolot. Helsinki, Tilastokeskus, Tutkimuksia, 1988, 138.
7. Työvoimatutkimus 1984, Työvoimaan kuuluvuus 1985. Helsinki, Tilastokeskus, Tilastotiedotuksia TY 1985:28.
8. Työolotutkimus 1984, Työn haitat 1985. Helsinki, Tilastokeskus, Tilastotiedotuksia TY 1985:18.
9. C. A. Zimmermann. The Life-Cycle Concept as a Tool for Travel Research. *Transportation*, Vol. 11, 1982, pp. 51–69.
10. K. Train. *Qualitative Choice Analysis—Theory, Econometrics, and an Application to Automobile Demand*. MIT Press, Cambridge, Mass., 1986.
11. MVA Consultancy et al. *Value of Travel Time Savings*. Department of Transport, Great Britain. Newbury, Policy Journals, 1987.
12. R. E. Schuler and J. W. Coulter. The Effect of Socio-Economic Factors on the Value of Time in Commuting to Work. *Transportation*, Vol. 7, 1978, pp. 381–401.
13. A. Cubukgil and E. J. Miller. Occupational Status and the Journey-to-Work. *Transportation*, Vol. 11, 1982, pp. 251–276.

Publication of this paper sponsored by Committee on Traveler Behavior and Values.

Convergent Algorithm for Dynamic Traffic Assignment

BRUCE N. JANSON

A link flow formulation and a convergent solution algorithm for the dynamic user equilibrium (DUE) traffic assignment problem for road networks with multiple trip origins and destinations are presented. The link flow formulation does not implicitly assume complete enumeration of all origin-destination paths as does the equivalent path flow formulation. DUE is a temporal generalization of the static user equilibrium (SUE) assignment problem with additional constraints to ensure temporally continuous paths of flow. Whereas SUE can be solved by methods of linear combinations, these methods can create temporally discontinuous flows if applied to DUE. This convergent dynamic algorithm (CDA) uses the Frank-Wolfe method of linear combinations to find successive solutions to DUE while holding node time intervals fixed from each origin. In DUE, the full assignment period of several hours is discretized into shorter time intervals of 10 to 15 min each, for which trip departure matrices are assumed to be known. The performance of CDA is compared with that of a heuristic solution procedure called DTA. CDA can be applied to solving DUE on large networks, and the examples presented show that CDA consistently converges to solutions that closely satisfy the DUE optimality conditions. With computational advances such as parallel computing, CDA can be run in near real-time on large-scale networks and used with in-vehicle route advisory systems for traffic management during evacuations and special events.

Dynamic traffic assignment procedures are needed to evaluate the impacts of alternative travel demand management strategies during peak periods in urban areas and will play a key role in the development of real-time traffic management and in-vehicle route guidance systems. Merchant and Nemhauser (*1*) formulated a system-optimal version of the dynamic assignment problem in which there can be multiple origins but only one destination. A global optimum is difficult to obtain because of the nonconvexity of that formulation (*2,3*). Carey (*4*) shows that the Merchant and Nemhauser formulation satisfies a certain constraint qualification needed for Kuhn-Tucker optimality conditions to exist at the optimum. Carey (*5*) presents an alternative formulation of the Merchant and Nemhauser problem with only one destination that is convex and can be made piecewise linear for solution purposes.

The dynamic user equilibrium (DUE) assignment problem is defined in this paper as follows: Given a set of zone-to-zone trip tables containing the number of vehicle trips departing from and headed toward each zone in successive time intervals of 10 to 15 min each, determine the volume of vehicles on each link in each time interval of a network connecting these zones that satisfy the following two conditions:

Department of Civil Engineering, University of Colorado, Denver, Colo. 80217-3364.

1. All paths between a given pair of zones used by trips departing in a given time interval must have equal travel impedances.
2. All paths between a given pair of zones not used by trips departing in a given time interval cannot have lower travel impedances.

These two conditions for DUE given fixed trip departure times, which will be derived later from the link flow formulation of DUE presented herein, are temporal generalizations of Wardrop's conditions (*6*) for static user equilibrium (SUE). Other authors have defined DUE to include variable trip departure times, which are assumed to be fixed in this paper. Janson (*7*) formulates and presents a convergent solution algorithm for the combined problem of dynamic traffic assignment and trip distribution with variable trip departure or arrival times.

Friesz et al. (*8*) present an optimal control theory formulation of dynamic traffic assignment in continuous time for which the equilibrium conditions are a variation of the foregoing conditions. The optimality conditions of their model are that all used paths between any two nodes must have equal impedances at any given instant and unused paths cannot have lower impedances. These conditions allow complete origin-to-destination paths used by trips to have unequal impedances for any given departure time. These conditions are not user optimal, because travelers can switch to paths with lower impedances. Once trips depart on separate paths from a given origin, these paths need not have equal impedances for the remainders of their journeys.

DUE is a temporal generalization of the SUE assignment problem of which SUE is the special case with one long assignment period. In DUE, the full assignment period of several hours is discretized into shorter time intervals of 10 to 15 min each, for which trip departure matrices are assumed to be known. With multiple time intervals, DUE requires nonlinear mixed-integer constraints with "node time intervals" that ensure temporally continuous paths of flow. Although DUE is nonconvex over the domain of feasible node time intervals for trip paths from each origin, DUE is convex with a unique global optimum for any given set of fixed node time intervals. The optimality conditions of DUE are later derived from a qualified statement of DUE in which only temporally continuous paths can be used in the optimum solution according to prespecified node time intervals.

The steady-state flow assumption of SUE allows it to be formulated with all linear constraints. Whereas SUE can be solved efficiently by methods of linear combinations, these methods create temporally discontinuous flows if applied to

DUE. Janson (*9*) formulated DUE as a nonlinear mixed-integer program in terms of path flows and described a dynamic traffic assignment (DTA) heuristic that generates approximate solutions to DUE for large networks. DTA is not a convergent solution algorithm for DUE, but instead was designed to produce traffic assignments that tend to satisfy the DUE optimality conditions stated earlier. In test applications, DTA produced both static and dynamic assignments that approximately satisfied the user equilibrium conditions of those problems.

Both SUE and DUE can be formulated equivalently in terms of either path or link flows, but the link flow formulations of these problems do not implicitly assume complete enumeration of all possible paths between zone pairs. More important, the link flow formulation of DUE decomposes directly into two subproblems solved successively by the convergent dynamic algorithm (CDA) presented herein. CDA solves DUE with fixed node time intervals by the Frank-Wolfe (F-W) method of linear combinations and then updates the node time intervals with which to generate the next F-W solution. The procedure terminates when the number of node time interval changes (or other measure of convergence tolerance) is acceptable. Several examples indicate that CDA consistently converges toward optimal DUE solutions. CDA can be applied to large planning networks, and convergence difficulties that might occur on small, specially configured networks are less likely to occur on larger networks in which paths from many origins share common links.

DUE ASSIGNMENT PROBLEM

A temporal generalization of SUE in terms of path or link flows requires that a time interval superscript be added to each link flow variable and impedance function. A departure time superscript must also be added to each origin-specific link flow variable and to each element of the trip matrix Q. Superscripts representing time intervals of link use and trip departure must also be added to each node time interval variable to indicate whether trips departing from origin zone r in Time Interval d reach Node i in Time Interval t. Denoting a link as a node pair ij instead of with an arc subscript k is necessary for the derivation of optimality conditions from this link flow formulation of DUE given by Equations 1 through 10.

$$\text{DUE} \quad \text{Minimize} \sum_{ij \in A} \sum_{t \in D} \int_0^{x_{ij}^t} \mathrm{f}_{ij}^t(w)\, dw - \sum_{r \in Z} \sum_{i \in N} \sum_{d \in D} b_{ri}^d \tag{1}$$

subject to

$$x_{ij}^t = \sum_{r \in Z} \sum_{d \le t} v_{rij}^{dt}\, \alpha_{ri}^{dt} \qquad \text{for all } ij \in A,\ t \in D \tag{2}$$

$$q_{rn}^d = \sum_{t \ge d} \left(\sum_{in \in A} v_{rin}^{dt}\, \alpha_{ri}^{dt} - \sum_{nj \in A} v_{rnj}^{dt}\, \alpha_{rn}^{dt} \right)$$

$$\text{for all } r \in Z,\ n \in N,\ d \in D \tag{3}$$

$$v_{rij}^{dt} \ge 0 \qquad \text{for all } r \in Z,\ ij \in A,\ d \in D,\ t \in D \tag{4}$$

$$\alpha_{ri}^{dt} = (0, 1) \qquad \text{for all } r \in Z,\ i \in N,\ d \in D,\ t \in D \tag{5}$$

$$\sum_{t \in D} \alpha_{ri}^{dt} = 1 \qquad \text{for all } r \in Z,\ i \in N,\ d \in D \tag{6}$$

$$(b_{rj}^d - b_{ri}^d)\, \alpha_{ri}^{dt} \le f_{ij}^t(x_{ij}^t)\, \alpha_{ri}^{dt}$$

$$\text{for all } r \in Z,\ ij \in A,\ d \in D,\ t \in D \tag{7}$$

$$[b_{ri}^d - (t - d + l)\Delta t]\alpha_{ri}^{dt} \le 0$$

$$\text{for all } r \in Z,\ i \in N,\ d \in D,\ t \in D \tag{8}$$

$$[b_{ri}^d - (t - d)\Delta t]\alpha_{ri}^{dt} \ge 0$$

$$\text{for all } r \in Z,\ i \in N,\ d \in D,\ t \in D \tag{9}$$

$$b_{rr}^d = 0 \qquad \text{for all } r \in Z,\ d \in D \tag{10}$$

where

N = set of all nodes,
Z = set of all zones (i.e., trip-end nodes),
A = set of all links (directed arcs),
Δt = duration of each time interval (same for all t),
D = set of all time intervals in the full analysis period (e.g., eighteen 10-min intervals for a 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}^{dt} = number of vehicle trips departing from origin zone r in Time Interval d assigned to Link ij in Time Interval t (variable),
$f_{ij}^t(x_{ij}^t)$ = travel impedance on Link k 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 (zero for any node $n \in Z$) (fixed),
b_{ri}^d = travel time along any path used from origin zone r to Node i by trips departing in Time Interval d (variable), and
α_{ri}^{dt} = zero-one variable indicating whether trips departing from origin zone r in Time Interval d reach Node i in Time Interval t (henceforth called a "node time interval") (0 = no, 1 = yes) (variable).

This formulation of DUE assumes that a directed network $G(N, A)$ is given, where N is the set of nodes and A is the set of directed arcs or links. Zones (denoted by the set Z) are nodes at which trips originate or terminate. Equation 2 defines the total flow on Link ij in Time Interval t to be the sum of flows departing from any origin r in any time interval d that use Link ij in Time Interval t in order to formulate the objective function as given by Equation 1. It is not necessary to multiply v_{rij}^{dt} by α_{ri}^{dt} in Equation 2, because flows departing from origin r in Time Interval d will only be assigned to Link ij in Time Interval t allowed by the α_{ri}^{dt} term equal to 1 in the nodal conservation-of-flow Equation 3. 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.

The second part of the DUE objective function is not required in SUE because origin-to-node travel times are not needed to calculate steady-state link volumes. In DUE, origin-to-node travel times are used to determine the node time intervals $\{\alpha_{ri}^{dt}\}$ in Equations 5 through 10. In DUE, each node time interval α_{ri}^{dt} cannot be prespecified with a fixed value because node time intervals are affected by travel times, which are affected by link loadings. The node time intervals are endogenous variables in DUE, creating nonlinear flow conservation constraints and requiring DUE to have additional constraints (Equations 5 through 10) to ensure temporally continuous paths.

Equation 5 defines each node time interval to indicate whether trips departing from origin zone r in Time Interval d reach Node i in Time Interval t. (Strictly speaking, each node time interval value α_{ri}^{dt} can only be 0 or 1, which then indicates the node time interval index t.) Equation 6 requires that there be only one time interval t in which trips departing from a given origin r in a given time interval d can reach a node. The assumption here is that any Link ij incident from Node i can only be used (if used at all) in the time interval t in which Node i is reached by trips departing from origin r in time interval d. Equation 7 determines the origin-to-node travel times, which could be any arbitrary values that optimize the objective function if these times were not maximized by the DUE objective function. Equations 8 and 9 then use the node time interval values to identify time interval indices that are compatible with the path travel times to each node.

Each intrazonal travel time b_{rr}^{d} must be set to zero or to some other fixed value as shown by Equation 10 to prevent the maximization of origin-to-node travel times (or the minimization of negative times) in the DUE objective function from having an infinite solution. Though counterintuitive, the maximization of travel times as given by Equation 1 subject to Constraints 5 through 10 is the correct objective for determining shortest travel time paths in this formulation. The shortest path problem is often formulated to find unit link flows that minimize the use of arc lengths. The second part of Equation 1 plus Equations 5 through 10 constitute the dual of that formulation, which is to maximize zone-to-node path lengths subject to fixed arc lengths. Because many hundreds of vehicles travel each path, the first part of the objective function will dominate the assignment process, whereas the second part causes the correct shortest path travel times to be found. If vehicle units are small, the second part of the objective function can be multiplied by any small constant and it will still achieve its desired result.

According to Equations 7 through 9, links are traversed within the time intervals that trip paths reach their tail nodes. For 10-min intervals, Interval 1 begins at 0 min, Interval 2 begins at 10 min, Interval 3 begins at 20 min, and so forth. If a path reaches a node at the exact beginning of a time interval (to the degree of floating point precision being used), the solution algorithm can be coded to have the path use the link in that time interval rather than in the previous interval. Note that Equations 7 through 9 also work for the static case, albeit unnecessary, so long as the duration Δt of the single time interval exceeds the longest trip length in the network (measured in time). Equations 1 through 10 exactly define SUE if there is only one long time interval and all node time intervals given by array $\{\alpha_{ri}\}$ (disregarding time) are set to 1. Thus, Equations 1 through 10 are a complete formulation of both the static and dynamic assignment problems, with the static problem being a special case.

Figure 1 shows how a node time interval depends on the time at which a path reaches the tail node of a link. The link number is indicated by k, and the time interval of link use is indicated by t. Link 1 is used by this path in Interval 1, although it overlaps into Interval 2. Links 2 and 3 are used in Interval 2, but Link 4 is used in Interval 3, because it begins exactly at the beginning of Interval 3. The portion of Link 1 that overlaps into Interval 2 is discussed next.

If link lengths are generally much shorter than the time interval duration (e.g., less than 20 percent), fractions of paths overlapping time intervals may not be significant. Suppose that a dynamic assignment has a mean link length of M min and a time interval duration of N min. The probability is M/N that a link of M min will overlap time intervals in any path that uses it. When overlap does occur, the average amount of overlap into the next time interval is one-half the average link length, or $M/2$ min. Thus, the average amount that used links overlap time intervals is $M^2/(2N)$, which is one-half the average link length ($M/2$) times the probability that an overlap occurs (M/N). For example, links of 2 min in time intervals of 10 min will, on the average, overlap intervals by 0.2 min, or 10 percent of the link length. The overlapping percentage of total trip length, which accounts for trip volumes on the links, was computed for the examples given later and was always found to fall below this formula's estimate.

Whereas the impedance functions in the SUE objective function are not restricted to travel time alone, inconsistencies could develop between path travel times and node time intervals in DUE if the impedance functions are not strictly measures of time. Equal paths according to a composite cost could reach nodes in different time intervals. The path flow formulation of DUE given by Janson (*9*) avoids this complication by explicitly computing the travel times along every possible O-D path. In the path flow formulation, the impedance functions in Equation 1 can include travel cost factors other than travel time that affect route choice. The path flow formulation only gains that advantage by requiring complete path enumeration.

The derivation of user equilibrium optimality conditions from the preceding formulation is complicated by the nonlinear mixed-integer constraints and integer node time intervals. Although optimality conditions cannot be derived from the general formulation with integer unknowns, they can be derived for a given set of node time intervals to which all temporally continuous paths in the optimal solution to the general problem must conform. Because node time intervals can be uniquely determined from a given set of link volumes, they

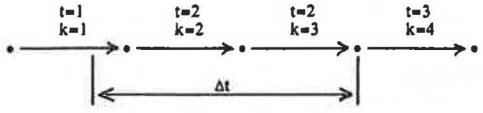

FIGURE 1 Example of time interval overlap at node.

can be assumed to be known in the derivation of optimality conditions for the global optimum.

Equation 11 is the Lagrangian of Equations 1 through 4 with linear constraints and only the first part of the objective function because of fixed node time intervals. Over the domain of variable integer values, Equation 11 is nonconvex, and there are many local optima that are inferior to the global optimum. For a set of fixed node time intervals, the bordered Hessian matrix of Equation 11 is positive definite, which means that there is a unique global optimum with no local optima (*10*). The Hessian matrix 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, which is assumed to be true in the foregoing formulation of DUE.

$$L(X, V, \lambda, \mu, \tau) = \sum_{ij \in A} \sum_{t \in D} \int_0^{x_{ij}^t} f_{ij}^t(w)dw$$

$$- \sum_{ij \in A} \sum_{t \in D} \lambda_{ij}^t \left(x_{ij}^t - \sum_{r \in Z} \sum_{d \le t} v_{rij}^{dt} \alpha_{ri}^{dt} \right)$$

$$+ \sum_{r \in Z} \sum_{n \in N} \sum_{d \in D} \mu_{rn}^d \left[q_{rn}^d - \sum_{t \ge d} \left(\sum_{in \in A} v_{rin}^{dt} \alpha_{ri}^{dt} - \sum_{nj \in A} v_{rnj}^{dt} \alpha_{rn}^{dt} \right) \right]$$

$$+ \sum_{r \in Z} \sum_{ij \in A} \sum_{d \in D} \sum_{t \in D} \tau_{rij}^{dt} (-v_{rij}^{dt}) \tag{11}$$

The optimality conditions are given by Equations 12 through 14.

$$\partial L / \partial x_{ij}^t \rightarrow f_{ij}^t(x_{ij}^t) = \lambda_{ij}^t \qquad \text{for all } ij \in A,\ t \in D \tag{12}$$

$$\partial L / \partial v_{rij}^{dt} \rightarrow (\mu_{rj}^d - \mu_{ri}^d)\alpha_{ri}^{dt} = \lambda_{ij}^t \alpha_{ri}^{dt} - \tau_{rij}^{dt}$$

$$\text{for all } r \in Z,\ ij \in A,\ d \in D,\ t \in D \tag{13}$$

$$\tau_{rij}^{dt} v_{rij}^{dt} = 0 \qquad (\tau_{rij}^{dt} \ge 0)$$

$$\text{for all } r \in Z,\ ij \in A,\ d \in D,\ t \in D \tag{14}$$

where $\tau_{rij}^{dt} = 0$ if $v_{rij}^{dt} > 0$, positive otherwise; impedance difference from Node i to Node j via used path versus by Link ij if used or unused in Time Interval t for trips departing from Zone r in Time Interval d.

The last part of Equation 11 ensures nonnegative link flows and results in a third optimality condition given by Equation 14, which requires τ_{rij}^{dt} to be zero if any trips departing from origin zone r in Time Interval d are assigned to Link ij in Time Interval t, and nonnegative otherwise. According to Equation 12, the optimal solution has a unique equilibrium impedance for each link in each time interval. According to Equations 13 and 14, for any given pair of nodes, all used paths from a given origin for a given departure time have the same travel impedance, and any unused path between these nodes cannot have a lower impedance.

The optimality conditions for DUE can be stated similarly to Wardrop's statement of necessary conditions for SUE (*6*). For trips from Zone r departing in Time Interval d, let μ_{ri}^d be the equilibrium travel impedance of used paths to Node i. Also, let λ_{ij}^t be the equilibrium travel impedance of Link ij equal to the right side of Equation 13, and v_{rij}^{dt} be the equilibrium flow on Link ij in Time Interval t of trips from Zone r that depart in Time Interval d. At equilibrium, all paths from Zone r to Node j used by trips departing in Time Interval d have impedance μ_{rj}^d, and no unused path from Zone r to Node n for this departure time can have a lower impedance. These two conditions are given by Equations 15 and 16.

$$\mu_{ri}^d + \lambda_{ij}^t = \mu_{rj}^d \qquad \text{if } v_{rij}^{dt} > 0 \text{ and } \tau_{rij}^{dt} = 0$$

$$\text{for all } r \in Z,\ ij \in A,\ d \in D,\ t \in D \tag{15}$$

$$\mu_{ri}^d + \lambda_{ij}^t \ge \mu_{rj}^d \qquad \text{if } v_{rij}^{dt} = 0 \text{ and } \tau_{rij}^{dt} \ge 0$$

$$\text{for all } r \in Z,\ ij \in A,\ d \in D,\ t \in D \tag{16}$$

Equations 15 and 16 and equivalent to SUE conditions if there is only one time interval for the full analysis period so that the time interval superscripts can be removed from all terms. When SUE is solved by a method of linear combinations, one measure of how close a solution is to equilibrium is the sum of trip impedance differences from shortest path impedances between all zone pairs in each iteration. This measure of convergence is often referred to as the duality gap or impedance gap. In solving DUE, this gap is the sum of trip impedance differences from shortest path impedances between all zone pairs for trips departing in each time interval, which is one way in which the example results presented later are evaluated.

EXAMPLE OF TEMPORALLY DISCONTINUOUS TRIP PATHS

Methods of linear combinations (e.g., F-W and PARTAN), which apply to nonlinear programs with all linear constraints, are used to solve SUE by combining link volumes without regard to when they occur because they are assumed to occur continuously. The difficulty of ensuring temporally continuous flows arises because the proper time interval superscript at each node n is unknown with respect to flows departing from origin r in different time intervals. Although conservation of flow at each node and zone is stated in terms of link flows, Equation 3 would not ensure temporally continuous flows without Equations 5 through 10 defining the values of the integer node time intervals. The summation over intervals $t \ge d$ in Equation 3 only prevents trips from using links earlier than their trip departure times.

In Figure 2, Link k' incident to Node n in Time Interval t may have a short travel time so that its flow also passes onto Link k in Time Interval t. However, the travel time of Link k' or links before it may grow longer because of congestion such that its flow passes onto Link k in Time Interval $t + 1$. As a result, flows that satisfy Constraint Equation 3, but not Equations 7 through 9, can skip over time intervals at any node or even enter nodes later than they exit.

A simple example is given to demonstrate the difference in solving DUE with and without Equations 5 through 10. Figure 2 shows a path from Zone r to Zone s taken by 3,000 vehicles departing in Time Interval 1 between 7:00 and 7:10 a.m., and no trips depart in the only other time interval from 7:10 to 7:20 a.m. Using the BPR impedance function with free-flow travel times and capacities shown for the links, and without Equations 5 through 10 to ensure temporally contin-

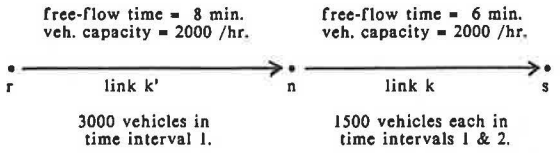

FIGURE 2 Example of temporally discontinuous flows.

uous flows, the optimal solution is to subdivide the flow at Node n into two parts of 1,500 vehicles each. One part is assigned to Link k in Time Interval 1, and the other part is assigned to Link k in Time Interval 2. The F-W method will assign these invalid flows, and once created in the solution, they are never eliminated as the algorithm proceeds. Other examples with two or more routes also show that methods of linear combinations create temporally discontinuous flows when used to solve DUE.

CDA PROCEDURE

In the CDA procedure, DUE is decomposed into two subproblems. The preceding derivation of DUE optimality conditions from Equations 1 through 4 with fixed node time intervals corresponds to the first subproblem, P1, in which the first part of the DUE objective function and Equations 1 through 4 are solved with a fixed set of node time intervals using the standard F-W algorithm or another method of linear combinations. The second part of the DUE objective function and Equations 5 through 10 are then solved in Subproblem P2 to obtain a new set of node time intervals that form temporally continuous shortest paths given the link travel times obtained from solving P1. The iterative process of solving P1 and then P2 terminates when the convergence criterion is satisfied, such as when changes in assigned link flows or node time intervals are acceptably low. Cycling may occur in the node time interval values between successive solutions of the F-W algorithm, but such cycling can be detected and avoided if it occurs. Moreover, such cycling is unlikely to prevent reasonable convergence in larger networks where trips from many origins share common links. An acceptable degree of convergence was obtained in each of the examples given later.

To clarify the explanation of CDA, Subproblems P1 and P2 are formulated below as Equations 17 and 2 through 4, and as Equations 18 and 5 through 10, respectively.

$$\text{P1} \qquad \text{Minimize} \sum_{ij \in A} \sum_{t \in D} \int_0^{x_{ij}^t} f_{ij}^t(w)\, dw \tag{17}$$

subject to Equations 2 through 4, where all x_{ij}^t are variable and all α_{ri}^{dt} are fixed in solving P1, and all other values are fixed or variable as they were defined in Equations 1 through 10.

$$\text{P2} \qquad \text{Maximize} \sum_{r \in Z} \sum_{i \in N} \sum_{d \in D} b_{ri}^d \tag{18}$$

subject to Equations 5 through 10, where all x_{ij}^t are fixed and all α_{ri}^{dt} are variable in solving P2, and all other values are fixed or variable as they were defined in Equations 1 through 10.

With fixed node time intervals, Subproblem P1 is solved without fixing which links are used but only fixing the time intervals in which links are used by trips depending on their origins and departure times. Subproblem P2 is solved with a label-setting or label-correcting shortest path algorithm adapted for temporally dependent arc lengths. Both types of shortest path algorithms will correctly find temporally continuous shortest paths given dynamic arc lengths with the restriction that vehicles do not pass each other along any link. An equivalent assumption when dealing with aggregate vehicle flows is that vehicles make only one-for-one (or zero-sum) exchanges of places in traffic along any link. This assumption is acceptable and even expected in aggregate traffic models.

CDA converges toward a DUE solution for the following reasons. First, if node time intervals corresponding to a true equilibrium are known, the F-W algorithm will reproduce the equilibrium link volumes from which these node time intervals can be calculated. That convergence proof is equivalent to the convergence proof of the F-W algorithm for SUE. Second, given node time intervals that do not correspond to a true dynamic equilibrium, the F-W algorithm will produce link volumes that shift the node time intervals toward their correct values. For example, if a particular node time interval is too early, then paths to that link will be assigned more traffic by the F-W algorithm such that the node time interval becomes later when recalculated. Oppositely, if a node time interval is too late, paths to that link will be assigned less traffic by the F-W algorithm such that the node time interval becomes earlier when recalculated. Hence, CDA converges toward a set of node time intervals that, when used to assign trips to the network, result in temporal link volumes that give rise to the same node time intervals. A similar case is the convergence proof given by Evans for combined distribution and assignment (*11*) in which a trip distribution is found that when assigned to the network reproduces O-D impedances that give rise to the same distribution.

DTA PROCEDURE

Janson (*9*) developed and evaluated the performance of a heuristic solution procedure for DTA that requires much less memory and computational effort than the just-mentioned approach of successively solving constrained specifications of DUE to which methods of linear combinations can be applied. In example applications, DTA was shown to produce good approximate solutions to DUE for both static and dynamic travel demands. Thus, DTA can serve as an effective means of providing an initial starting solution to the convergent algorithm. The remainder of this section reviews the DTA procedure.

A key assumption of the DTA procedure described here is that route choice decisions are made at the time of trip departure on the basis of projected link impedances that account for changes in travel demand over future time intervals. This assumption offers a great deal of computational efficiency, because the trip departure matrix assigned in each time interval is only a zone-to-zone trip matrix, not a node-to-zone trip matrix as would be required to track trips at nodes through the network by their destinations so that trip paths could be revised en route.

The way in which current link volumes are projected into future time intervals for the purpose of finding trip assignment paths in DTA requires a brief preview. The reason for projecting current link volumes into future time intervals is that some trips departing in later intervals from other origins will use the links concurrently with trips from the present origin for which paths are being found. A premise of DTA is that future link volumes can be adequately estimated from current link volumes on the basis of relative levels of travel demand in future time intervals. These projections are made only for path-finding purposes and are not factored into the actual assignment of link volumes.

The following notation is used to describe the DTA procedure:

x_k^t = assigned volume on Link k in the current Time Interval t after trips departing in Time Intervals 1 through $t - 1$ have been assigned, and while trips departing in Time Interval t are being assigned;

x_k^{t-1} = assigned volume on Link k in Time Interval $t - 1$ (i.e., just before the current time interval) after all trips departing in Time Intervals 1 through $t - 1$ have been assigned;

y_k^{t+n} = projected volume on Link k in interval $t + n$ (where $n \geq 0$ and $t + n \in D$) after trips departing in Time Intervals 1 through $t - 1$ have been assigned and while trips departing in Interval t are being assigned;

$f_k^{t+n}(y_k^{t+n})$ = projected impedance of Link k in the current or future Time Interval $t + n$ computed directly as a function of the projected volume y_k^{t+n}, where $n \geq 0$ and $t + n \in D$; and

Q^t = total number of trips departing from all zones in Time Interval t (i.e., total inflow to the network in Time Interval t).

The superscript of each link volume term always indicates the time interval of link use, whereas the superscript of each trip matrix term always indicates the time interval of trip departure. The DTA procedure assumes that all trip departures are known and fixed for all time intervals and O-D pairs. However, only the trip departure matrix of the current time interval must be stored in random access memory during each iteration of the DTA procedure. The other trip departure matrices can reside in permanent memory until needed.

The steps of the DTA procedure are as follows:

1. Read network data (link-node incidences, free-flow impedances, and practical capacities adjusted to the time interval duration Δt for the link impedance functions). Initialize the link volumes in each time interval to 0, or read in a set of starting volumes if known. Specify as NTREES the number of shortest path trees to be assigned trips from each origin zone in each time interval. Initialize the current time interval t to 0.

2. Increment the current time interval counter t to $t = t + 1$. Read in matrix of trip departures between each O-D pair in Time Interval t.

3. Randomly select an origin zone for which all trips departing in the current time interval t have not yet been assigned. Find a shortest path tree from this origin zone to all other zones on the basis of the projected link impedances in the current and future time intervals. The path search routine finds shortest paths based on link impedances as they are projected to exist during the time intervals in which they are traversed. To minimize array space allocation, each projected link volume and link impedance can be calculated as it is needed in the shortest path routine. The projected link volumes are calculated according to Equation 19, where θ^t is the percentage of Q^t that has not yet been assigned to the network. Define $\omega_{t-1}^{t+n} = Q^{t+n}/Q^{t-1}$ as a measure of systemwide travel demand and projected traffic volumes in interval $t + n$ relative to $t - 1$, and similarly for $t + n$ versus t. The projected volume on Link k in Time Interval $t + n$ is equal to

$$y_k^{t+n} = \theta^t \omega_{t-1}^{t+n} x_k^{t-1} + (1 - \theta^t) \omega_t^{t+n} x_k^t$$

$$\text{for all } k \in A,\ n \geq 0,\ t + n \in D \tag{19}$$

Hence, for each link, the current and projected link volumes are estimated as weighted combinations of the final volume assigned to that link in the previous interval $t - 1$ and the volume assigned thus far in the current interval t, weighted by ratios of total trip departures from all origins in intervals $t - 1$, t, and $t + n$. Each projected link impedance is computed directly from its projected volume using its impedance function.

4. Assign 1/NTREES of the trips departing in Interval t from the current origin to the shortest path tree found in Step 3. Store the assigned link volumes x_k^{t+n} by time of link use for all $n \geq 0$ and $t + n \in D$. If all trips departing from all zones in Time Interval t have been assigned, go to Step 5. Otherwise, return to Step 3 to process the next origin zone.

5. If all departure time intervals in the analysis period have been processed, STOP the program. Otherwise, write out the assigned link volumes for the current time interval t to a disk file and copy the link volumes of each future time interval into the link volume array space of the preceding interval to economize on array space. Return to Step 2.

Trips are assigned from origins chosen in a geographically random order to randomize the order of link loadings. A strategy that reduces random variability in the link volumes from one time interval to the next (as opposed to variations between intervals caused by changes in travel demand) is to find NTREES shortest path trees from each origin in each time interval and to assign 1/NTREES of the trip departures from each origin to each tree. In each time interval, Steps 3 and 4 are repeated NTREES times for all origins chosen randomly without replacement until all trips from all origins in that time interval have been assigned in random order.

The DTA procedure was tested on two networks described later with NTREES values of 2, 3, and 4. As expected, the DTA assignments always improved in terms of satisfying the desired user equilibrium conditions as NTREES was increased. For these networks, loading three trees from each origin in each time interval produced significantly better results than loading only two. However, increasing NTREES from 3 to 4 achieved a much smaller improvement, which indicates a decreasing marginal rate of improvement for the additional burden of finding more trees from each origin for each departure interval. All of the DTA assignments presented later were generated with NTREES equal to 3.

In assigning trips from each origin zone during the tree-by-tree assignment process in DTA, it is unknown how current and future link volumes will be affected by trips assigned from other zones. After each incremental assignment of trips from a given origin in the current time interval, projections must be made of currently assigned link volumes into future time intervals. The technique used in DTA to project current link volumes into future time intervals is to use a weighted combination of current link volumes assigned thus far and final link volumes from the previous interval. This combination is weighted 100 percent toward the just-previous link volumes when assigning the first fraction of trips from the first randomly selected origin and changes to 100 percent of the current link volumes when assigning the last fraction of trips from the last origin. In between, the weight given to current link volumes equals the percentage of total trip departures that have been assigned thus far.

Janson (*9*) evaluated the performance of DTA in several test applications, including comparisons with steady-state equilibrium assignment. When applied to the Pittsburgh network used by Janson et al. (*12*) in a validation of equilibrium assignment, DTA was found to generate link volumes and travel times that compared favorably with both equilibrium assignment and observed link counts and travel times. DTA can also be used to generate a good starting solution for CDA.

COMPARISON OF DTA AND CDA PERFORMANCE RESULTS

DTA and CDA results for steady-state travel demands are first compared with equilibrium assignment for the well-known Sioux Falls network having 76 one-way links, 24 nodes, and 24 origin-destination zones (*13*). Equilibrium assignment results for this network using the standard F-W algorithm have been described by numerous authors, including Fukushima (*14*), LeBlanc et al. (*15*), and Rose et al. (*16*). The standard F-W algorithm was also used to generate the equilibrium assignments with which CDA results are compared in this paper.

Even for cases of steady-state travel demands, the incremental loading of trips causes DTA link volumes to vary from one time interval to the next, and unless perfectly converged, CDA link volumes will also exhibit some variation between time intervals. Thus, for static assignments, it is important (*a*) to examine the degree to which DTA and CDA link volumes vary between time intervals and (*b*) to compare DTA and CDA link volumes over the assignment period with final F-W link volumes. Two measures of variation in DTA or CDA link volumes (called APV1 and APV2) are defined for this purpose. The degree to which DTA or CDA link volumes vary between time intervals is measured by their average percent variation (APV1) from their means as given by Equation 20.

$$\text{APV1} = (100/TX) \sum_{k \in A} \sum_{t \in D} |x_k^t - m_k| \tag{20}$$

where

APV1 = the average percent variation between time intervals of DTA or CDA link volumes from their means,
x_k^t = DTA or CDA volume on Link k in Time Interval t,
m_k = mean DTA or CDA volume on Link k over the time intervals in D, and
TX = total DTA or CDA link volumes in all time intervals = $\sum_{k \in A} \sum_{t \in D} x_k^t$.

A second measure of DTA or CDA link volume variation (APV2) is used to evaluate the disparity between DTA or CDA and F-W link volumes. APV2 is the average percent variation or absolute difference between DTA or CDA link volumes and final F-W link volumes. APV2 is computed in the same way as APV1, except that (*a*) m_k in APV1 is replaced in APV2 by the final F-W volume on Link k, where each F-W link volume is divided by the number of time intervals in D to represent the same time units as m_k and (*b*) TX in APV1 is replaced in APV2 by the sum of F-W link volumes for the full assignment period.

APV2 is expected to exceed APV1, because APV2 indicates variation from values not derived from the DTA or CDA link volumes. APV2 is affected by both the between-interval variation of the DTA or CDA link volumes and the differences between DTA or CDA mean volumes and F-W volumes. Both APV1 and APV2 are reported for each DTA or CDA static assignment.

How well these assignments satisfy the desired SUE or DUE conditions must also be assessed. One measure of user equilibrium is the size of the SUE or DUE objective function, which is Equation 1 summed over one time interval in the static case. Equation 1 is applied to final F-W link volumes, whereas Equation 1 is computed with the link volumes assigned by DTA or CDA in each time interval. For static assignments, Equation 1 could be applied to average DTA or CDA link volumes over all time intervals, but that would produce a falsely lower objective function and make the DTA or CDA results appear too favorable.

Another standard measure of user equilibrium for an assignment is the duality gap (DG), which is the difference between the sum of assigned trip impedances and the sum of shortest path trip impedances based on the assigned link loadings (*16*). The time dimension of DG, defined by Equation 21 for a dynamic assignment, can be disregarded in computing this measure for a static assignment with only one time period.

$$\text{DG} = (100/TC) \left[\sum_{k \in A} \sum_{t \in D} x_k^t f_k^t (x_k^t) \right] - \left[\sum_{r \in Z} \sum_{s \in Z} \sum_{d \in D} q_{rs}^d c_{rs}^d \right] \tag{21}$$

where

DG = the duality gap of a dynamic assignment,
q_{rs}^d = number of trips from Zone r to Zone s departing in Time Interval d via any path,
c_{rs}^d = shortest path impedance from Zone r to Zone s for trips departing in Time Interval d through a network of assigned link loadings, and
TC = total trip impedance if all trips were to use their shortest paths through a network of assigned link loadings = $\sum_{r \in Z} \sum_{s \in Z} \sum_{d \in D} q_{rs}^d c_{rs}^d$.

The left-most bracketed term of Equation 21 is the system-optimal objective function of DUE or SUE. The right-most bracketed term, which equals *TC*, is a strict lower bound on the optimal value of the DUE or SUE objective function for a given feasible solution with no temporally discontinuous paths. The duality gap decreases toward zero, although not strictly monotonically, as the F-W algorithm converges. The duality gap equals zero for a true equilibrium solution in which the impedance of every used path between each pair of zones equals the shortest path impedance. *TC* is not a strict lower bound on the optimal value of the DUE objective function if it is based on an infeasible solution with temporally discontinuous paths. The percentage of temporally discontinuous node time intervals (called temporal node violations) is reported later for each final solution as one qualification of this lower bound.

Although the convergence rate of the standard F-W algorithm can be improved, these improvements would not affect the comparisons in this paper because the F-W algorithm was run to a high degree of convergence in each case. When applied to the Sioux Falls network to generate assignment results for this paper, the F-W algorithm was halted when the greatest single link volume change was less than 1 percent between iterations. This degree of convergence for the Sioux Falls network required 76 iterations of the standard F-W algorithm starting from free-flow impedances. Comparisons of F-W, DTA, and CDA results can also depend on the initial link volumes or impedances used in the F-W algorithm. The outcomes reported in this paper would not be significantly affected by different F-W starting solutions because of the high degree of F-W convergence required in each case. Rose et al. (*16*) found that final link volumes for the Sioux Falls network had less than a 0.5 percent coefficient of variation between solutions when they applied the 1 percent link volume change stopping criterion to the F-W algorithm with different starting solutions.

An important consideration in developing test problems for DTA and CDA is the time interval duration, or the number of time intervals in the analysis period. An interval duration of 10 min was used in all of the following test cases. This duration was chosen after observing that the mean F-W link impedance for the Sioux Falls network was 6 min. DTA and CDA link volumes show less variation between intervals when the time interval duration is at least four to five multiples of the mean link impedance. The time interval duration would have to be 24 to 30 min to achieve this multiple for the Sioux Falls network. However, most transportation planning networks used in practice generally have shorter links, such as the Pittsburgh network used later in which the mean F-W link impedance is only 0.6 min.

To obtain initial link loadings on the network prior to the first time interval of the analysis period, both DTA and CDA were run to include ½ hr (or three 10-min intervals) of average travel demand in direct proportion to the full trip matrix, but at a lower departure rate than for the peak-hour intervals. Because the majority of trips in both the Sioux Falls and Pittsburgh networks have trip impedances within ½ hr, both DTA and CDA needed roughly ½ hr of initial trip departures for initial link loadings to be obtained. Likewise, each DTA or CDA assignment assigned six time intervals of average travel demand after the six peak-hour time intervals to allow the peak-hour trips to clear the network.

To obtain steady-state assignments from DTA or CDA, a uniform fraction of the 1-hr trip matrix was assumed to depart in each time interval of the analysis period. In DTA, each interval of trip departures was assigned incrementally to three shortest path trees (corresponding to an NTREES value of 3) by executing successive tree-by-tree assignments in each interval and loading one-third of each zone's origins to each tree. Hence, DTA could assign no more than three different paths between each pair of zones in each time interval. For this reason, fewer paths are assigned trips by DTA than by either the F-W or CDA algorithms.

Table 1 presents a comparison of the F-W, DTA, and CDA static assignments for the Sioux Falls network. The total 1-hr link volumes of the F-W, DTA, and CDA assignments were 969, 960, and 972, respectively, which are essentially equal. Trips departing in the six time intervals of the 1-hr analysis period have mean trip impedances of 14.96, 15.84, and 14.91 min for the F-W, DTA, and CDA assignments, respectively. The objective function values of the F-W, DTA, and CDA assignments, which are directly comparable, are 50.05, 50.74, and 50.22, respectively.

The F-W and CDA objective function values are nearly equal, and the F-W algorithm required 23 iterations to achieve a lower objective function value than the DTA procedure. However, the F-W algorithm requires less computational effort than either DTA or CDA. Using the Sioux Falls network and executing all procedures on the same 80486-based microcomputer with all output suppressed, DTA required 12 sec to perform 15 time intervals of dynamic assignment, whereas the F-W algorithm required only 5 sec to execute 23 iterations of equilibrium assignment. CDA required 40 sec to converge to less than a 3 percent maximum link volume change between iterations, which required 57 iterations. Thus, DTA and CDA required 2.4 times and 8 times as much computational effort, respectively, as the F-W algorithm to generate a comparable 1-hr static assignment for the Sioux Falls network. All programs were coded by the author in FORTRAN using many of the same subroutines and compiled with the same compiler.

In order to assess the degree of steady-state equilibrium obtained by DTA and CDA, Table 1 gives the values of APV1, APV2, and DG as defined earlier. The average percent link volume variations between time intervals of the DTA and CDA assignments from their own link volumes means

TABLE 1 SUMMARY OF SIOUX FALLS STATIC ASSIGNMENT RESULTS

Evaluation Measure	F-W	DTA	CDA
Total Link Volume (TX)	969	960	972
Mean Trip Impedance (min)	14.96	15.84	14.91
SUE or DUE Objective Func.	50.05	50.74	50.22
Volume Variation 1 (APV1)	----	3.00%	2.34%
Volume Variation 2 (APV2)	----	5.88%	2.84%
Duality Gap (DG)	----	1.51%	0.426%
Temporal Node Violations	----	7.3%	7.6%
80486 Computation Time	5 sec	12 sec	40 sec

(APV1) were 3.00 and 2.34 percent, respectively. The average percent link volume variations of the DTA and CDA assignments from the final F-W link volumes were 5.88 and 2.85 percent, respectively. The duality gaps of the DTA and CDA assignments are 1.51 and 0.426 percent, respectively. The duality gap of the F-W assignment from its own final shortest path impedances after 76 iterations was 0.2 percent. Temporal node violations are the number of times that shortest paths traverse nodes in incorrect time intervals in the final solution (as a percentage of total node time intervals). DTA had 7.3 percent violations, whereas CDA had 7.6 percent.

Overall, both DTA and CDA produced static assignments for the Sioux Falls network that compared favorably with F-W results. A network of the Pittsburgh eastern travel corridor was used next to examine the performances of DTA and CDA with both static and dynamic travel demands. This network contains 807 one-way links, 372 nodes, and 30 origin-destination zones. Both DTA and CDA were applied to the Pittsburgh network using 10-min intervals, with three initial intervals, six analysis period intervals, and six network-clearing intervals. Both F-W and CDA were run until no single link volume varied by more than 3 percent between iterations. F-W required 33 iterations, but CDA required only 16. However, CDA again used eight times as much CPU time as the F-W algorithm, whereas DTA used only two times as much.

Table 2 compares the F-W, DTA, and CDA static assignment results for the Pittsburgh network. Total 1-hr link volumes in the F-W, DTA, and CDA assignments were 411,178, 408,482, and 410,355, respectively. The objective function values and the mean trip impedances of the three assignments are nearly equal. Although the mean DTA and CDA trip impedances are not much greater than one time interval, many of the routes assigned by DTA and CDA were between 40 and 50 min in length. Table 2 also gives the link volume variations and duality gap measures for the DTA and CDA assignments. The link volume variations are slightly greater in this case than for the Sioux Falls network, whereas the duality gaps are much lower. The duality gap of the F-W assignment from its own final shortest path impedances after 33 iterations was less than 0.1 percent. As far as temporal node violations in the final solutions, DTA had 4.3 percent violations, whereas CDA had only 1.9 percent.

As explained in the previous section, link volumes from an earlier run of DTA can be used as the projected link volumes in future time intervals instead of projecting link volumes on the basis of the current assignment. If the network's structure or travel demands have been altered, link volumes from a previous run may not be usable, depending on the extent of the changes. However, it is instructive to see whether the duality gap of dynamic assignment route impedance differences can be improved by using link volumes from a previous run as prior information. DTA was applied 10 more times to the Sioux Falls static assignment problem, with the link volumes from each run used in the next. Values of DG for these next 10 runs were 1.778, 1.732, 1.478, 1.632, 1.545, 1.419, 1.660, 1.419, 1.718, and 1.502 percent, respectively, with a mean of 1.588 percent. Compared with a value of 1.511 percent for the initial run of DTA described previously for which there were no previous link volumes, using link volumes from previous runs did not significantly affect DG for the Sioux Falls static assignment. This insignificant effect was to be expected for a static assignment, because previous link volumes add no information to the projection formula given by Equation 19 if travel demands are constant.

TABLE 2 SUMMARY OF PITTSBURGH STATIC ASSIGNMENT RESULTS

Evaluation Measure	F-W	DTA	CDA
Total Link Volume (TX)	411178	408482	410355
Mean Trip Impedance (min)	11.25	11.23	10.91
SUE or DUE Objective Func.	3884	3873	3885
Volume Variation 1 (APV1)	----	3.84%	0.473%
Volume Variation 2 (APV2)	----	7.28%	3.61%
Duality Gap (DG)	----	0.354%	0.126%
Temporal Node Violations	----	4.3%	1.9%
80486 Computation Time	25 sec	50 sec	200 sec

To test whether previous link volumes improved the duality gap of the DTA assignment with time-varying travel demands, a Sioux Falls assignment was run for six 10-min intervals using trip departure percentages of 12.5, 16.5, 21, 21, 16.5, and 12.5. These percentages were estimated for the Pittsburgh network examples as explained later and used here for example purposes. DTA was applied 10 more times to the Sioux Falls dynamic assignment, with the final link volumes from each run used in the next. Values of DG for these next 10 runs were 3.666, 3.089, 2.722, 2.461, 2.767, 2.214, 2.226, 2.181, 2.564, and 2.652 percent, respectively, with a mean of 2.654 percent. Compared with a value of 3.425 percent for the initial DTA run for which there were no previous link volumes, using link volumes from previous runs improved the duality gap for the Sioux Falls dynamic assignment. However, a low DG value had already been achieved by DTA in the initial run.

The DTA and CDA procedures were also applied to the Sioux Falls and Pittsburgh networks using 6-min time intervals. As expected, both DTA and CDA did not perform as well with 6-min intervals on the Sioux Falls network because the average F-W link length was also 6 min, and both procedures produce better assignments when the time interval duration is at least four to five times greater than the average link length. DTA and CDA produced similar static assignments with both 6-min and 10-min intervals on the Pittsburgh network in which the average F-W link length was only 0.6 min. Thus, these few tests may indicate that DTA and CDA results are not greatly affected by moderate changes in the time interval duration so long as it remains several magnitudes greater than the average link length.

Next, DTA and CDA were applied to the Pittsburgh network with dynamic travel demands over the peak hour. Instead of assigning one-sixth of the peak-hour trip matrix to the network every 10 min, the matrix was assigned in six successive trip departure percentages equal to 12.5, 16.5, 21.0, 21.0, 16.5, and 12.5 percent of the trip matrix. These trip departure percentages are based on travel data collected for a study of highway reconstruction impacts in the Pittsburgh eastern corridor (*17*) and compared with percentages re-

ported by Hendrickson and Plank (*18*) for work trips commuting to Pittsburgh from the south. These percentages are used in the following example, but they may not be representative of the entire study region or any particular zone within it, because they are based on limited data.

Total 1-hr link volumes assigned in the dynamic case by DTA and CDA to the Pittsburgh network were 404,845 and 401,854, respectively, which are slightly below the total 1-hr link volume assigned by any of the procedures in the static case. However, the mean travel impedances of the DTA and CDA dynamic assignments are 12.3 and 12.1 min, respectively, which are 1 min greater than the mean travel impedances of the static assignments due to nonlinear link impedance functions causing greater than linear increases in trip impedances with increasing travel demands. The objective function values are 3,922 and 3,724 for the DTA and CDA dynamic assignments, respectively.

Additional comparisons between the Pittsburgh dynamic and static assignments are limited to a few of the evaluation measures. Neither of the two link volume variations (APV1 and APV2) is meaningful with time-varying travel demands. The duality gaps of the DTA and CDA dynamic assignments are 0.483 and 0.138 percent, respectively, which are only slightly greater than their respective duality gaps of 0.354 and 0.126 percent for the DTA and CDA static assignments. Concerning temporal node violations in the final solutions, DTA had 6.7 percent violations, whereas CDA had only 4.2 percent. Overall, CDA performed slightly better that DTA in achieving a lower duality gap, but CDA required roughly six times as much CPU time as DTA for the dynamic assignment to converge on this size network.

Table 3 indicates how the total link volumes and mean travel impedances varied over the six time intervals of the Pittsburgh DTA and CDA dynamic assignments. The third and fourth columns give the trip departures in each interval, both in numbers of vehicle trips and as percentages of total 1-hr trip departures. The remaining columns give the mean trip impedances and 10-min link volumes at 10 screenline locations for DTA and CDA, respectively. The screenline locations are on freeways and major arterials along a radius approximately 3 mi from the central business district and essentially capture all significant volumes of traffic approaching downtown Pittsburgh from the eastern communities.

In a previous study using this network, Janson et al. (*12*) found that equilibrium assignment produced reasonable estimates of link volumes but underestimated link impedances, both before and during a major freeway project. That study found the equilibrium assignment link volumes to differ from observed morning peak-hour link volumes at the 10 screenline locations by an APV of 16.2 percent. The level of convergence obtained in the Pittsburgh F-W assignment for this paper is greater than that obtained by Janson et al. (*12*). As a result, the F-W link volumes found here differ from the observed morning peak-hour link volumes at the 10 screenline locations by an APV of 15.1 percent.

By comparison, the DTA static assignment link volumes for the full 1-hr period differed along the screenline from observed traffic counts by an APV of 20.4 percent and from F-W link volumes by an APV2 of 6.0 percent. The CDA static assignment link volumes for the full 1-hr period differed along the screenline from observed traffic counts by an APV of 17.1 percent and from F-W link volumes by an APV2 of 2.5 percent. Thus, CDA produced 1-hr link volumes that are very slightly different from F-W link volumes and only slightly less accurate when compared with actual counts. The average 1-hr screenline crossings from each assignment was exactly 1,000, which is 10 percent different from the observed value of 1,112.

One additional comparison is the extent to which these two procedures achieve similar impedances for alternative routes used between a given origin-destination pair of zones. An examination of used trip paths indicated that four alternative routes connecting a residential zone east of Pittsburgh with the downtown central business district were assigned trips by the F-W algorithm. Routes A and B use all arterial streets, whereas Routes C and D use arterials and portions of a major freeway called the Parkway East. Each route is roughly 8 mi and contains between 19 and 24 links in the coded network. As expected of a highly converged F-W solution, all four routes had the same impedance of 12.3 min (to one decimal place accuracy). The DTA static assignment resulted in average impedances of 12.1, 12.1, 12.5, and 12.2 min for Routes A, B, C, and D, respectively, over the 1-hr assignment period. The CDA static assignment resulted in average impedances of 12.3, 12.1, 12.0, and 12.3 min for Routes A, B, C, and D, respectively, over the 1-hr assignment period. However, the

TABLE 3 SUMMARY OF PITTSBURGH DYNAMIC ASSIGNMENT RESULTS

Time Interval	Time of Day	Trip Depart Prct	Number of Trip Departs	DTA Mean Trip Time	DTA Total Link Volume*	CDA Mean Trip Time	CDA Total Link Volume*
1	6:50	12.5%	3051	10.2 min	1180	10.3 min	1209
2	7:00	16.5%	4027	10.9 min	1478	10.8 min	1511
3	7:10	21.0%	5125	13.1 min	1948	12.9 min	1916
4	7:20	21.0%	5125	14.8 min	2079	14.5 min	2048
5	7:30	16.5%	4027	12.1 min	1699	11.9 min	1680
6	7:40	12.5%	3051	10.5 min	1377	10.4 min	1309

* Total link volume in each interval for the 10 screenline locations.

CDA impedances for these routes were almost exactly constant over the period, whereas the DTA impedances showed significant variation.

Figures 3 and 4 show the travel times of these four routes for trips departing in each interval resulting from DTA and CDA, respectively. Travel times from DTA for these routes are higher and more varied than the travel times from CDA. The travel times all begin at around 11 min (slightly below the SUE travel time of 12 min due to the low initial departure rate of 12.5 percent), rise to between 13 and 14 min, and then decrease to their initial levels as travel demand falls off to a lower, steady-state level. Although observed travel times are not available for each 10-min departure interval, the author's experience with these routes, having lived in the area for 7 years, is that these times underestimate actual times but show relatively valid magnitudes of peak-period variability between intervals. In validating this network for a previous study, Janson et al. (*12*) found that a SUE assignment from the F-W algorithm also tended to underestimate actual impedances along routes where travel time runs had been made.

CONCLUSIONS AND FUTURE RESEARCH

Overall, the F-W, DTA, and CDA procedures were shown to generate similar steady-state assignments in these examples. However, only the CDA procedure steadily converges toward a dynamic equilibrium solution as formulated in this paper. Compared with DTA, CDA was shown to produce assignments that more closely satisfy DUE conditions as measured by the duality gap. Data are unavailable for the networks used in this study with which to validate the DTA and CDA results against observed link counts and travel times in each time interval. The author is arranging to obtain 10-min traffic counts and a coded network from a major metropolitan planning organization for validation purposes.

The link flow formulation of DUE presented in this paper for multiple origins and destinations and the equivalent path flow formulation presented by Janson (*9*) prevent temporally discontinuous flows from entering the solution. An example was given earlier of how methods of linear combinations can easily cause such flows when applied to this problem.

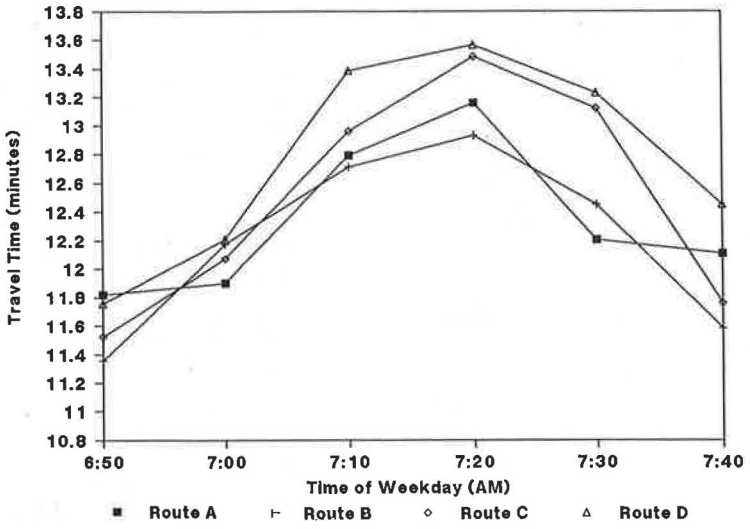

FIGURE 3 DTA impedances of alternative routes.

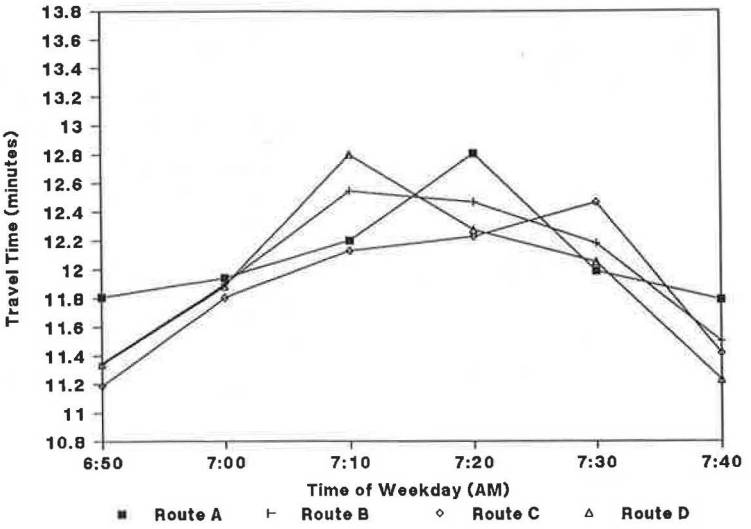

FIGURE 4 CDA impedances of alternative routes.

Hamerslag (*19*) presented results of applying the F-W algorithm to a small network to generate a dynamic assignment but did not evaluate the solution's degree of optimality or temporal continuity. CDA did not exhibit convergence difficulties in the test cases of this paper, and it was relatively quick to converge. With regard to convergence difficulties, Janson and Zozaya-Gorostiza (*20*) show that the F-W algorithm introduces cyclic flows to an assignment that retard its convergence, and this problem will be compounded by the creation of temporally discontinuous flows in dynamic assignments.

The test assignments evaluated in this paper were made with the usual BPR impedance function. Research has been conducted to test whether other functions, such as the Davidson function, may be superior when used in aggregate assignment models (*21*). Hungerink (*22*) suggests a modification of the usual link impedance function as one method of approximating queuing delays in capacity-restrained assignments so that link volumes in excess of capacity affect the impedances of inflow links. Such impedance function alterations might be considered for dynamic assignment procedures, because travel times directly affect the temporal incidences of competing flows on links common to paths from different origins.

A practical advantage of dynamic traffic assignment as formulated and solved in this paper is that it builds directly on the transportation planning data sets and solution algorithms familiar to transportation planners and software developers. DUE and CDA can also be integrated with other travel forecasting procedures. Janson and Southworth (*23*) show how trip departure times can be estimated with dynamic traffic assignment and observed traffic counts on selected links in each time interval. CDA can be run on large networks essentially in real-time using high-speed computers. Thus, CDA is one approach to implementing real-time traffic assignment and route guidance systems on urban transportation networks and to evaluating plans for traffic management during evacuations and special events.

ACKNOWLEDGMENTS

This research was funded by the Director's Research and Development Fund of Oak Ridge National Laboratory and the Federal Emergency Management Agency.

REFERENCES

1. D. K. Merchant and G. L. Nemhauser. A Model and an Algorithm for the Dynamic Traffic Assignment Problem. *Transportation Science*, Vol. 12, 1978, pp. 183–199.
2. D. K. Merchant and G. L. Nemhauser. Optimality Conditions for a Dynamic Traffic Assignment Model. *Transportation Science*, Vol. 12, 1978, pp. 200–207.
3. J. K. Ho. A Successive Linear Optimization Approach to the Dynamic Traffic Assignment Problem. *Transportation Science*, Vol. 14, 1980, pp. 295–305.
4. M. Carey. A Constraint Qualification for a Dynamic Traffic Assignment Model. *Transportation Science*, Vol. 20, No. 1, 1986, pp. 55–58.
5. M. Carey. Optimal Time-Varying Flows on Congested Networks. *Operations Research*, Vol. 35, No. 1, 1987, pp. 58–69.
6. J. G. Wardrop. Some Theoretical Aspects of Road Traffic Research. *Proc., Institution of Civil Engineers*, Vol. 1, Part 2, 1952, pp. 325–378.
7. B. N. Janson. Combined Dynamic Distribution and Assignment with Variable Departure or Arrival Times. Working paper. University of Colorado at Denver, Denver, 1991 (in review).
8. T. L. Friesz, 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. B. N. Janson. Dynamic Traffic Assignment for Urban Road Networks. *Transportation Research B* (forthcoming).
10. M. A. Taha. *Operations Research: An Introduction*. Macmillian Publishing Co., New York, 1982.
11. S. P. Evans. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research*, Vol. 10, 1976, pp. 37–57.
12. B. N. Janson, S. P. T. Thint, and C. T. Hendrickson. Validation and Use of Equilibrium Network Assignment for Urban Highway Reconstruction Planning. *Transportation Research*, Vol. 20A, No. 1, 1986, pp. 61–73.
13. L. J. LeBlanc. An Algorithm for the Discrete Network Design Problem. *Transportation Science*, Vol. 9, 1975, pp. 183–199.
14. M. Fukushima. A Modified Frank-Wolfe Algorithm for Solving the Assignment Problem. *Transportation Research*, Vol. 18B, 1983, pp. 169–177.
15. L. J. LeBlanc, R. V. Helgason, and D. E. Boyce. Improved Efficiency of the Frank-Wolfe Algorithm for Convex Network Programs. *Transportation Science*, Vol. 19, No. 4, 1985, pp. 445–462.
16. G. Rose, M. S. Daskin, and F. S. Koppelman. An Examination of Convergence Errors in Equilibrium Traffic Assignment Models. *Transportation Research*, Vol. 22B, No. 4, 1988, pp. 261–274.
17. C. Hendrickson, R. Carrier, T. Dubyak, and R. Anderson. Traveler Responses to Reconstruction of Parkway East (I-376) in Pittsburgh. *Transportation Research Record 890*, TRB, National Research Council, Washington, D. C., 1982, pp. 33–39.
18. C. Hendrickson and E. Plank. The Flexibility of Departure Times for Work Trips. *Transportation Research*, Vol. 18A, No. 1, 1984, pp. 25–36.
19. I. R. Hamerslag. Dynamic Assignment in the Three-Dimensional Timespace: Mathematical Specification and Algorithm. *Proc., U.S.-Italy Joint Seminar on Urban Traffic Networks: Dynamic Control and Flow Equilibrium*, Capri, Italy, 1989.
20. B. N. Janson and C. Zozaya-Gorostiza. The Problem of Cyclic Flows in Traffic Assignment. *Transportation Research*, Vol. 21B, No. 4, 1987, pp. 299–310.
21. D. E. Boyce, B. N. Janson, and R. W. Eash. The Effects on Equilibrium Trip Assignment of Different Link Congestion Functions. *Transportation Research*, Vol. 15A, No.3, 1981, pp. 223–232.
22. G. Hungerink. Q-Net: Assignment on Over-Congested Networks by Link Inflow Constraint. *Proc., U.S.-Italy Joint Seminar on Urban Traffic Networks: Dynamic Control and Flow Equilibrium*, Capri, Italy, 1989.
23. B. N. Janson and F. Southworth. Estimating Departure Times from Traffic Counts Using Dynamic Assignment. *Transportation Research A* (forthcoming).

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

Dynamic Analysis of User-Optimized Network Flows with Elastic Travel Demand

BYUNG-WOOK WIE

An equivalent continuous time optimal control problem is formulated for dynamic user-optimized traffic assignment with elastic travel demand. Using the Pontryagin minimum principle, optimality conditions are derived and economic interpretations that correspond to a dynamic generalization of Wardrop's first principle are provided. The existence and optimality of singular controls are examined. Under steady-state assumptions, the model is shown to be a proper dynamic extension of Beckmann's equivalent optimization problem for static user-optimized traffic assignment with elastic demand. Finally, limitations and extensions of the present model are discussed.

There has been a great deal of interest in dynamic network equilibrium models. The interest derives from a growing recognition that steady-state network equilibrium is not typically reached during peak hours of commuting and that demand and supply characteristics of urban transportation networks are inherently time-varying in certain situations. With this recognition, a number of researchers have developed dynamic network equilibrium models from different perspectives. Friesz (*1*), Alfa (*2*), and Wie et al. (*3*) provide literature reviews of the dynamic network equilibrium models proposed to date.

In this paper, a simplistic dynamic extension of the static user equilibrium traffic assignment model with elastic demand, which was first formulated as an equivalent optimization problem by Beckmann et al. (*4*), is analyzed. The key simplification is related to the assumption that adjustments from one system state to another may occur instantaneously as traffic conditions on the network change. The problem of dynamic user-optimized traffic assignment with elastic demand is not only to predict time-varying traffic flows on congested networks, but also to predict the temporal distribution of travel demand from each origin node in response to dynamic changes in traffic conditions. This paper should be regarded as an extension of the dynamic user-optimized traffic assignment model presented by Friesz et al. (*5*). The particular extension is to include elastic time-varying travel demand, which leads to the implicit consideration of departure time changes. The previous model considered only the route choice decision-making process because travel demand at each instant was assumed to be known and inelastic with respect to changes in travel costs. This paper is a direct extension of Wie (*6*), which was restricted to a very simple network with one origin-destination pair connected by parallel arcs to a network with many origins and a single destination.

School of Travel Industry Management, University of Hawaii, Honolulu, Hawaii 96822.

To describe individual behaviors of departure time and route choices, we assume that each driver receives complete information on the current state of the network at each instant through an in-car computer connected to a traffic information center. The current traffic information may include instantaneous measures of arc densities and arc capacity changes due to traffic accidents, weather conditions, or road construction. On the basis of continuously updated traffic information, each driver can estimate the instantaneous expected unit path costs from an origin node or any en route intersection node to the destination node. We further assume that all network users attempt to minimize individual travel costs by changing routes and departure times. The problem considered in this paper corresponds to a type of noncooperatively dynamic game in which network users act independently without collaboration and compete with one another for limited network capacity through route and departure time choices.

Our dynamic model has different behavioral assumptions compared with the dynamic models that have been developed by Hendrickson and Kocur (*7*), de Palma et al. (*8*), Mahmassani and Herman (*9*), Ben-Akiva et al. (*10*), Newell (*11*), and Arnott et al. (*12*). The choice of route or departure time, or both, in these dynamic models is generally based on the trade-off between travel time and schedule delay (i.e., a penalty for late or early arrival). In contrast, our dynamic model cannot explicitly treat schedule delay; in other words, the choice of departure time is not based on the trade-off between travel time and schedule delay. Our dynamic model handles departure time and route choices in a sequential manner. At each instant, travel demand—that is, the rate of departure from each origin node—is endogenously determined as a function of the instantaneous expected travel cost between the associated origin-destination pair. Each driver who decides to depart then chooses the shortest path with minimum instantaneous expected unit travel cost to the destination. It is assumed that no driver has information as to how travel costs for further downstream arcs may change by the time of arrival at those arcs. However, as a driver moves downstream, he is free to revise his route choice at any en route intersection node if his current route is no longer optimal on the basis of updated traffic information.

Another important difference in behavioral assumptions is associated with the dynamic user equilibrium conditions. The instantaneous expected travel cost defined in this paper is not the cost actually experienced on that particular day, but an estimate based on current traffic information. Therefore, our

dynamic model cannot predict time-varying traffic flows and elastic travel demands satisfying the dynamic user equilibrium conditions such that all network users with the same desired arrival time and same origin-destination pair experience equal travel cost to the destination regardless of route and departure time chosen. Because the optimality conditions of our dynamic model require only equalization of instantaneous expected unit path costs, it is possible that some drivers can reduce individual travel costs by unilaterally changing routes or departure times.

An equivalent continuous time optimal control problem that corresponds to the problem of dynamic user-optimized traffic assignment with elastic travel demand is formulated. We derive the optimality conditions using the Pontryagin minimum principle, and we examine the existence and optimality of singular controls. The economic interpretation of the optimality conditions is given as a dynamic generalization of Wardrop's first principle. Our dynamic model is also analyzed under steady-state assumptions to show that it is a proper dynamic extension of Beckmann's equivalent optimization problem for static user-optimized traffic assignment with elastic demand. Finally, limitations and extensions of our dynamic model are discussed.

MODEL FORMULATION

Assume a network represented by a directed graph $G(N, A)$, where N is the set of nodes and A is the set of arcs. The cardinality of the set N is denoted by $|N| = n$. Nodes $1, 2, \ldots, n-1$ are origins, whereas n is the only destination. The set of all origins is denoted by M. In general, we use the index a to denote an arc, k a node, and p a path. The set of all paths connecting Node k and Node n is denoted by P_{kn}. We consider a fixed time horizon of length T; that is, all activities occur at some time $t \in [0, T]$.

Let $x_a(t)$ denote the number of vehicles traveling on Arc a at Time t, which will be referred to as the traffic volume on Arc a at Time t. We assume that the instantaneous expected travel cost for a driver (or drivers) entering Arc a at Time t is dependent on $x_a(t)$ and that the instantaneous expected unit cost functions $c_a[x_a(t)]$ are positive, nondecreasing, differentiable, and convex for all $x_a(t) \geq 0$ and $t \in [0, T]$. Link interactions are not considered in this model. To depict the physical phenomenon of traffic congestion on each arc, the exit functions $g_a[x_a(t)]$ are assumed to be nonnegative, nondecreasing, differentiable, and concave for all $x_a(t) \geq 0$ and $t \in [0, T]$ with the additional restriction that $g_a(0) = 0$ for all $a \in A$. A functional form of the exit functions can be represented as

$$g_a[x_a(t)] = g_a^{\max} \cdot \{1 - \exp[-x_a(t)/\beta_a]\} \qquad \forall a \in A \quad \forall t \in [0, T] \qquad (1)$$

where $g_a^{\max}$ is the maximum number of vehicles that can exit from Arc a at each instant and β_a is a parameter that varies with road type and traffic signal system.

The dynamic evolution of the state of each arc is described by first-order nonlinear differential equations:

$$dx_a(t)/dt = \dot{x}_a(t) = u_a(t) - g_a[x_a(t)] \qquad \forall a \in A \quad \forall t \in [0, T] \qquad (2)$$

where

$x_a(t)$ = the state variable denoting the traffic volume on Arc a at Time t;
$u_a(t)$ = the control variable denoting the traffic flow entering Arc a at Time t; and
$g_a[x_a(t)]$ = the traffic flow exiting Arc a at Time t.

Throughout the paper, Equation 2 will be called the state equation. In addition, we assume that the traffic volume on Arc $a \in A$ is a known nonnegative constant at the initial time $t = 0$:

$$x_a(0) = x_a^0(t) \geq 0 \qquad \forall a \in A \qquad (3)$$

Different initial values of traffic volumes may lead to different predictions of time-varying traffic flow patterns.

Let $S_k(t)$ denote the instantaneous travel demand generated (i.e., the rate of departure) from an origin node k at Time t. We assume that $S_k(t)$ is endogenously determined as a function of the instantaneous expected unit travel cost between an origin node k and the destination node n at Time t. It follows that

$$S_k(t) = D_k[t, \mu_k(t)] \qquad \forall k \in M \qquad \forall t \in [0, T] \qquad (4)$$

where $D_k(\cdot)$ is the instantaneous demand function for travel between Nodes k and n at Time t and $\mu_k(t)$ is the minimum instantaneous expected travel cost between Nodes k and n at Time t. We assume that the instantaneous demand functions are nonnegative and monotonically decreasing and that they can continuously change in functional form over the time interval $[0, T]$ to represent time-varying price elasticity of demand for travel between each origin-destination pair. We are unable to discuss a functional form of the instantaneous demand functions and their calibrations. A more realistic dynamic model should also consider cross elasticity of demand, implying that $S_k(t)$ is determined as a function of the trajectory of $\mu_k(t)$ over the time interval $[0, T]$. At this point, we are unable to model this case. Furthermore, we assume that the inverse of the instantaneous demand function is well defined and exists as follows:

$$\mu_k(t) = \Phi_k[t, S_k(t)] \qquad \forall k \in M \qquad \forall t \in [0, T] \qquad (5)$$

The flow conservation constraints are stated as follows:

$$S_k(t) + \sum_{a \in B(k)} g_a[x_a(t)] - \sum_{a \in A(k)} u_a(t) = 0 \qquad \forall k \in M \qquad \forall t \in [0, T] \qquad (6)$$

where

$S_k(t)$ = the control variable, denoting the instantaneous travel demand generated from Node k at Time t;
$A(k)$ = the set of arcs whose tail node is k; and
$B(k)$ = the set of arcs whose head node is k.

We ensure that both the state and control variables are nonnegative:

$$x_a(t) \geq 0 \qquad \forall\, a \in A \qquad \forall\, t \in [0,\ T] \tag{7}$$

$$u_a(t) \geq 0 \qquad \forall\, k \in M \qquad \forall\, t \in [0,\ T] \tag{8}$$

$$S_k(t) \geq 0 \qquad \forall\, k \in M \qquad \forall\, t \in [0,\ T] \tag{9}$$

However, we will not consider the nonnegatively of the state variables in an explicit manner because the assumption $g_a(0) = 0$ ensures that the state variables are always nonnegative.

We define $x(t) \equiv (\ldots, x_a(t), \ldots)$, $u(t) \equiv (\ldots, u_a(t), \ldots)$, and $S(t) \equiv (\ldots, S_k(t), \ldots)$. In the sequel, we employ the set of feasible solutions

$$\Omega \equiv \{[x(t),\ u(t),\ S(t)]\text{: Expressions 2, 3, 6, 8, and 9 are satisfied}\} \tag{10}$$

for economy of notation. We are now ready to formulate the dynamic user-optimized traffic assignment problem with elastic demand as an equivalent continuous time optimal control problem:

$$\text{Minimize } J = \sum_{a \in A} \int_0^T \int_0^{x_a(t)} c_a(\omega)[dg_a(\omega)/d\omega]\, d\omega\, dt - \sum_{k \in M} \int_0^T \int_0^{S_k(t)} \Phi_k(t,\ \eta)\, d\eta\, dt \tag{11}$$

subject to

$$[x(t),\ u(t),\ S(t)] \in \Omega$$

where ω and η are dummy variables of integration. The performance index J is a scalar function that has no intuitive economic interpretation. It should be viewed strictly as a mathematical construction. The derivation of J is analogous to that of the objective function of Beckmann's equivalent optimization problem for a static user equilibrium traffic assignment with elastic demand (*4*).

OPTIMALITY CONDITIONS

The Pontryagin minimum principle (*13*) is used to derive the necessary conditions for an optimal solution of the control problem (Equation 11). We first construct the Hamiltonian function:

$$\mathrm{H}[x(t),\ u(t),\ S(t),\ \lambda(t),\ \mu(t)] = \sum_{a \in A} \int_0^{x_a(t)} c_a(\omega)[dg_a(\omega)/d\omega]\, d\omega - \sum_{k \in M} \int_0^{S_k(t)} \Phi_k(t,\ \eta)\, d\eta + \sum_{a \in A} \lambda_a(t)\left\{u_a(t) - g_a[x_a(t)]\right\} + \sum_{k \in M} \mu_k(t)\left\{S_k(t) + \sum_{a \in B(k)} g_a[x_a(t)] - \sum_{a \in A(k)} u_a(t)\right\} \tag{12}$$

where $\lambda_a(t)$ is the costate variable associated with the state equation (Equation 2) and $\mu_k(t)$ is the Lagrange multiplier associated with the flow conservation constraints (Equation 6). Note that $\lambda(t) \equiv (\ldots, \lambda_a(t), \ldots)$ and $\mu(t) \equiv (\ldots, \mu_k(t), \ldots)$.

The Pontryagin minimum principle states that the dynamic evolution of the costate variables are governed by the following first-order differential equations:

$$-\dot{\lambda}_a(t) = \partial \mathrm{H}[\mathrm{x}(t),\ u(t),\ S(t),\ \lambda(t),\ \mu(t)]/\partial x_a(t) = \left\{c_a[x_a(t)] - \lambda_a(t) + \mu_k(t)\right\} g_a'[x_a(t)] \qquad \forall\, a \in B(k) \qquad \forall\, k \in M \qquad \forall\, t \in [0,\ T] \tag{13}$$

where $g_a'[x_a(t)] = dg_a[x_a(t)]/dx_a(t)$. Equation 13 will be called the costate equation. Let $\phi_a[x_a(T)]$ denote the salvage value function when the terminal state is $x_a(T)$. Because we impose no constraint on the values of the state variables at the terminal time T, the value of $\phi_a[x_a(T)]$ must be equal to zero for all $a \in A$. Hence, the terminal boundary conditions on the costate variables are given as follows:

$$\lambda_a(T) = \partial \phi_a[x_a(T)]/\partial x_a(T) = 0 \qquad \forall\, a \in A \tag{14}$$

Equation 14 is often called the transversality conditions. In addition, the state equation (Equation 2) can be expressed in terms of the Hamiltonian as follows:

$$\dot{x}_a(t) = \partial \mathrm{H}[x(t),\ u(t),\ S(t),\ \lambda(t),\ \mu(t)]/\partial \lambda_a(t) = u_a(t) - g_a[x_a(t)] \qquad \forall\, a \in A \qquad \forall\, t \in [0,\ T] \tag{15}$$

In optimal control theory, the differential equations for the state variables and the differential equations for the costate variables plus all boundary conditions are called the canonical equations, which give rise to the two-point boundary value problem.

The Pontryagin minimum principle also requires that the Hamiltonian (Equation 12) be minimized by choice of the optimal control variables at each point along the optimal state trajectories. The control problem (Equation 11) can thus be converted into an infinity of constrained static optimization problems for each instant $t \in [0,\ T]$ as follows:

$$\text{Minimize } \mathrm{H}[x(t),\ u(t),\ S(t),\ \lambda(t),\ \mu(t)] \tag{16}$$

subject to

$$u_a(t) \geq 0 \qquad \forall\, a \in A$$

$$S_k(t) \geq 0 \qquad \forall\, k \in M$$

While holding $x(t)$ and $\lambda(t)$ constant, the Kuhn-Tucker necessary conditions for $u(t)$ and $S(t)$ to be optimal are readily obtained:

$$\partial \mathrm{H}/\partial S_k(t) = \mu_k(t) - \Phi_k[t,\ S_k(t)] \geq 0 \qquad \forall\, k \in M \qquad \forall\, t \in [0,\ T] \tag{17}$$

$$S_k(t)\,[\partial H/\partial S_k(t)] = S_k(t)\left\{\mu_k(t) - \Phi_k[t, S_k(t)]\right\} = 0$$

$$\forall\, k \in M \qquad \forall\, t \in [0, T] \qquad (18)$$

$$\partial H/\partial u_a(t) = \lambda_a(t) - \mu_k(t) \geq 0$$

$$\forall\, a \in A(k) \qquad \forall\, k \in M \qquad \forall\, t \in [0, T] \qquad (19)$$

$$u_a(t)[\partial H/\partial u_a(t)] = u_a(t)[\lambda_a(t) - \mu_k(t)] = 0$$

$$\forall\, a \in A(k) \qquad \forall\, k \in M \qquad \forall\, t \in [0, T] \qquad (20)$$

$$\partial H/\partial \mu_k(t) = S_k(t) + \sum_{a \in B(k)} g_a[x_a(t)] - \sum_{a \in A(k)} u_a(t) = 0$$

$$\forall\, k \in M \qquad \forall\, t \in [0, T] \qquad (21)$$

The complementary slackness conditions (Equations 17 and 18) indicate that if the optimal value of the control variable $S_k(t)$ is positive, Equation 17 must hold as an equality. In other words

$$\mu_k(t) = \Phi_k[t, S_k(t)] \qquad \forall\, k \in M \qquad \forall\, t \in [0, T] \qquad (22)$$

Both sides of Equation 22 can be inverted to obtain

$$S_k(t) = D_k[t, \mu_k(t)] \qquad \forall\, k \in M \qquad \forall\, t \in [0, T] \qquad (23)$$

Hence, if the traffic flow generated at Node k at Time t is positive, it must be determined by the instantaneous demand function $D_k[t, \mu_k(t)]$. If, however, $\mu_k(t) > \Phi_k[t, S_k(t)]$, then $S_k(t) = 0$, meaning that the minimum instantaneous expected travel cost between Origin k and Destination n at Time t may be too high to induce any departure. Because the instantaneous demand function $D_k[t, \mu_k(t)]$ is assumed to be monotonically decreasing, it follows that its inverse, $\Phi_k[t, S_k(t)]$, should be a decreasing function. The integral of a decreasing function is strictly concave, and the negative of the sum of concave functions is a strictly convex function. Thus, the second term in the Hamiltonian (Equation 12) is strictly convex, implying that the optimization problem (Equation 16) has a unique solution in terms of $S(t)$.

From the complementary conditions (Equations 19 and 20), we know for all $a \in A(k)$ and $k \in M$ that if $\lambda_a(t) > \mu_k(t)$, then $u_a(t) = 0$; if $\lambda_a(t) = \mu_k(t)$, then $u_a(t) \geq 0$. Obviously, the optimal value of the control variable $u_a(t)$ is influenced by the sign of $[\lambda_a(t) - \mu_k(t)]$, which is called the switching function. If, however, $\lambda_a(t) = \mu_k(t)$ for some $a \in A(k)$ and $k \in M$ during a finite time interval, the minimization of the Hamiltonian (Equation 12) leads to nonunique determination of the optimal value of $u_a(t)$, that is, singular control. As a result, the Pontryagin minimum principle yields no useful information to determine the optimal value of the control variable $u_a(t)$. In this circumstance, an additional necessary condition is required to replace the optimality conditions (Equations 19 and 20) so that the singular controls could be tested for optimality. To this end we proceed to derive an expression for the singular control and to ensure that the generalized Legendre-Clebsch condition is satisfied. If $\lambda_a(t) = \mu_k(t)$ for some $a \in A(k)$ and $k \in M$ during a finite time interval $[t_1, t_2] \subseteq [0, T]$, it follows at once that

$$\frac{d}{dt}\left[\frac{\partial L}{\partial u_a(t)}\right] = \dot{\lambda}_a(t) - \dot{\mu}_k(t) = 0 \qquad (24)$$

$$\frac{d^2}{dt^2}\left[\frac{\partial L}{\partial u_a(t)}\right] = \ddot{\lambda}_a(t) - \ddot{\mu}_k(t) = 0 \qquad (25)$$

Consider an arc whose tail node is k and head node is s, that is, $a = (k, s) \in A$. Using the costate equation (Equation 13), we may rewrite Equations 24 and 25 as follows:

$$-\left\{c_a[x_a(t)] - \lambda_a(t) + \mu_s(t)\right\} g_a'[x_a(t)] - \dot{\mu}_k(t) = 0 \qquad (26)$$

$$-\left\{c_a'[x_a(t)]\,\dot{x}_a(t) - \dot{\lambda}_a(t) + \dot{\mu}_s(t)\right\} g_a'[x_a(t)]$$

$$-\left\{c_a[x_a(t)] - \lambda_a(t) + \mu_s(t)\right\} g_a''[x_a(t)]\dot{x}_a(t) - \ddot{\mu}_k(t) = 0 \qquad (27)$$

where $dc_a[x_a(t)]/dx_a(t) = c_a'[x_a(t)]$. For the moment, we suppress the time notation (t) and the traffic volume notation $x_a(t)$. Substitution of Equations 2 and 24 into Equation 27 yields an expression for the singular control as a feedback control:

$$u_a[x(t), \lambda(t), \mu(t)] =$$

$$\frac{(c_a' g_a + \dot{\mu}_k - \dot{\mu}_s) g_a' + (c_a - \mu_k + \mu_s) g_a'' g_a - \ddot{\mu}_k}{c_a' g_a' + (c_a - \mu_k + \mu_s) g_a''}$$

$$\forall\, a = (k, s) \in A \qquad \forall\, k \in M \qquad \forall\, s \in M$$

$$\forall\, t \in [t_1, t_2] \subseteq [0, T] \qquad (28)$$

When a singular control exists, an additional test is needed to determine whether this singular control is optimizing or not. Usually, the optimality of singular controls given by Equation 28 can be tested by using the generalized Legendre-Clebsch condition (*14*) as follows:

$$\frac{\partial}{\partial u_a(t)}\left\{\frac{d^2}{dt^2}\left[\frac{\partial L}{\partial u_a(t)}\right]\right\} = -(c_a - \mu_k + \mu_s) g_a'' - c_a' g_a' \leq 0$$

$$\forall\, a = (k, s) \in A \qquad \forall\, k \in M \qquad \forall\, s \in M$$

$$\forall\, t \in [t_1, t_2] \subseteq [0, T] \qquad (29)$$

Because $c_a[x_a(t)]$ and $g_a[x_a(t)]$ are assumed to be nondecreasing for all $x_a(t) \geq 0$, it follows that $c_a'[x_a(t)]\, g_a'[x_a(t)]$ is nonnegative. We also know that $g_a''[x_a(t)] \leq 0$ because $g_a[x_a(t)]$ is concave for all $x_a(t) \geq 0$. It remains to be proven that $c_a[x_a(t)] - \mu_k(t) + \mu_s(t) \leq 0$. However, this condition is not strictly satisfied because it requires from the costate equation (Equation 13) that $d\lambda_a(t)/dt$ always be nonnegative. Certainly, we know that $d\lambda_a(t)/dt$ can be negative during a finite time period. Hence, we are not able to conclude that the singular controls expressed in Equation 28 are optimal. We reserve this issue for future research.

ANALYSIS OF OPTIMALITY CONDITIONS

Let us consider a path p connecting an origin node k and the destination node n, expressed in generic form as

$$p \equiv [k = k_0, a_1, k_1, \ldots, k_{m-1}, a_m, n = k_m] \quad (30)$$

Using the costate equation (Equation 13), we define the instantaneous unit path travel cost function:

$$\Psi_p(t) = \sum_{a \in A(p)} \left\{ c_a[x_a(t)] + \frac{\dot{\lambda}_a(t)}{g_a'[x_a(t)]} \right\}$$

$$\forall\, p \in P_{kn} \qquad \forall\, t \in [0, T] \quad (31)$$

where $A(p)$ denotes the set of arcs composing a path p. We are now ready to state and prove the following theorem:

Theorem 1. If $u_a(t) > 0$ for all $a \in A(p)$ at some time $t \in [0, T]$, then $\Psi_p(t) = \inf \{\Psi_b(t): \forall\, b \in P_{kn}\}$ for an optimal solution of the dynamic user-optimized traffic assignment problem with elastic demand (Expression 11).

Proof: Using the costate equation (Equation 13), we may rewrite Equation 31 as

$$\Psi_p(t) = \sum_{i=1}^{m} [\lambda_{a_i}(t) - \mu_{k_i}(t)] \quad (32)$$

We know from Equation 20 that if $u_a(t) > 0$ for all $a \in A(p)$,

$$\lambda_{a_i}(t) = \mu_{k_{i-1}}(t) \qquad \text{for } i = 1, \ldots, m \quad (33)$$

Substituting Equation 33 into Equation 32, it follows at once that for a path $p \in P_{kn}$

$$\begin{aligned} \Psi_p(t) &= [\mu_{k_0}(t) - \mu_{k_1}(t)] + [\mu_{k_1}(t) - \mu_{k_2}(t)] + \ldots \\ &\quad + [\mu_{k_{m-1}}(t) - \mu_{k_m}(t)] \\ &= \mu_k(t) - \mu_n(t) \end{aligned} \quad (34)$$

The theorem follows immediately. Q.E.D.

Provided that $\mu_n(t) = 0$, the economic interpretation of the Lagrange multiplier $\mu_k(t)$ can be given as a minimum instantaneous expected unit travel cost between Origin k and Destination n at Time t:

$$\mu_k(t) = \inf \left(\sum_{a \in A(p)} \left\{ c_a[x_a(t)] + \frac{\dot{\lambda}_a(t)}{g_a'[x_a(t)]} \right\} : \forall\, p \in P_{kn} \right) \quad (35)$$

Substituting Equation 35 into Equation 13 yields

$$\begin{aligned} \lambda_a(t) &= c_a[x_a(t)] + \left\{ \dot{\lambda}_a(t)/g_a'[x_a(t)] \right\} + \mu_s(t) \\ &= c_a[x_a(t)] + \left\{ \dot{\lambda}_a(t)/g_a'[x_a(t)] \right\} \\ &\quad + \inf \{\Psi_p(t): \forall\, p \in P_{sn}\} \end{aligned}$$

$$\forall\, a = (k, s) \in A \qquad \forall\, s \in M \qquad \forall\, t \in [0, T] \quad (36)$$

The costate variable $\lambda_a(t)$ can be interpreted as a minimum instantaneous expected unit travel cost between Origin k and Destination n at Time t with only the restriction that $a = (k, s)$ must be the first arc to traverse. Therefore, $\Psi_p(t)$ may be viewed as consisting of static and dynamic components. The term $\sum_{a \in A(p)} c_a[x_a(t)]$ is considered to be static. On the other hand, the term $\sum_{a \in A(p)} [d\lambda_a(t)/dt] g_a'[x_a(t)]$ is considered to be dynamic in that it includes the rate of change of $\lambda_a(t)$ with respect to time, which represents the dynamics of the costate variables, whereas $g_a'[x_a(t)]$ is interpreted as a scaling of $d\lambda_a(t)/dt$.

Theorem 1 requires equilibration of instantaneous unit travel costs for all the paths that are being used at each instant in time for a given origin-destination pair as follows:

$$\Psi_1(t) = \ldots = \Psi_j(t) \le \Psi_{j+1}(t) \le \ldots \le \Psi_J(t) \quad (37)$$

$$u_a(t) > 0 \text{ for all } a \in \{A(p) \mid p = 1, \ldots, j\} \quad (38)$$

$$u_a(t) = 0 \text{ for some } a \in \{A(p) \mid p = j+1, \ldots, J\} \quad (39)$$

where J is the cardinality of the set P_{kn}. Hence, Theorem 1 can be interpreted as a dynamic generalization of Wardrop's first principle (*15*) such that if, at each instant in time, for each origin-destination pair, the instantaneous unit travel costs for all the paths that are being used are identical and equal to the minimum instantaneous unit travel cost, the corresponding time-varying traffic flow pattern is said to be user optimized.

EQUIVALENCY UNDER STEADY-STATE ASSUMPTIONS

We shall establish that the control problem (Equation 11) is a proper dynamic extension of Beckmann's mathematical programming problem for a static user equilibrium traffic assignment with elastic demand (*4*). To this end we analyze our dynamic model under the following steady-state assumptions: first, that $S_k(t)$ and $c_a[x_a(t)]$ are time-invariant for all $a \in A$ and $k \in M$, and second, that $dx_a(t)/dt = 0$ and thus $u_a(t) = g_a[x_a(t)]$ for all $a \in A$ and $t \in [0, T]$. By changing the variables of integration, we may rewrite the first term in the Hamiltonian (Equation 12) and have the following relation:

$$\sum_{a \in A} \int_0^{x_a} c_a(\omega)[dg_a(\omega)/d\omega]\, d\omega = \sum_{a \in A} \int_0^{g_a(x_a)} c_a(\xi)\, d\xi \quad (40)$$

Let f_a denote $g_a(x_a)$, meaning the traffic flow rate on Arc a. Under the steady-state assumptions, the continuous time optimal control problem (Equation 11) becomes a series of constrained static optimization problems that are identical at each instant during the fixed time interval $[0, T]$ as follows:

Minimize $Z(x) =$

$$\sum_{a \in A} \int_0^{f_a} c_a(\omega) d\omega - \sum_{k \in M} \int_0^{S_k} \Phi_k(\eta)\, d\eta \quad (41)$$

subject to

$$S_k + \sum_{a \in B(k)} f_a - \sum_{a \in A(k)} f_a = 0 \qquad \forall\, k \in M$$

$$f_a \geq 0 \qquad \forall\, a \in A$$

$$S_k \geq 0 \qquad \forall\, k \in M$$

The Kuhn-Tucker necessary conditions for the problem (Equation 41) can be readily obtained as

$$f_a[c_a(f_a) - \mu_k] = 0 \qquad \forall\, a \in A \tag{42}$$

$$c_a(f_a) - \mu_k \geq 0 \qquad \forall\, a \in A \tag{43}$$

$$S_k[\mu_k - \Phi_k(S_k)] = 0 \qquad \forall\, k \in M \tag{44}$$

$$\mu_k - \Phi_k(S_k) \geq 0 \qquad \forall\, k \in M \tag{45}$$

$$S_k + \sum_{a \in B(k)} f_a - \sum_{a \in A(k)} f_a = 0 \qquad \forall\, k \in M \tag{46}$$

$$f_a \geq 0 \qquad \forall\, a \in A \tag{47}$$

$$S_k \geq 0 \qquad \forall\, k \in M \tag{48}$$

where μ_k is the Lagrange multiplier associated with the flow conservation constraints, denoting the minimum unit travel cost between Origin k and Destination n. Because the optimality conditions (Equations 42 through 48) are identical to user equilibrium conditions (*16*), the control problem (Equation 11) is proven to be a proper dynamic extension of Beckmann's equivalent optimization problem for a static user equilibrium traffic assignment with elastic demand.

CONCLUSION

We have shown that an equivalent continuous time optimal control problem can be formulated to model the problem of dynamic user-optimized traffic assignment with elastic travel demand. The optimality conditions were derived and given economic interpretations. We have also shown that the model presented in this paper is, under steady-state assumptions, a proper dynamic extension of Beckmann's equivalent optimization problem for static user-optimized traffic assignment with elastic demand.

The dynamic model encounters several limitations. First, the instantaneous costs are used as a criterion for departure time and route choices. A more realistic model should use the anticipated costs as the approximations of the actually experienced costs. Second, the instantaneous travel demand is determined as a function of the instantaneous perceived unit cost between the associated origin-destination pair. This assumption implies that no cross elasticity of demand is considered in our dynamic model. Third, the choice of departure time and route is not based on the trade-off between travel time and schedule delay. It means that our model cannot explicitly handle schedule delay as a penalty for early or late arrival at the destination. Fourth, the state equation (Equation 2) is not empirically tested to answer whether it provides an adequate representation of reality. No consideration of cross-link interactions further simplifies the dynamics of traffic flow on each arc. Finally, a functional form of the instantaneous demand function is not specified in this paper. The question arises as to whether time-varying elasticity of demand can be well represented in its functional form.

Our future research includes the following important issues. First, we should be able to test the optimality of singular controls using the generalized convexity condition, and we need to investigate transition process from a singular control to a nonsingular control. Second, our dynamic model should be extended to analyze traffic flows in a congested network with multiple origins and multiple destinations. However, as discussed by Wie et al. (*3*) and Wie (*17*), there are some difficulties in generalizing the present model to a multiple destination case. Its generalization requires the linear exit function assumptions to keep the property of separability, which is crucial to show equilibration of instantaneous unit path costs. Last, an efficient algorithm must be developed for the computation of our dynamic model. Recently, a gradient algorithm based on the discrete maximum principle was developed by Wie (*18,19*) to solve the problem of dynamic user-optimized traffic assignment with fixed demand. It is hoped that the algorithm can be modified to solve the problem with elastic demand.

REFERENCES

1. T. L. Friesz. Transportation Network Equilibrium, Design and Aggregation: Key Developments and Research Opportunities. *Transportation Research*, Vol. 19A, 1985, pp. 413–427.
2. A. S. Alfa. A Review of Models for the Temporal Distribution of Peak Traffic Demand. *Transportation Research*, Vol. 20B, 1986, pp. 491–499.
3. B. W. Wie, T. L. Friesz, and R. L. Tobin. Dynamic User Optimal Traffic Assignment on Congested Multidestination Networks. *Transportation Research*, Vol. 24B, No. 6, 1990, pp. 431–442.
4. M. Beckmann, C. McGuire, and C. Winston. *Studies in the Economics of Transportation*. Yale University Press, New Haven, Conn., 1956.
5. T. L. Friesz, F. J. Luque, R. L. Tobin, and B. W. Wie. Dynamic Network Traffic Assignment Considered as a Continuous Time Optimal Control Problem. *Operations Research*, Vol. 37, No. 6, 1989, pp. 893–901.
6. B. W. Wie. 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.
7. C. Hendrickson and G. Kocur. Schedule Delay and Departure Time Decisions in a Deterministic Model. *Transportation Science*, Vol. 15, 1981, pp. 62–77.
8. A. de Palma, M. Ben-Akiva, C. Lefevre, and N. Litinas. Stochastic Equilibrium Model of Peak Period Traffic Congestion. *Transportation Science*, Vol. 17, 1983, pp. 430–453.
9. H. S. Mahmassani and R. Herman. Dynamic User Equilibrium Departure Time and Route Choice on Idealized Traffic Arterials. *Transportation Science*, Vol. 18, 1984, pp. 362–384.
10. M. Ben-Akiva, A. de Palma, and P. Kanaroglou. Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates. *Transportation Science*, Vol. 20, 1986, pp. 164–181.
11. G. F. Newell. The Morning Commute for Nonidentical Travelers. *Transportation Science*, Vol. 21, 1987, pp. 74–82.
12. R. Arnott, A. de Palma, and R. Lindsey. Schedule Delay and Departure Time Decisions with Hetereogeneous Commuters. In *Transportation Research Record 1197*, TRB, National Research Council, Washington, D.C., 1989, pp. 56–67.
13. L. S. Pontryagin, V. A. Boltyanski, R. V. Gankrelidge, and E. F. Mishenko. *The Mathematical Theory of Optimal Process*. John Wiley and Sons, New York, 1962.
14. A. E. Bryson and Y.-C. Ho. *Applied Optimal Control* (revised version). John Wiley and Sons, New York, 1975.

15. J. G. Wardrop. Some Theoretical Aspects of Road Traffic Research. *Proc., Institution of Civil Engineers*, Vol. 2, No. 1, 1952, pp. 235–378.
16. Y. Sheffi. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1985.
17. B. W. Wie. *Dynamic Models of Network Traffic Assignment: A Control Theoretic Approach*, Ph.D. dissertation. University of Pennsylvania, 1988.
18. B. W. Wie. *A Gradient Algorithm for Dynamic User Optimized Traffic Assignment*. Working paper. University of Hawaii, Honolulu, 1990.
19. B. W. Wie. *A Comparison of Dynamic User-Optimum vs System-Optimum Traffic Assignment*. Working paper. University of Hawaii, Honolulu, 1990.

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

Multicriteria Decision Making in Location Modeling

ALI E. HAGHANI

Public- and private-sector location decisions are generally made in a multiobjective planning environment. Examples include location of emergency medical service facilities and warehouses. Various criteria considered in making these decisions may include minimization of costs, maximization of demands that are satisfied within a prespecified time or distance, and minimization of average distance from demand points to the nearest facility. Generally, location problems are formulated and solved as mathematical optimization problems. Attempts have been made to solve location problems through multiobjective optimization. Generally, these models account for more than one objective by using a variety of weighting schemes or by concentrating on one objective and incorporating the others into the model as constraints. An alternative approach to consider several objectives simultaneously is presented. The approach is based on the analytical hierarchy process, which is a useful tool in multicriteria decision making. The approach is demonstrated by a numerical example.

Engineers and planners in both the public and the private sectors are frequently faced with deciding where to locate facilities and how to allocate the resources for these facilities among competing demands. Public-sector examples of such decisions include where to locate emergency medical service (EMS) vehicles to serve a community, how such vehicles should be dispatched to incidents, how many fire stations are needed so that all points in an urban area may be reached within a prespecified time, and where public schools should be located. Private-sector examples include where to locate maintenance facilities over an airline or railroad network and how many warehouses are needed in a distribution system and where they should be located. In all cases, the location of the facilities significantly affects both the operating costs of the system and the ability of the system to satisfy the demands placed on it.

Facility location decisions in both public and private sectors are generally multiobjective in nature. For example, in locating emergency medical service facilities, the objectives are to minimize the average travel time to an incident and the number of vehicles deployed, and hence the operating cost of the system. There may also be a need to consider many other objectives, such as balancing the work load, increasing the number of people served by the system, and equity. Some objectives may be conflicting in nature. For example, minimizing the number of deployed vehicles conflicts with maximizing the number of people served.

Generally, location-allocation problems have been formulated and solved as mathematical optimization problems. Different formulation approaches and the use of different objective functions have resulted in a variety of location models.

Department of Civil Engineering, University of Maryland, College Park, Md.

Although attempts have been made to deal with location decisions from a multiobjective point of view (*1–5*), most location models are formulated and solved in a single-objective framework. Among the single-objective models, a few studies attempt to deal with the location-allocation problem in a hierarchical modeling approach, and mostly from a dual objective point of view (*6–9*). This paper presents an alternative approach for making multiobjective location decisions. Theoretically, the approach enables analysts and decision makers to simultaneously consider as many objectives as they wish without defining any a priori weights for the objectives or setting any constraints on the level of achievement of those objectives.

PROBLEM STATEMENT

Location problems are frequently confronted in both the public and the private sectors. The most common location problems in the public domain include choosing fire station sites and determining the number and location of EMS vehicles. Private-sector location problems include determining the number of maintenance facilities needed on an airline or railroad network and the optimal location of these facilities and identifying optimal warehousing locations in product distribution networks.

The allocation problem is closely related to the location problem. The allocation problem deals with optimal assignment of the demands to service centers. The location-allocation problem derives its importance from two sources. First, in many cases, the location of service centers in a network and the allocation of demands to service facilities has a direct effect on system operating costs. This is particularly true in private-sector applications. The second reason for interest in the location-allocation problem is that, in some cases, the nature of the demands and of the service depends dramatically on the ability of the customers to obtain service quickly. An example is medical emergencies, in which timely response is clearly needed.

Research on the location-allocation problem must recognize that location decisions are generally made in a multiobjective planning framework. For example, in locating EMS vehicles, the objectives are to minimize the average travel time to an incident, the maximum travel time to an incident, and service differences between geographic areas of the city; maximize the number of people or potential demands that can be served by the system within a given time limit; and minimize the number of vehicles deployed. In fact, these are only representative of the many objectives that public decision makers must consider in locating and in choosing response

districts for EMS vehicles. Similarly, private location-allocation decisions are also multiobjective in nature. In choosing repair centers for an airline network, for example, the objectives are to minimize the amount of deadheading required to reach the facilities and the construction and operating costs of the facilities.

One important difference between private- and public-sector location problems is that it is generally easier to collapse the different objectives into a single objective in private-sector location problems. For example, deadheading may be converted into an operating cost. By using appropriate weighting and discounting factors, the deadheading, construction, and operating costs in the airline example may be collapsed into a single cost figure. In the EMS vehicle location example, however, it would be virtually impossible to collapse measures of service inequities across geographic areas and measures of vehicle work loads into a single value to be optimized. Even if it were possible, it would require the analyst to implicitly weight the two objectives. This task is better left to the public decision makers. This, however, demands that the analyst develop techniques capable of clearly displaying the trade-offs between objectives.

This paper presents an approach that, theoretically, enables simultaneous consideration of as many objectives as desired in making location decisions. This approach is based on the analytical hierarchy process (AHP). AHP is selected as a vehicle for evaluation and ranking of alternative sites (or combinations of sites) on the basis of multiple objectives. It has proven to be a valuable tool in multicriteria decision making. In this paper we deal with objectives that are generally used in public-sector planning. However, the methodology can easily be generalized to include as many and as different objectives as needed. The paper focuses on discussion of issues involved in implementation of AHP in multiobjective location modeling.

DETERMINISTIC LOCATION MODELS

Facility location is one of the most important long-term logistical decisions faced in the public or private sector. Most facility location models are concerned with network locations rather than plane locations. The location models can be categorized on the basis of such important criteria as type of objective function and the deterministic versus stochastic models. In this section we focus on deterministic location models and distinguish them by their objective function. Stochastic models are beyond the scope of this study.

Three general types of objective functions have been used the most in the literature concerning public-sector location decisions. Covering models locate facilities on a network such that the demands are covered within a prespecified critical time or distance. Median or minisum models locate the facilities such that the weighted average distance between the facilities and the demands served by those facilities is minimized. Center or minimax models locate facilities to minimize the weighted maximum distance from the facilities to the demands served by them. Median, covering, and center models encompass a large portion of the location models, and therefore we focus on these models. The interested reader is referred to Handler and Mirchandani (*10*), Tansel et al. (*11*), or Brandeau and Chiu (*12*) for a more complete review. In what follows, we describe covering, median, and center location models and then briefly review some of the multiobjective approaches to location modeling.

Covering Models

The inputs to location models include a network with a set of N nodes and A links. Associated with each node is a demand to be served by the facilities. Demands are assumed to be generated at nodes only. The travel time (or travel distance) between Nodes i and j is represented by a shortest path matrix. Also, a prespecified critical time (or distance) is defined, which is the maximum time (or distance) limit.

The simplest covering model is the location set covering model, which was originally formulated by Toregas et al. (*13*). This formulation attempts to minimize the number of service centers on a network such that all demands are covered within the critical time or distance. This model has been used in locating ambulances (*14*) and fire stations (*15,16*). Plane and Hendrick (*17*) studied a dual objective formulation of the set covering problem. They minimize the number of fire stations needed and maximize the number of existing stations in the solution. Daskin and Stern (*6*) proposed an alternative dual objective formulation that minimizes the number of service centers and maximizes the amount of backup coverage. Demand-weighted backup coverage has also been formulated (*18,19*).

The set covering model fails to recognize that demands are generated at the nodes at different rates. The maximum covering location model formulated by Church and ReVelle (*20*) accounts for this by trying to maximize the number of demands that are within the critical time (or distance) of one of the P facilities that are to be located. Maximum covering models have been used in practice to locate service facilities (*21–23*).

Minisum or Median Models

The P-median or minisum problem was originally formulated by Hakimi (*24*). This model minimizes the weighted average distance from a demand node i to the facility to which it is assigned. In the absence of capacity constraints or other complications, demands are assigned to the nearest facilities. In using this model for an EMS facility location problem, the objective is to locate facilities so that the best average behavior of the system is obtained.

Hakimi (*24*) has shown that at least one optimal solution to this problem consists of locating only on the nodes of the network. Hakimi's result has been generalized to a number of extensions of the P-median problem (*25,26*). A variety of heuristic methods have been proposed for solving the P-median problem (*10,27*).

Minimax or Center Models

These models minimize the weighted maximum distance from the demand points to the nearest facilities. When facilities are to be located on nodes of a network, minimax models are referred to as vertex-center models. Center models attempt

to locate facilities over the network so that the level of service in the worst possible case is as good as possible (*10*).

The objectives incorporated in the covering, median, and center models represent three of the most important objectives that can be considered in location of public facilities—in particular, EMS facilities such as ambulance and fire stations. It is clear that the best solution under one modeling approach and for one of the foregoing objectives may not be the best or even a good solution for the other objectives.

Furthermore, although these objectives are among the most important, they are not necessarily the only ones to be considered. In public-sector facility location many other objectives may be equally important. Balancing the work load among facilities, providing equitable service, providing as much backup service as possible, and political considerations are other objectives that may be considered. Some objectives may not even be quantifiable and cannot be incorporated into a mathematical model. It is desirable, therefore, to approach location decisions in a multiobjective framework that addresses these issues.

Multiobjective Approaches

One of the important aspects of location decisions is their long-term effects on the service provided to the public. Several models have been formulated to deal with the public facility location problem in a multiobjective framework. Multiobjective approaches have been used to locate energy facilities (*28–30*) and fast food restaurants (*31*). Cohon et al. (*29*) formulated a multiobjective linear programming model for locating energy facilities. Mladineo et al. (*30*) propose a multicriteria approach for ranking the alternative sites. Min's model (*31*) considers the behavioral and spatial aspects of location scenarios. Fortenberry et al. (*32*) and Heller et al. (*33*) propose models to locate emergency medical service facilities in a multiobjective environment. Fortenberry et al. (*32*) use linear programming to determine optimal locations, and Heller et al. (*33*) propose a model that minimizes the mean response time and balances the facility work load. The results of the latter model are validated by simulation techniques.

Ross and Soland (*34*) present an interactive algorithm for multicriteria optimization that solves a finite sequence of generalized assignment problems for location of public facilities. Schilling (*35*) proposed a dynamic location model to locate public facilities. His approach uses multiobjective analysis to plan public-sector facility systems that operate in a dynamic environment.

Goal programming is used as a technique for approaching location decisions in a multicriteria environment (*36–38*). Some researchers have developed interactive models for locating facilities (*39,40*). Multicriteria approaches have also been used in locating private-sector facilities (*36,41–43*). Buhl (*44*) presents several theorems characterizing single-objective reductions of multiobjective problems and shows that the objective functions used in location theory contain both implicit and explicit value judgments.

In general, all but a few of the multiobjective approaches to location modeling use an optimization framework. Either all of the objectives to be considered are collapsed into a single objective with a weighting scheme or thresholds are defined for all but one objective and the problem is approached by optimizing that objective while the others are constrained within their predefined thresholds (as in goal programming). In these approaches, nonquantifiable objectives are either ignored or dealt with exogenously.

Although efforts have been made to formulate and solve location models that deal with more than one objective, this area of research is still promising. In particular, a multiobjective approach that can incorporate several objectives, provide a systematic evaluation and ranking of alternatives, and, particularly, handle nonquantifiable objectives would be desirable.

In contrast to other multiobjective approaches, optimization is proposed as an alternative generation tool while at the same time a framework is devised enabling us to implement a multicriteria decision-making tool such as AHP to select the best alternative from those generated by the optimization approach. This approach theoretically enables us to consider as many objectives as desired simultaneously with no prior weighting schemes or threshold definitions.

MULTIOBJECTIVE APPROACH TO LOCATION DECISIONS

In this section a multiobjective approach to location modeling incorporating the maximum covering, median, and center is presented. We also incorporate cost considerations as the fourth criterion to be considered in the location of facilities. The approach is based on AHP (*45–50*). The procedure has proven useful in multicriteria decision making (*51*).

AHP

AHP provides a way to organize complex decision-making problems in a manner that allows for interaction and interdependence among factors influencing the decisions and still allows the analyst to think about these factors in a simple way. It enables the analyst to make effective decisions on complex issues by simplifying and expediting the natural decision-making process.

AHP is based on three principles: decomposition, comparative judgments, and synthesis of priorities. A complex, unstructured problem is decomposed into its component parts, which are further arranged into a hierarchic order. The elements in the hierarchy define the problem. A matrix of pairwise comparisons of the relative importance of the elements in a level of hierarchy with respect to the elements in the level immediately above it is then set up. Finally, the global or composite priorities of elements at the lowest level of the hierarchy (alternative solutions) are synthesized.

AHP also provides an effective structure for group decision making. It enables decision makers to represent the simultaneous interactions of many factors in complex, unstructured situations and helps to identify and set priorities on the basis of various objectives. A detailed description of AHP is beyond the scope of this paper. The interested reader is referred to Saaty (*45,48*). However, the main steps of the process can be summarized as follows (*46*):

1. Define the problem and specify the alternative solutions.
2. Break the problem down into a system of components at different levels of hierarchies. Each component in a higher level of hierarchy encompasses some or all of the components in the next lower level.
3. Construct a pairwise comparison matrix of the relevant contributions or impacts of each component on each governing component (criterion) in the next higher level. In this matrix, pairs of elements are compared with respect to a criterion in the superior level. The numbers in the cells of this matrix represent the superiority (inferiority) of each component compared with the others, with respect to their contribution to the governing component. The numbers can be based on either subjective judgments or available numerical data that measure the performance of the alternatives with respect to the criteria under consideration.
4. Obtain all judgments required to develop the set of matrices in the third step.
5. Establish the priorities of components at each level of hierarchy with respect to the criterion or component in the higher level encompassing those components.
6. Repeat Steps 3 through 5 until the priorities are established for all levels of hierarchy.
7. Vectors of priorities are then weighted by the weight of the criteria of each level, and this process is repeated until a priority vector for the lowest level of hierarchy (the alternative solutions) is obtained.
8. The final decision or outcome depends on the vector of priorities for the lowest level of hierarchy and can be evaluated on the basis of consistency measures.

AHP in Location Analysis

The facility location problem can be approached as a multiobjective decision-making process using AHP. The problem can be decomposed to three levels of hierarchy. The first, finding the best sites among the candidates for locating the facilities, is the focus of the problem. The second level of hierarchy consists of the factors or objectives affecting this decision. Finally, the third level of hierarchy includes alternative sites or combinations of sites. According to AHP, matrices of pairwise comparisons between various candidate sites can be established on the basis of individual objectives. The matrices allow the analyst to rank the alternative sites according to the individual objectives. Then an overall vector of priorities (or ranking of alternatives) can be obtained by weighting the priorities according to the individual objectives by the relative weights of the objectives. Note that a vector of priorities for the different objectives can also be obtained through construction of a pairwise comparison matrix for the objectives themselves. The priorities can then be used as weights for the corresponding objectives. Construction of this pairwise comparison matrix requires judgment by the decision maker or the analyst on the relative importance of the objectives. The important issue, however, is that the value judgments can be made by comparing the relative merits of only two objectives at a time, a much easier task than comparing all objectives simultaneously. When these pairwise value judgments are obtained and the comparison matrix is completed, the overall weights for all of the objectives can be obtained. Therefore no a priori weights for the objectives need to be known and no threshold on the level of achievements of the objectives needs to be defined.

There are three major issues in implementation of a tool like AHP in location decision making, in particular when we are dealing with multiple facility location. The first is the hierarchical structure to be used. In a single-family location case, a hierarchical structure like the one mentioned earlier can easily be used. However, in a more complex multifacility location problem it is not clear that such a structure is the best. The second major issue, and perhaps the most important, is to generate the set of alternatives to be evaluated using AHP. In cases where a single facility is to be located, candidate sites can be used as individual alternatives. When we are dealing with large networks on which multiple facilities must be located, we have a combinatorial problem, and identification of the alternatives is not a simple task. Construction of the pairwise comparison matrices is the third issue. If these issues are successfully resolved, AHP will offer tremendous advantages in its simplicity of application, its ability to deal with qualitative objectives, and its theoretical ability to deal simultaneously with as many objectives as desired.

In this paper, we try to deal with these issues in a simple example. Further research is required to address them in a more general case, along with development of other multiobjective approaches that could present similar advantages. In the next section we present the implementation of a multiobjective approach to locating facilities over a small network based on AHP. The example, although simple, provides useful insights into the multiobjective nature of location decisions.

A Simple Example

Consider the problem of locating a single facility or two facilities on the nodes of the network shown in Figure 1. The facilities must respond to the demands generated at the nodes of the network. The shortest path distance between the nodes of the network, along with the demand for service at each node and the cost of operation and maintenance of a facility located at each of the nodes, is given in Table 1.

Assume that in this location decision we need to consider four different criteria: (*a*) achievement of lowest operation and maintenance costs, (*b*) coverage of as much demand as possible within a critical distance of 10 units, (*c*) provision of service such that the average distance from the located facilities to the demand locations is minimized, and (*d*) provision of service such that the level of service in the worst possible case is as good as possible. As mentioned before, consideration of these criteria individually generally results in location decisions that are drastically different. In fact, locations that might be optimal under one criterion may not even be good locations under another. Therefore, we must consider the trade-offs among the criteria. All nodes are considered to be candidate sites for the location of the facilities.

First, we explore the results of different approaches to the single-facility location problem. Using the individual criteria, this simple location problem can be formulated as either a cost minimization, a maximum covering, a 1-median, or a vertex-1 center problem. Table 2 gives the value of each of

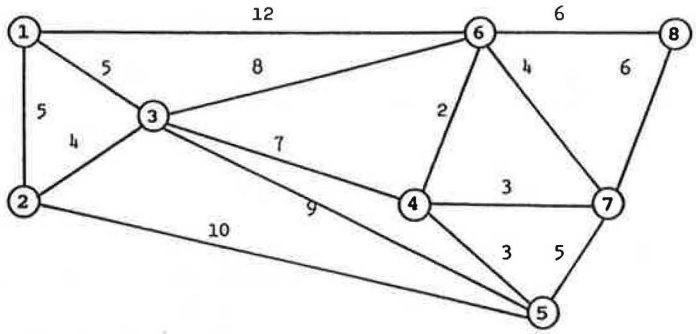

FIGURE 1 Example network.

TABLE 1 DISTANCES, DEMANDS, AND COSTS

Nodes	1	2	3	4	5	6	7	8
1	-	5	5	12	14	12	15	18
2	5	-	4	11	10	12	14	18
3	5	4	-	7	9	8	10	14
4	12	11	7	-	3	2	3	8
5	14	10	9	3	-	5	5	11
6	12	12	8	2	5	-	4	6
7	15	14	10	3	5	4	-	6
8	18	18	14	8	11	6	6	-
Demands	9	13	6	3	8	5	7	10
Costs	9	10	13	12	12	8	9	7

TABLE 2 OBJECTIVE FUNCTION EVALUATION

Location	Vertex-1 Center	1-Median	Maximum Covering	Cost
1	180	588	28	9
2	180	520	36	10
3	140	440	51	13
4	143	428	39	12
5	130	489	42	12
6	156	446	39	8
7	182	506	39	9
8	234	664	25	7
Optimal Location	5	4	3	8

the individual objectives when the facility is located at each of the nodes. The solution of the single-facility location problem is trivial and can be determined by inspection in this case. Table 2 also gives the best location on the basis of the individual criteria.

Note that on the basis of each of these criteria, the optimal location of the facility is a different node. It is clear that each location, optimal on the basis of an individual objective, may not be optimal when all four objectives are considered simultaneously.

We now present the results of an AHP-based approach. For this simple problem, the hierarchies can be structured as shown in Figure 2.

The next step is to set up the pairwise comparison matrices for the alternative sites on the basis of the individual objectives. The cells of each matrix indicate whether or not the site represented by the row is superior or inferior to the site represented by the column on the basis of the objective represented by that matrix. A cell value greater than (less than) 1.0 indicates that the site represented by the row in the matrix is superior (inferior) to the site represented by the column. A cell value of 1.0 indicates that the two sites are equivalent with respect to that objective.

In general, to fill the comparison matrices, Saaty (*46*) suggests using values from 1 to 9 in comparing two alternatives. The value 1 indicates that the two alternatives are equivalent with respect to the criterion under consideration, and the value 9 indicates that one alternative has the highest possible priority over the other. The diagonal elements of these matrices are all 1's, and when the value for Cell (i, j) is determined, the value for Cell (j, i) is the reciprocal of the value for Cell (i, j).

An important issue in setting up the comparison matrices is consistency in judgment when comparing alternatives. Consistency means that if Alternative i is preferred twice as much as Alternative j, and Alternative j is preferred twice as much as Alternative k, then Alternative i should be preferred four times as much as Alternative k. In many cases in the real world enforcing perfect consistency in judgment is not possible; however, relative consistency is desirable so that the judgments do not appear to be random. This is particularly true in dealing with qualitative judgments rather than judgments based on quantitative measures. Fortunately, AHP provides a measure to determine the overall consistency of

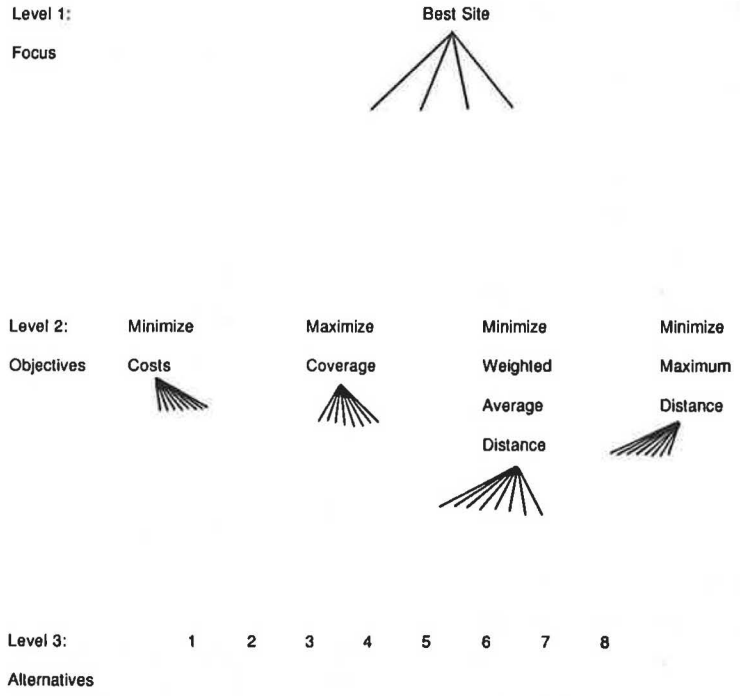

FIGURE 2 Levels of hierarchy for the location problem.

judgments by means of a consistency ratio. The details of calculation of this ratio are given in Saaty (*46*), and a consistency ratio of 0.10 or less is deemed to represent good consistency (a consistent matrix has a consistency ratio of 0).

To construct the comparison matrices, our preference among alternative sites with respect to the individual criterion is based on the value of the mathematical objective function that represents the criterion under consideration. The criteria we are considering in our problem are represented by cost minimization, maximum covering, 1-median, and vertex-1 center objective functions. Therefore, the level of preference of one site over the other with respect to a particular criterion is determined by the ratio of the values of the objective function that represents that criterion when those sites are chosen for locating the facility. For example, when comparing Site 3 with Site 1 with respect to the center objective function, we use the ratio 180/140 in Cell (3, 1), which indicates that Site 3 is preferred 1.286 times as much as Site 1 for this objective, and we use the reciprocal of this value (0.778) in Cell (1, 3). Note that we are implicitly assuming that our preference among the alternatives is directly proportional to their degree of attainment of the objective under consideration. In this case, where the objectives are readily quantifiable, such an assumption may be appropriate. However, in a general case when we are dealing with nonquantifiable objectives, the preference structure may be complex, and such an assumption cannot be made. In that case, obtaining numerical values for the cells of the comparison matrices depends on comparing the values of the objectives among the alternatives, which by itself is an important task. Even when the objectives are quantifiable, it does not necessarily mean that an alternative with twice the objective function value of another alternative will be preferred twice as much. These are important issues that must be addressed in future research. In any event, this assumption can easily be relaxed by incorporating other preference structures. The advantage of AHP is that, as long as a preference structure is agreed upon by the decision makers, it provides an excellent framework for further analysis.

This approach has the advantage that the comparison matrices are perfectly consistent because all of the cells represent the ratios of the objective function values. In cases where the criteria are not easily quantifiable or calculation of numerical values is difficult, Saaty's general procedure could be used. These matrices are shown in Table 3, and the vectors of priorities synthesized on the basis of these comparison matrices are shown in Table 4 [details on how the vectors of priorities are synthesized are given by Saaty (*46*)]. The underlined cell in each column indicates the most attractive alternative according to the objective represented by that column. The most attractive alternatives identified by AHP on the basis of the individual objectives correspond to the optimal locations identified in Table 2.

The overall ranking of the alternatives is obtained by multiplying the priority of each alternative based on each of the criteria by the weight of that criterion and summing the results for each alternative over all criteria.

As long as a simple pairwise comparison of alternatives can be made, there is no need to know the weights of the different criteria a priori. The weights can be determined in the same way as the vectors of priorities for individual criteria, by setting up a comparison matrix for the criteria themselves. The cells of this matrix determine whether a particular criterion is superior or inferior to another and show the preference structure among the criteria.

In this example, if all criteria have the same level of importance, they will have equal weights of 0.25 each, and the vector of overall priorities is (0.112, 0.118, 0.138, 0.130, 0.131, 0.138, 0.125, 0.109). This means that both Nodes 3 and 6 are

TABLE 3 COMPARISON MATRICES BASED ON DIFFERENT OBJECTIVES

Nodes	1	2	3	4	5	6	7	8
1	1.00	1.00	0.78	0.79	0.72	0.87	1.01	1.30
2	1.00	1.00	0.78	0.79	0.72	0.86	1.01	1.30
3	1.29	1.29	1.00	1.02	0.93	1.11	1.30	1.67
4	1.26	1.26	0.98	1.00	0.91	1.09	1.27	1.64
5	1.39	1.39	1.08	1.10	1.00	1.20	1.40	1.80
6	1.15	1.15	0.90	0.92	0.83	1.00	1.17	1.50
7	0.99	0.99	0.77	0.79	0.71	0.86	1.00	1.29
8	0.77	0.77	0.60	0.61	0.56	0.67	0.78	1.00

(a) Comparison Based on Center Objective Function

Nodes	1	2	3	4	5	6	7	8
1	1.00	0.88	0.75	0.73	0.83	0.76	0.86	1.12
2	1.13	1.00	0.85	0.82	0.94	0.86	0.97	1.28
3	1.34	1.18	1.00	0.97	1.11	1.01	1.15	1.51
4	1.37	1.22	1.03	1.00	1.14	1.04	1.18	1.55
5	1.20	1.06	0.90	0.88	1.00	0.91	1.04	1.36
6	1.32	1.17	0.99	0.96	1.10	1.00	1.14	1.49
7	1.16	1.03	0.87	0.85	0.97	0.88	1.00	1.31
8	0.89	0.78	0.66	0.65	0.74	0.67	0.76	1.00

(b) Comparison Based on Median Objective Function

Nodes	1	2	3	4	5	6	7	8
1	1.00	0.79	0.55	0.72	0.67	0.72	0.72	1.12
2	1.29	1.00	0.71	0.92	0.86	0.92	0.92	1.44
3	1.82	1.42	1.00	1.31	1.21	1.31	1.31	2.04
4	1.39	1.08	0.77	1.00	0.93	1.00	1.00	1.56
5	1.50	1.17	0.82	1.08	1.00	1.08	1.08	1.68
6	1.39	1.08	0.77	1.00	0.93	1.00	1.00	1.56
7	1.39	1.08	0.77	1.00	0.93	1.00	1.00	1.56
8	0.89	0.69	0.49	0.64	0.60	0.64	0.64	1.00

(c) Comparison Based on Covering Objective Function

Nodes	1	2	3	4	5	6	7	8
1	1.00	1.11	1.44	1.33	1.33	0.89	1.00	0.78
2	0.90	1.00	1.30	1.20	1.20	0.80	0.90	0.70
3	0.69	0.77	1.00	0.92	0.92	0.62	0.69	0.54
4	0.75	0.83	1.08	1.00	1.00	0.67	0.75	0.58
5	0.75	0.83	1.08	1.00	1.00	0.67	0.75	0.58
6	1.13	1.25	1.63	1.50	1.50	1.00	1.13	0.88
7	1.00	1.11	1.44	1.33	1.33	0.89	1.00	0.78
8	1.29	1.43	1.86	1.71	1.71	1.14	1.29	1.00

(d) Comparison Based on Cost Minimization Objective Function

TABLE 4 VECTORS OF PRIORITIES ACCORDING TO OBJECTIVES

	Objective Function			
Nodes	Cost	Covering	Median	Center
1	0.133	0.094	0.106	0.113
2	0.120	0.120	0.120	0.113
3	0.092	0.171	0.142	0.145
4	0.100	0.130	0.146	0.142
5	0.100	0.140	0.128	0.157
6	0.150	0.130	0.140	0.131
7	0.133	0.130	0.123	0.112
8	0.171	0.084	0.094	0.087

the preferred sites if all of the criteria are of equal importance. However, if the cost considerations are twice as important as the other criteria (i.e., have a weight of 0.4 while the other three have a weight of 0.2), the overall priority vector becomes (0.116, 0.119, 0.129, 0.124, 0.125, 0.140, 0.126, 0.121), which suggests that Node 6 is the preferred node on the basis of this preference among the criteria. Note that this site is different from those selected on the basis of consideration of the individual objectives. Also note that although Node 6 is not the optimal site when we consider the individual objectives, it performs well compared with the other sites; therefore, the AHP approach has resulted in a logical choice considering all of the criteria.

To see the effects of inconsistency in judgments, we changed some of the cells in the comparison matrices and introduced some inconsistency in those matrices. However, the changes were made only to the extent that the overall consistency ratio for the matrices remained under 0.1, so that we still had relatively consistent matrices. With these new matrices, the priorities of some of the sites changed. However, the preferred sites remained the same. This suggests that AHP results are not sensitive with respect to consistency in judgments, and if the comparison matrices are relatively consistent, the AHP provides overall results similar to those provided when perfect consistency exists.

The single-facility location problem was an excellent tool to show the implementation of AHP. However, when we are dealing with multifacility location problems, the task is not so easy. In multifacility location problems we are dealing with combinatorial problems that may have numerous alternative solutions. One must note that AHP is not a tool for generating alternative good or optimal solutions; rather, it provides a framework for evaluating and ranking alternative solutions on the basis of multiple criteria. Therefore, in a multifacility location problem, the major issue is how to generate good alternative solutions that can later be evaluated using AHP.

A naive approach to the multifacility location problem can be as follows. Assume for the moment that we want to locate P facilities. We can first rank all of the individual sites according to all of the criteria (as we did in the single-facility

location problem); then the P sites with the highest ranks could be selected for location of facilities. This method is a type of greedy adding heuristic which will result in a solution quickly; however, there is no guarantee that we have the best possible set of sites. To obtain better results, the alternative solutions that are input to AHP must be combinations of P out of N sites that perform well with respect to the individual criteria.

To generate good alternative solutions that can be input to AHP, we can use single-objective mathematical optimization. This can be done in two ways. A simple approach is to find a number of optimal and near-optimal sites on the basis of the individual objectives. This can easily be achieved by solving a sequence of uniobjective mathematical programming problems. When a number of good solutions based on each objective are identified, they are combined in a global set of alternatives, which can then be evaluated and ranked using AHP as discussed previously.

For example, assume that we want to locate three facilities in a 15-node network using the same criteria as in the single-facility location problem. We first identify a set of optimal or near-optimal alternatives on the basis of the individual criteria. This can be achieved through an iterative process of solving the individual optimization problems, recording the optimal solution, and forcing it out of the solution space in the next iteration. Assume that the following sets of alternatives were identified:

- Cost considerations: {1, 4, 7}, {2, 4, 8}, and {4, 7, 8};
- Coverage criterion: {5, 6, 9}, {6, 8, 4}, and {1, 2, 5};
- Median criterion: {6, 7, 9}, {2, 4, 8}, and {1, 8, 11}; and
- Center criterion: {5, 6, 8}, {7, 8, 11}, and {4, 8, 9}.

Then the global set of alternatives to be considered would be the combination of all 12 alternative sets of sites.

A more involved approach is to identify the individual sets of alternatives as discussed. However, to find the global set of alternatives, consider all of the sites providing the individual alternatives and examine all possible combinations of these sites. For example, in the previous problem, the sites providing the individual alternatives are {1, 2, 4, 5, 6, 7, 8, 9, 11}. We can therefore consider all possible combinations of three of these sites. This results in 84 alternatives. This approach identifies many more alternatives, which in turn results in a more thorough evaluation of the alternative space, but if the set of sites is large, the number of alternatives to be considered becomes impractical. This approach is particularly useful when a relatively small number of sites provide all individual alternatives.

To do a preliminary evaluation of these approaches, we considered a two-facility location problem on the network of Figure 2, with all data and criteria being the same.

On the basis of the vector of overall priorities, the naive approach suggests that the two highest-ranked sites are Nodes 6 and 3 if all criteria are of equal importance. To implement the first alternative generation approach, we identified several alternatives under each criterion. These alternatives are as follows:

- Cost: {1, 6}, {1, 7}, {2, 6}, {2, 7}, and {6, 7};
- Coverage: {1, 5}, {1, 6}, {1, 7}, {2, 5}, {2, 6}, {2, 7}, {3, 5}, {3, 6}, and {3, 7};
- Median: {1, 7}, {2, 6}, {2, 7}, and {3, 7}; and
- Center: {1, 6}, {1, 7}, {2, 6}, {2, 7}, {3, 6}, and {3, 7}.

This results in 10 distinct alternatives, which were evaluated using AHP. On the basis of equal preference among the criteria, Nodes 2 and 6 were identified as the best locations.

To implement the second alternative generation approach, we used the set of sites providing all of the preceding alternatives. This set is {1, 2, 3, 5, 6, 7}. All combinations of two sites out of these six were identified and considered as possible alternatives. This resulted in 15 alternatives, which were examined using AHP. Again on the basis of equal preference among criteria, Nodes 2 and 6 were the best sites. Nodes 2 and 6 are one set of the optimal locations based on center and covering objectives. The optimal locations based on median and cost objectives are {2, 7} and {1, 6} or {6, 7}, respectively.

Although this problem is a small one and does not fully serve the purpose of evaluating these approaches, it provides useful insights into the applicability of the alternative generation approaches and the AHP. The power of AHP in multicriteria decision making is further realized where the presence of qualitative objectives complicates the application of traditional multiobjective optimization techniques, such as weighting and constraint methods.

CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

In this paper we presented an approach to dealing with location decisions in a multiobjective planning framework. The approach is based on AHP, which is a useful tool for multicriteria decision making. The approach was implemented in the context of a single- and a two-facility location problem. The example, although simple, provides useful insights into the multiobjective nature of location decisions and clearly shows that when location decisions are made in view of a single objective, the selected sites are likely to be inferior with respect to other objectives. In some cases, as the example indicates, the preferred location when considering all objectives may not even be the optimal location when considering any of those objectives individually.

The paper indicates that AHP is a promising approach in dealing with location decisions in a multiobjective planning framework. We presented a brief sensitivity analysis of the AHP results with respect to the numerical values representing the comparative judgments among alternative sites (the cells of the comparison matrices). More detailed sensitivity analyses and exploration of methods of generating these numerical values other than those presented in this paper are important areas of research. In particular, the implicit assumption regarding the preference structure among the various criteria must be examined, and the sensitivity of AHP results with respect to changes in this preference structure should be analyzed.

Alternative generation techniques that could provide better results should also be explored. The results of this approach should be further tested on a larger network and in a multifacility location problem context. These results should also be compared with those obtained from implementation of other multiobjective optimization techniques, such as weight-

ing and bounding schemes. Finally, the results of this paper clearly indicate the multiobjective nature of location decisions. Development of other tools enabling analysts to approach location problems from the point of view of several simultaneous objectives is yet another promising research area.

REFERENCES

1. M. W. Cooper. Modelling Nonlinearities in Multicriteria Locational Decisions. *Mathematical Modelling*, Vol. 8, 1987, pp. 207–208.
2. B. C. Tansel, R. L. Francis, and T. J. Lowe. A Biobjective Multifacility Minimax Location Problem on a Tree Network. *Transportation Science*, Vol. 16, No. 4, 1982, pp. 407–429.
3. D. J. Engberg, J. L. Cohon, and C. S. ReVelle. Multiobjective Siting Model of a Natural Gas Pipeline System. Proceedings of the Modelling and Simulation Annual Pittsburgh Conference, May 1–2, Pittsburgh, Pa., 1980, pp. 1593–1598.
4. G. R. Bitran and K. D. Lawrence. Locating Service Offices: A Multicriteria Approach. *Omega*, Vol. 8, No. 2, 1980, pp. 201–206.
5. K. G. Zagrafos. *Multiobjective Hierarchical Models for Locating Public Facilities on a Transportation Network: A Goal Programming Approach*. Ph.D. dissertation. University of Connecticut, Storrs, 1986.
6. M. S. Daskin and E. H. Stern. A Hierarchical Objective Set Covering Model for Emergency Medical Service Vehicle Deployment, *Transportation Science*, Vol. 15, 1981, pp. 137–152.
7. A. E. Haghani. Capacitated Maximum Covering Problem. Working paper. Department of Civil Engineering, University of Pittsburgh, Pittsburgh, Pa., 1988.
8. M. J. Hodgson. Hierarchical Facility Location Models for Primary Health Care Delivery in Developing Areas. Presented at Fourth International Symposium on Locational Decisions, ISOLDE IV, Namur, Belgium, 1987.
9. D. A. Schilling and H. Pirkul. The Capacitated Maximal Covering Location Problem with Backup Service. Presented at Fourth International Symposium on Locational Decisions, ISOLDE IV, Namur, Belgium, 1987.
10. G. Y. Handler and P. B. Mirchandani. *Location on Networks: Theory and Algorithms*. MIT Press, Cambridge, Mass., 1979.
11. B. C. Tansel, R. L. Francis, and T. J. Lowe. Location on Networks: A Survey. I. The P-Center and P-Median Problem. *Management Science*, Vol. 29, No. 4, 1983, pp. 482–497.
12. M. L. Brandeau and S. S. Chiu. An Overview of Representative Problems in Location Research. *Management Science*, Vol. 35, 1989, pp. 645–674.
13. C. Toregas, R. Swain, C. ReVelle, and L. Bergman. The Location of Emergency Service Facilities. *Operations Research*, Vol. 19, 1971, pp. 1363–1373.
14. G. N. Berlin and J. C. Liebman. Mathematical Analysis of Emergency Ambulance Location. *Socio-Economic Planning Science*, Vol. 8, 1971, pp. 323–328.
15. P. Kolesar and W. E. Walker. An Algorithm for the Dynamic Relocation of Fire Companies. *Operations Research*, Vol. 22, 1974, pp. 249–274.
16. W. E. Walker. Using the Set-Covering Problem to Assign Fire Companies to Fire Houses. *Operations Research*, Vol. 22, 1974, pp. 275–277.
17. D. R. Plane and T. E. Hendrick. Mathematical Programming and the Location of Fire Companies. *Operations Research*, Vol. 25, 1977, pp. 563–578.
18. J. M. Benedict. *Three Hierarchical Objective Models Which Incorporate the Concept of Excess Coverage to Locate EMS Vehicles or Hospitals*. M. S. thesis. Northwestern University, Evanston, Ill., 1983.
19. K. Hogan and C. ReVelle. *Concepts and Applications of Backup Coverage*. Operations Research Group Report Series. Report 84-05. Johns Hopkins University, Baltimore, Md., 1984.
20. R. L. Church and C. ReVelle. The Maximum Covering Location Problem. *Papers of the Regional Science Association*, Vol. 32, 1974, pp. 101–118.
21. V. L. Bennett, D. J. Eaton, and R. L. Church. Selecting Sites for Rural Health Workers. *Soc. Sci. Med.*, Vol. 16, 1982, pp. 63–72.
22. D. J. Eaton, M. S. Daskin, D. Simmons, B. Bulloch, and G. Jansma. Determining Emergency Medical Service Vehicle Deployment in Austin, TX. *Interfaces*, Vol. 15, 1985, pp. 96–108.
23. S. Belardo, J. Harrald, W. A. Wallace, and J. Ward. A Partial Covering Approach to Siting Response Resources for Major Maritime Oil Spills. *Management Science*, Vol. 30, 1984, pp. 1184–1196.
24. S. L. Hakimi. Optimum Locations of Switching Centers and the Absolute Centers and Medians of a Graph. *Operations Research*, Vol. 12, 1964, pp. 450–459.
25. J. R. Weaver and R. L. Church. Computational Procedures for Location Problems on Stochastic Networks. *Transportation Science*, Vol. 17, 1983, pp. 168–180.
26. P. Hansen, M. Labbe, and M. Minoux. The p-Center Sum and p-Median Max Location Problems. Presented at Fourth International Symposium on Locational Decisions, ISOLDE IV, Namur, Belgium, 1987.
27. M. B. Tietz and P. Bart. Heuristic Methods for Estimating the Generalized Vertex Median of a Weighted Graph. *Operations Research*, Vol. 16, 1968, pp. 955–961.
28. R. L. Church, J. L. Cohon, and C. S. ReVelle. Energy Facility Siting by Multiobjective Location Analysis. *Proceedings of the Modelling and Simulation Annual Pittsburgh Conference*, Pittsburgh, Pa., 1978, pp. 1–10.
29. J. L. Cohon, C. S. ReVelle, J. Current, T. Eagles, R. Eberhart, and R. Church. Application of a Multiobjective Facility Location Model to Power Plant Siting in a Six-State Region of the U.S. *Computers and Operations Research*, Vol. 7, No. 1/2, 1980, pp. 107–123.
30. N. Mladineo, J. Margeta, J. P. Barns, and B. Mareschal. Multicriteria Ranking of Alternative Locations for Small Scale Hydro Plants. *European Journal of Operational Research*, Vol. 13, No. 2, 1987, pp. 215–222.
31. H. Min. A Multiobjective Retail Service Location Model for Fastfood Restaurants. *Omega*, Vol. 15, No. 5, 1987, pp. 429–441.
32. J. C. Fortenberry, A. Mitra, and R. D. M. Willis. A Multicriteria Approach to Optimal Emergency Vehicle Location Analysis. *Computers & Industrial Engineering*, Vol. 16, No. 2, 1989, pp. 339–347.
33. M. Heller, J. L. Cohon, and C. S. ReVelle. The Use of Simulation in Validating a Multiobjective EMS Location Model. *Annals of Operations Research*, Vol. 18, No. 1–4, 1989, pp. 303–322.
34. G. T. Ross and R. M. Soland. A Multicriteria Approach to the Location of Public Facilities. *European Journal of Operational Research*, Vol. 4, No. 5, 1980, pp. 307–321.
35. D. A. Schilling. Dynamic Location Modelling for Public Sector Facilities: A Multicriteria Approach. *Decision Sciences*, Vol. 11, No. 4, 1980, pp. 714–724.
36. P. Nijkamp and J. Spronk. Analysis of Production and Location Decisions by Means of Multicriteria Analysis. *Engineering & Process Economics*, Vol. 4, No. 2/3, 1979, pp. 285–302.
37. R. P. Mohanty and N. Rathnakumar. Multicriteria Location/Allocation Problem Analysis. *J. Inst. Eng. India Part ME 1*, Vol. 65, 1984, pp. 1–5.
38. H. Min. The Dynamic Expansion and Relocation of Capacitated Public Facilities: A Multiobjective Approach. *Computers and Operations Research*, Vol. 15, No. 3, 1988, pp. 243–252.
39. P. Nijkamp and J. Spronk. Interactive Multidimensional Programming Models for Locational Decisions. *European Journal of Operational Research*, Vol. 6, No. 2, 1981, pp. 220–223.
40. J. W. Hultz, D. D. Klingman, G. T. Ross, and R. M. Soland. An Interactive Computer System for Multicriteria Facility Location. *Computers and Operations Research*, Vol. 8, No. 4, 1981, pp. 249–261.
41. G. I. Green. A Multicriteria Warehouse Location Model. *International Journal of Physical Distribution & Materials Management*, Vol. 11, No. 1, 1981, pp. 5–13.
42. S. Eilon. Multicriteria Warehouse Location. *International Journal of Physical Distribution & Materials Management*, Vol. 12, No. 1, 1982, pp. 42–45.

43. S. M. Lee and R. L. Luebbe. The Multicriteria Warehouse Location Problem Revisited. *International Journal of Physical Distribution & Material Management*, Vol. 17, 1987, pp. 56–59.
44. H. U. Buhl. Axiomatic Considerations in Multiobjective Location Theory. *European Journal of Operational Research*, Vol. 33, No. 3, 1988, pp. 363–367.
45. T. L. Saaty. *The Analytic Hierarchy Process*. McGraw-Hill Book Company, New York, 1980.
46. T. L. Saaty. *Decision Making for Leaders, The Analytical Hierarchy Process for Decisions in a Complex World*. 1986.
47. T. L. Saaty. *Multicriteria Decision Making: The Analytic Hierarchy Process*. 1987.
48. T. L. Saaty. *Decision Making for Leaders* (second edition). 1988.
49. P. T. Harker and L. G. Vargas. The Theory of Ratio Scale Estimation: Saaty's Analytic Hierarchy Process. *Management Science*, Vol. 33, 1987, pp. 1383–1403.
50. F. Zahedi. The Analytic Hierarchy Process—A Survey of the Method and its Applications. *Interfaces*, Vol. 16, 1986, pp. 96–108.
51. P. T. Harker. State of the Art in the Analytic Hierarchy Process. Presented at ORSA-TIMS Joint National Meeting, Washington, D.C., 1988.

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

Forecasting Short-Term Demand for Empty Containers: A Case Study

MICHEL GENDREAU, TEODOR GABRIEL CRAINIC, PIERRE DEJAX, AND HÉLÈNE STEFFAN

Issues related to the modeling and estimation of short-term demand for empty containers for subsequent movements on international shipping lines are explored. The study is part of a larger research effort directed toward the development of models, methods, and integrated planning tools to address typical problems related to the management of the land distribution and transportation of containers. The study is based on actual operational data coming from a large international maritime shipping company. The information and the extensive manipulations required to obtain a suitable data set for statistical analyses are described, including analyses of both daily and weekly demand for various combinations of container category and customer aggregation. The difficulties and requirements associated with forecasting short-term demand for containers are discussed.

A study aimed at the modeling and estimation of short-term demand for empty containers within the land networks of international shipping lines is presented. The study is part of a larger research effort directed toward the development of models, methods, and integrated planning tools to address typical problems encountered by international shipping companies, which operate large-scale maritime and land networks, in the management of their container fleet and their land distribution and transportation operations.

The planning of these operations is an extremely complex activity, especially if the aim is to simultaneously optimize the cost and service aspects of the company's operations in a competitive environment. Crainic et al. (*1*) describe the various operations and planning issues involved and present the methodological framework that we propose to address them. To facilitate understanding of the context of the demand study, we present a brief overview of operations and the proposed methodology.

Arriving ships carry containers, which come in several sizes and types and are loaded with imported goods, and empty containers returning from previous exports. Loaded containers are moved to their final destinations (import customers), and the empty containers are dispatched wherever they are needed for subsequent exports. Once unloaded, empty containers at the customer's site return either to the port of origin or to another depot. On the other hand, exporting customers require empty containers. Once loaded, containers are transported to the port and loaded on ships together with empty containers sent abroad to cope with the worldwide supply-demand imbalance. Note the importance of empty-container movements. First, every commercial (profitable) movement of a loaded container generates, almost automatically, a nonprofitable empty-container movement. Second, significant regional imbalances between imports and exports result in empty containers being moved over relatively long distances, usually directly between depots. Thus, up to 40 percent of land container traffic is made up of movements of empty containers, which represent a significant portion of the total system cost (*2,3*).

The methodology proposed to improve the management of empty-container movements is based on an integrated multilevel approach, which reflects the observed hierarchy in the decision process and the flow of information. Its first main component is a strategic-tactical model, formulated as a multimode, multicommodity location-distribution problem with interdepot balancing requirements [(*4*); see Crainic et al. (*1*) for references concerning algorithms developed for this model]. The output of this model consists of the set of depots to be used for the duration of the planning horizon (several months to a year), the customer-to-depot allocation rules, and the main interdepot empty-container balancing flows.

The second component of our method corresponds to the level of the operational (day-to-day) planning of the company's activities. At this level, demand must be satisfied and the most effective routes and means of transportation must be selected and used. Two models are used: an empty-container allocation model and an empty- and loaded-container routing model. The allocation model aims to determine the "best" distribution of empty containers that satisfies known and forecast customer demands at lowest total system cost; it is formulated as a stochastic, dynamic network model (*5*). The routing model strives to minimize the overall transportation cost of the loaded and empty containers from their origin to their destination while ensuring on-time delivery (*6*).

Demand data, especially short-term forecasts, are essential for these operational models because, besides regular customers with sufficiently well-known behavior (and possible long-term contracts), a significant part of the total service requests come from irregular customers, who order rarely, or at irregular intervals, or both. Yet, to our knowledge, there does not exist any model with which short-term (daily or

M. Gendreau, Centre de recherche sur les transports, Université de Montréal, Montréal, Canada. T. G. Crainic, Centre de recherche sur les transports, Université de Montréal and Département des sciences administratives, Université du Québec à Montréal, Montréal, Canada. P. Dejax, Laboratoire Economique, Industriel et Social, Ecole Centrale Paris, Chatenay-Malabry, France. H. Steffan, Centre de recherche sur les transports, Université de Montréal, Montréal, Canada and Laboratoire Economique, Industriel et Social, Ecole Centrale Paris, Chatenay-Malabry, France.

weekly) container demand can be estimated. Hence, we have initiated an empirical investigation aimed at determining possible formulas for the average demand and the associated probability distributions. Ultimately, we look for models that may help forecast demand by container type and geographical zone, for a very short horizon. This paper gives an account of the first phase of this study, which was dedicated to the exploration of the available information, the construction of a reliable data set, and its statistical analysis. The plan of the paper parallels this sequence.

AVAILABLE DATA

The initial data available for this study come from a company operating worldwide maritime shipping lines to and from some 20 major European ports while servicing customers throughout most of Western Europe. The data correspond to the land distribution and transportation operations of the company for 1986. We concentrate on a network covering France, parts of Germany (corresponding to the former Federal Republic), the Netherlands, Belgium, and Luxemburg. This area corresponds to about 80 percent of the total container movements the company performed in Europe in that particular year.

The 224,374 records of the data file stand for as many land movements of one or several loaded or empty containers. Among the various pieces of information recorded for each movement, the most significant for the demand study are the following:

- Transportation mode identifies the actual mode used for transportation and the type of movement. The two main transportation means are private trucking and railways, including multimodal truck-rail combinations. Movements may be import or export, loaded or empty, and commercial or technical (balancing flows, movements of damaged or rented containers, etc.).
- Commercial mode specifies the type of contract between company and customer. Two modes are prevalent: carrier haulage (70 percent of the performed movements), in which the company manages and pays for the transportation of the containers; and merchant haulage (about 20 percent of the performed movements), in which the customer is entirely in charge of moving the containers and only the loaded movements are registered (with zero cost) in the company's records. The remaining traffic moves under mixed mode agreements, in which the company and the customer are each responsible for a part of the trip.
- The container category used for a given movement is identified by a combination of size and type characteristics. In 1986, the company used containers of 20 different categories (see Table 1). Note that, with the exception of the "bulk" and $9m^3$ containers, the company used exclusively 20′ and 40′ containers (in Europe, a truck is 20′ or 40′ long, whereas a rail car measures 60′). Distinctions among these containers are then induced by the particular usage (shipping line or operational characteristics) they are intended for.

Several manipulations of this initial information had to be performed to obtain a data set suitable for demand analyses (7). After cleaning up the file, we identified the origins and destinations of movements as customers or depots. We then determined a customer (depot) zone for each customer (depot) by selecting some 300 points where container traffic is most intense and building zones around these points that are consistent both geographically and commercially. Thus, because no aggregation of records has been performed, we ensure that, in most cases, a meaningful number of occurrences (observations) for each possible movement type are available for statistical analyses. The next step consists in estimating the daily supply of and demand for empty containers, for each customer zone and container category.

TABLE 1 DESCRIPTION OF CONTAINER CATEGORIES

Size	Type	Designation	Particularities
05	00	$9m^3$ containers	one 20′ = 3 $9m^3$
20	00	20′ general	All heights
22	00	20′ general	8′6 height (Zeebruge)
43	00	40′ general	
20	51	20′ open top	
40	51	40′ open top	
20	60	20′ flat	
45	00	40′ flat	
20	40	20′ isotherm	Not allowed on French Caribbean lines
42	40	40′ isotherm	Not allowed on French Caribbean lines
22	49	20′ isotherm	Dedicated to French Caribbean lines
42	49	40′ isotherm	Dedicated to French Caribbean lines
22	32	20′ refrigerated	
43	32	40′ refrigerated	
20	72	20′ tanker	
41	CT	40′ tanker	
49	RW	ACL trailer	ACL shipping line
VR		bulk	
NU		empty trailer	Truck movement without a container
22	80	20′ special bulk	
22	02	20′ open side	

Two approaches may be used to estimate the supply of and demand for empty containers. The first approach counts the empty containers that arrive (the empty customer demand) at and leave (the empty customer supply) each customer zone. This approach cannot be applied to movements performed under the merchant haulage commercial arrangement. The second method is based on the fact that each time a customer ships a loaded container, the company first had to deliver an empty one and, symmetrically, each loaded container delivered to a customer has to be picked up later on and moved away empty. Thus, the second approach counts the loaded containers that leave (the empty customer demand) and arrive (the empty customer supply) at each customer site.

Table 2 gives the results of the two methods for our data set. Both methods yield approximately the same figures. Yet, the total supply-demand is higher when the loaded method is used, which indicates that not all empty movements have been originally recorded in the company's data base. Consequently, we based our demand study on the daily estimates obtained by using this approach.

There are two sources of potential difficulties in the analysis of data and the interpretation of results. First, as emphasized by the yearly figures given in Table 2, not all container types are equally used. In fact, some are rarely called for. Consequently, 1 year's data may not be sufficient to obtain statistically significant results for some container types. Second, the available data do not reflect demand but rather the actual operations: the performed movements of loaded containers and some of the empty ones. Hence, the date of the request for a given movement is not recorded in the available data, and we miss the refused demand, as well as the container

TABLE 2 SUPPLY-DEMAND EVALUATIONS

Category	Carrier Haulage and Mixed Mode				Merchant Haulage	
	Empty		Loaded		Loaded	
	Demand	Supply	Supply	Demand	Demand	Supply
0500	29	8	8	27	12	505
2000	14183	6414	11361	18302	4674	13478
2040	410	2478	2920	416	628	193
2051	1004	482	520	1024	19	478
2060	223	90	93	242	80	439
2072	351	438	440	352	303	149
2200	3102	2732	3463	2931	549	1803
2232	1066	836	855	1080	378	1502
2249	6229	4568	6734	7181	14446	4320
2280	9	0	0	0	0	0
4051	1029	174	208	1013	10	157
41CT	3	11	11	3	1	0
4240	120	38	47	114	18	5
4249	1564	974	1346	1649	1950	746
4300	6465	2666	4117	8132	1137	2589
4332	1214	876	1015	1348	31	303
4500	581	56	36	560	7	1021
49RW	115	40	43	117	0	0

substitutions the company had to make to satisfy demand when the particular container category requested was not available. Furthermore, it has not been possible to establish exactly the meaning of the temporal information recorded with each movement. (At that time, the company was contracting out all its computer-related operations and was not using or validating in detail its past operational data.) On the basis of discussions with the company's management, we assumed that it indicates the date when the movement took place. This hypothesis is corroborated by the statistical analyses we performed. However, it may also indicate, at least for some movements, the date when the operation was recorded, thus introducing a certain level of uncertainty in our analysis.

In spite of these problems, it was decided to perform the analyses on the movement data, interpreting movements as demand, because this is the only information available. Furthermore, this is the type of data likely to continue to be available, because the planned management information system of the company was still meant to record the actual performed movements.

STATISTICAL ANALYSES

In spite of the aforementioned uncertainties regarding the exact timing of customer requests for empty containers, statistical analyses were performed on the final demand data set. They included analyses of both daily and weekly demand for various combinations of container categories (also called "product" in this section) and customer aggregation.

Daily Demand Analyses

These analyses were motivated by the fact that, in many fields of freight transportation, demand follows weekly cyclic patterns (*8*). More specifically, we were interested in determining whether a model of the following form could be fitted to the observed data:

$$D^p_{aw} = D^p_w \theta^p_d$$

where

D^p_{aw} = demand for Product p on Day d of Week w,
D^p_w = demand of Product p in Week w, and
θ^p_d = day-of-the-week adjustment factor for Product p and Day d.

Plotting demands for every day of the week over the 52 weeks of the year did not yield any clear pattern, whether aggregate demands for a product or demands for a specific container category–customer zone combination were considered, apart from the fact that observed values for Saturdays and Sundays were much lower.

To further investigate the impact of the day of the week on daily demands, linear regressions with dummy variables associated with the days of the week were performed on the data streams with the largest observed values (i.e., the aggregate demands per product and the demands for 24 product-zone pairs). In these regressions, the Saturday and Sunday values of each week were added together, yielding six observations per week for each data stream. The general form of the resulting regression equations was

$$D = aX_1 + bX_2 + cX_3 + dX_4 + eX_5 + f$$

where D is the daily demand for a given product or product-zone pair; X_i is 1 if the demand occurs on the ith day of the week and 0 otherwise; and a, b, c, d, e and f are the regression coefficients to be estimated.

The results of these regressions are summarized in Tables 3 and 4. For Table 3, it must be pointed out that for Product 0500 none of the independent variables X_i was sufficiently correlated with the dependent variable D to pass the tolerance tests of the stepwise regression method; thus in this case the regression is not significant. For the other products, the F-ratios indicate that the regressions are globally significant (the last column of the table gives the probability that the regression as a whole is not significant), but the R^2 coefficients, which correspond to the proportion of the total variation explained by the model, are not large. For the product-zone pairs (Table 4), though almost all regressions can be considered as significant, the R^2 coefficients are even smaller, which indicates a fairly poor fit.

TABLE 3 REGRESSION RESULTS ON UNNORMALIZED AGGREGATE DATA

Category	R^2	F-ratio	Probability
0500			
2000	41%	32.49	0.0000
2040	10%	7.18	0.0000
2051	4%	13.13	0.0003
2060	1%	4.63	0.0323
2072	14%	9.87	0.0000
2200	29%	24.86	0.0000
2232	9%	15.43	0.0000
2249	55%	76.54	0.0000
4051	16%	12.05	0.0000
4240	2%	7.21	0.0076
4249	39%	38.57	0.0000
4300	47%	53.61	0.0000
4332	20%	19.02	0.0000
4500	11%	7.89	0.0000
49RW	7%	7.70	0.0001

TABLE 4 REGRESSION RESULTS ON UNNORMALIZED DISAGGREGATE DATA

Category	Zone	R^2	F-ratio	Probability
2000	105	8%	5.54	0.0001
2000	107	13%	9.32	0.0000
2000	140	9%	6.47	0.0001
2000	154	6%	3.55	0.0039
2000	155	6%	3.81	0.0023
2000	210	8%	5.12	0.0002
2000	220	4%	2.27	0.0472
2000	222	4%	2.80	0.0268
2000	224	6%	3.85	0.0022
2000	237	17%	12.54	0.0000
2000	239	7%	4.84	0.0003
2232	208	12%	7.80	0.0000
2249	9	2%	1.60	0.1753
2249	140	23%	17.64	0.0000
2249	172	17%	12.49	0.0000
2249	192	8%	5.34	0.0001
2249	237	20%	15.27	0.0000
2249	274	33%	29.04	0.0000
2249	275	11%	7.59	0.0000
4300	107	7%	4.84	0.0003
4300	140	15%	10.69	0.0000
4300	208	21%	16.32	0.0000
4300	210	13%	9.13	0.0000
4300	239	2%	0.93	0.4586

Obviously, seasonal patterns could occur; if such was the case, they would probably be responsible for the large amount of unexplained variation. To account for such patterns, new data were created by dividing daily demands by weekly demands (the new observations thus represent the proportion of weekly demand taking place on a given day). The corresponding regression equation is then

$$\frac{D}{D_w} = a'X_1 + b'X_2 + c'X_3 + d'X_4 + e'$$

where D_w is the weekly demand and D, $X_1 \ldots, X_4$ are defined as before. (One independent variable had to be eliminated, because the observed values for every week now sum to 1.)

The results of these regressions for the global demands per product are given in Table 5. Surprisingly, the normalization of daily demands does not produce much larger R^2 values overall, and there are now two container categories for which the regressions are not significant (0500 and 4240).

TABLE 5 REGRESSION RESULTS ON NORMALIZED AGGREGATE DATA

Category	R^2	F-ratio	Probability
0500			
2000	42%	43.66	0.0000
2040	11%	7.42	0.0000
2051	16%	11.86	0.0000
2060	4%	5.89	0.0031
2072	13%	9.32	0.0000
2200	33%	29.99	0.0000
2232	17%	12.29	0.0000
2249	54%	73.41	0.0000
4051	15%	18.58	0.0000
4240			
4249	39%	39.84	0.0000
4300	50%	61.46	0.0000
4332	34%	31.30	0.0000
4500	12%	8.65	0.0000
49RW	8%	13.06	0.0000

It was pointed out earlier that demands on Saturdays and Sundays are, not unexpectedly, much lower. Because this could have had an impact on the previous analyses, it was decided to discard these observations and to perform a new series of regressions similar to the first one for the global demands per container category (i.e., with unnormalized values). These regressions yielded much lower R^2 coefficients (for instance, for Product 2000 the R^2 goes down from 41 to 18 percent and for Product 2249 from 55 to 17 percent). This confirmed that, apart from the difference between working days and weekends, the day-of-the-week effect is not important. This led us to abandon the model initially proposed and to focus our attention on weekly demands.

Weekly Demand Analyses

Whereas our analyses of daily demands were aimed at the identification of repetitive cyclic patterns, a different approach had to be used for weekly demands. For one thing, the available data (1 year) are clearly insufficient to provide any insight into possible seasonal patterns in any of the data streams. Furthermore, preliminary analyses of the aggregate demands per container category indicated that, if seasonal patterns were indeed present, these would be different from one product to another, thus making it useless to try to identify a general pattern for all categories. On the other hand, the observed variations in the demands may simply correspond to plain randomness in the underlying processes. To test this hypothesis, we first tried to fit normal and Poisson distributions to the demands per container category. The parameters used for the postulated distributions were the sample mean and standard deviation for the normal and sample mean for the Poisson. The goodness-of-fit test was the Kolmogorov-Smirnov test, which compares the observed cumulative distribution with the postulated cumulative distribution on the basis of the most extreme absolute difference (MEAD) between them. In fact, in this test, the MEAD is transformed into a standard normal statistic (Z value), from which the probability that the observed data follow the postulated distribution can be derived.

The results of these tests (see Tables 6 and 7) indicate that the normal distribution provides a good fit for four products (2000, 2200, 2249, and 4300) and a reasonable fit for two (2072 and 4051), whereas the demands for Container Categories 0500 and 4249 are approximately Poisson. It is interesting to note that the four container categories for which a good fit was obtained with the normal distribution are those with the largest sample means; there is thus a high level of aggregation of individual demands for these products and, therefore, this result could be expected.

The Kolmogorov-Smirnov test considers all observations as a static sample, neglecting any serial autocorrelation within the data streams. This is important because the presence of significant autocorrelations would invalidate the assumption of plain randomness. When we computed the autocorrelations and the partial autocorrelations for the 16 series, we found that there were no significant autocorrelations for 9 (Products 2000, 2051, 2060, 2072, 2232, 2249, 4051, 4240, and 49RW). It is thus possible to conclude that the series corresponding to Container Categories 2000, 2072, 2249, and 4051 are indeed

TABLE 6 KOLMOGOROV-SMIRNOV TEST: NORMAL DISTRIBUTION

Category	Mean	Standard Deviation	MEAD	Kolmogorov Smirnov Z	Probability
0500	0.53	1.03	0.364	2.597	0.000
2000	354.10	58.30	0.053	0.381	0.999
2040	8.08	5.11	0.153	1.094	0.182
2051	19.75	11.46	0.159	1.136	0.151
2060	4.51	4.91	0.205	1.467	0.027
2072	6.84	3.82	0.106	0.760	0.610
2200	57.18	20.90	0.066	0.469	0.980
2232	20.84	7.83	0.139	0.993	0.277
2249	138.69	23.12	0.076	0.541	0.932
4051	19.67	7.50	0.099	0.709	0.696
4240	2.18	2.46	0.215	1.534	0.018
4249	32.08	8.01	0.143	1.019	0.250
4300	158.27	27.57	0.069	0.494	0.968
4332	26.18	13.79	0.176	1.255	0.086
4500	10.82	6.29	0.116	0.830	0.496
49RW	2.27	2.59	0.218	1.556	0.016

TABLE 7 KOLMOGOROV-SMIRNOV TEST: POISSON DISTRIBUTION

Category	Mean	MEAD	Kolmogorov Smirnov Z	Probability
0500	0.53	0.078	0.555	0.918
2000	354.10	0.290	2.072	0.000
2040	8.08	0.136	0.973	0.300
2051	19.75	0.231	1.648	0.009
2060	4.51	0.253	1.808	0.003
2072	6.84	0.143	1.022	0.247
2200	57.18	0.263	1.878	0.002
2232	20.84	0.182	1.298	0.069
2249	138.69	0.186	1.330	0.058
4051	19.67	0.168	1.203	0.111
4240	2.18	0.169	1.207	0.108
4249	32.08	0.078	0.558	0.915
4300	158.27	0.226	1.617	0.011
4332	26.18	0.358	2.569	0.000
4500	10.82	0.198	1.414	0.037
49RW	2.27	0.193	1.376	0.045

"white noise" (i.e., sequences of independent identically distributed normal random variables).

Further analysis was required for the seven other container categories. This was done by using the well-known Box-Jenkins method, which allows characterization of sequential dependencies in time series through autoregressive integrated moving average (ARIMA) models (*9*). For two of the container categories (2040 and 4300), it was impossible to identify satisfactory models. Models were identified for the five remaining products, but they all displayed a high residual variance-to-mean ratio, which would make them almost useless in practice for predictive purposes (*7*).

We also applied the Box-Jenkins method to the disaggregate series (i.e., the series for the container category–customer zone pairs). For practical reasons, this analysis was restricted to the 24 pairs with annual demand larger than 300 containers. For these disaggregate series, the Box-Jenkins method yielded the following results (*7*):

- For 10 series, no ARIMA model could be identified.
- Four series displayed no significant autocorrelations, and it was possible to fit a normal distribution for three of them and a Poisson for the other.
- ARIMA models were identified and fitted for the other 10 series; however, these models were quite different from one another and, in most cases, the residual variances were large.

Overall, the results obtained for the disaggregate series confirm what could be suspected after the analysis of the aggregate series: the processes underlying the demand for empty containers are complex and their combined effect cannot be easily characterized in statistical terms, except when the level of aggregation is very high (that would be the case in the global demands for Products 2000, 2072, 2249, and 4051).

Discussion of Results

Several general remarks are in order. First, with only one year of data, it is not possible to detect any overall trend in the demands or to identify any seasonal patterns. Second, container categories with low traffic display high coefficients of variation, even for the weekly demands. This may suggest the use of different probability distributions, such as the negative binomial or the gamma, to represent the demand in these cases. Third, some external factors have an important effect on observed demands. For one thing, demands for empties are certainly driven to some extent by ship schedules, because a request for a container probably occurs a few days before the departure date of the ship on which it will sail. With regard to daily variations in demand, legal holidays and "traditional" vacation periods (such as most of July and August in most of Western Europe) must be taken into account. When we plotted the daily demands, we were easily able to pick up the dates of all major holidays on the graphs: in each case, there was a significant dip in demand. Given the large number of such holidays in the countries covered by the study, this in itself may have been sufficient to throw off the time series analyses.

CONCLUSIONS

The objective of the empirical study described in this paper was to gain some insight into the short-term demand process for empty containers and, if possible, to derive demand forecasting models.

We were successful in constructing a data set on demand for empties by using recorded information on loaded container movements. This data set, though not perfect, was adequate to allow for statistical analyses of demand.

The analyses indicated that predicting short-term demand for empty containers is extremely difficult. This confirms the conclusions of similar studies performed for the rail mode (*10,11*). However, when traffic is consistently large, we were able to fit probability distributions for the demand of specific container categories. The distributions can certainly be directly used in short-term planning models.

The situation is different for container categories with low traffic. For them, the available data were not sufficient to estimate distributions by using straightforward statistical techniques. Good predictions require more than 1 year of data, and all elements that may significantly affect the demand process in the specific case under study must be taken into account: ship schedules, holidays, vacations, substitution rules, and so forth. Note, however, that the lack of reliable demand distributions for these container categories is not a critical

issue in short-term planning. This is because conservative estimates of buffer stocks at depots can be derived from the means and variances of observed demand. Moreover, given the low level of demands (and stocks) for these containers, the inventory costs associated with overestimating demand form a small portion of total system costs.

ACKNOWLEDGMENTS

The authors are grateful to the Canadian National Science and Engineering Research Council, to the Fonds F.C.A.R. of the Province of Québec, and to the Programme de coopération France—Québec, who have provided the financial support for this research.

REFERENCES

1. T. G. Crainic, M. Gendreau, and P. J. Dejax. Modelling the Container Fleet Management Problem Using a Stochastic Dynamic Approach. In *Operational Research '90*, (H. E. Bradley, ed.), Pergamon Press, 1991, pp. 473–486.
2. P. Dejax, T. G. Crainic, L. Delorme, M. Bloise, E. de Tocqueville, and J. Hodgson. *Planification tactique du transport terrestre des conteneurs vides*. Ecole Centrale de Paris, Paris, 1987.
3. P. J. Dejax, T. G. Crainic, and L. Delorme. *Strategic/Tactic Planning of Container Land Transportation Systems*. Joint Meeting TIMS XXVIII/EURO IX, Paris, 1988.
4. T. G. Crainic, P. J. Dejax, and L. Delorme. Models for Multimode Multicommodity Location Problems with Interdepot Balancing Requirements. *Annals of Operations Research*, Vol. 18, 1989, pp. 279–302.
5. T. G. Crainic, M. Gendreau, and P. J. Dejax. *A Dynamic Stochastic Model for the Allocation of Empty Containers*. Publication 713. Centre de recherche sur les transports, Université de Montréal, Montréal, Québec, Canada, 1990.
6. T. G. Crainic, P. J. Dejax, M. Gendreau, and F. Benamar. *Multimodal Multiproduct Combined Vehicle Routing—Traffic Itinerary Generation*. ORSA/TIMS Joint National Meeting, Anaheim, Calif., 1991.
7. H. Steffan. *Etude de la demande de conteneurs vides a court terme*. Publication 651. Centre de recherche sur les transports, Université de Montréal, Montréal, Québec, Canada, 1989.
8. G. J. Beaujon and M. A. Turnquist. A Model for Fleet Sizing and Vehicle Allocation. *Transportation Science*, Vol. 25, No. 1, 1988, pp. 19–45.
9. G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.
10. W. K. Minger, D. J. Williams, and M. B. Hargrove. *Freight Car Demand Information and Forecasting Research Project. Phase I*. Report FRA-OPPD-76/9. Association of American Railroads, Washington, D.C., 1975.
11. W. K. Minger and M. B. Hargrove. *Freight Car Demand Information and Forecasting Research Project. Phase II*. Report FRA-OPPD-77/5. Associations of American Railroads, Washington, D.C., 1977.

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