TRANSPORTATION RESEARCH RECORD 895

Travel Demand Forecasting Procedures 1982

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY OF SCIENCES
WASHINGTON, D.C. 1982

Transportation Research Record 895
Price $7.80
Edited for TRB by Susan Singer-Bart

modes
1 highway transportation
2 public transit

subject areas
12 planning
13 forecasting

Library of Congress Cataloging in Publication Data
National Research Council. Transportation Research Board.
Travel demand forecasting procedures, 1982.

(Transportation research record; 895)
1. Choice of transportation–Forecasting–Congresses.
2. Transportation planning–United States–Congresses.
3. Urban transportation–United States–Planning–Congresses.
I. National Research Council (U.S.). Transportation Research Board. II. Series.
TE7.H5 no. 895 [HE336.C5] 380.5s [388.4] 83-8104
ISBN 0-309-03507-4 ISSN 0361-1981

Sponsorship of the Papers in This Transportation Research Record

GROUP 1–TRANSPORTATION SYSTEMS PLANNING AND ADMINISTRATION
Kenneth W. Heathington, University of Tennessee, chairman

Transportation Forecasting Section
George V. Wickstrom, Metropolitan Washington Council of Governments, chairman

Committee on Passenger Travel Demand Forecasting
David S. Gendell, Federal Highway Administration, chairman
Moshe E. Ben-Akiva, Daniel Brand, David J. Dunlap, Robert T. Dunphy, Raymond H. Ellis, Robert E. Gall, Thomas F. Golob, Walter G. Hansen, David T. Hartgen, Thomas J. Hillegass, Joel L. Horowitz, Stephen M. Howe, Lidia P. Kostyniuk, T. Keith Lawton, Steven Richard Lerman, Eugene J. Lessieu, Frederick A. Reid, Martin G. Richards, Gordon W. Schultz, Gordon A. Shunk, Bruce D. Spear, Anthony R. Tomazinis, Edward Weiner, Yacov Zahavi

Committee on Traveler Behavior and Values
David T. Hartgen, New York State Department of Transportation, chairman
Peter M. Allaman, Julian M. Benjamin, Katherine Patricia Burnett, Melvyn Cheslow, David Damm, Michael Florian, Thomas F. Golob, David A. Hensher, Joel L. Horowitz, Frank S. Koppelman, Steven Richard Lerman, Jordan J. Louviere, Arnim H. Meyburg, Richard M. Michaels, Alfred J. Neveu, Robert E. Paaswell, Earl R. Ruiter, Yosef Sheffi, Peter R. Stopher, Antti Talvitie, Mary Lynn Tischer

James A. Scott, Transportation Research Board staff

Sponsorship is indicated by a footnote at the end of each report. The organizational units, officers, and members are as of December 31, 1981.

Contents

Authors of the Papers in This Record

Aunet, Bruce, Division of Planning and Budget, Wisconsin Department of Transportation, P.O. Box 7913, Madison, WI 53707

Ben-Akiva, Moshe, Department of Civil Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

Benham, Joy L., Charles River Associates, Inc., John Hancock Tower, 200 Clarendon Street, Boston, MA 02116

Brand, Daniel, Charles River Associates, Inc., John Hancock Tower, 200 Clarendon Street, Boston, MA 02116

Cunagin, Wiley D., Texas Transportation Institute, Texas A&M University, College Station, TX 77843

Gur, Yehuda, Transportation Research Institute, Technion—Israel Institute of Technology, Haifa, Israel

Hyman, William, Division of Planning and Budget, Wisconsin Department of Transportation, P.O. Box 7913, Madison, WI 53707

Kocur, George, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; formerly at Thayer School of Engineering, Dartmouth College

Koppelman, Frank S., Department of Civil Engineering, Transportation Center, Northwestern University, 200 Sheridan Road, Evanston, IL 60201

Kumar, Ashok, Southern California Rapid Transit District, 425 South Main Street, Los Angeles, CA 90013; formerly with the Northeast Ohio Areawide Coordinating Agency

Litinas, Nicolaos, Athens Area Urban Transport Organization, Athens, Greece

Nihan, Nancy L., Department of Civil Engineering, University of Washington, Seattle, WA 98195

Rogness, Ramey O., North Dakota State University of Agriculture and Applied Science, Fargo, ND 58105

Sheffi, Yosef, Department of Civil Engineering and Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, MA 02139

Sugiyama, Morihisa, Department of Civil Engineering and Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, MA 02139

Wilmot, Chester G., Transportation Center, Northwestern University, Nathaniel Leverone Hall, 2001 Sheridan Road, Evanston, IL 60201

Procedure for Estimating Freeway Trip Tables

NANCY L. NIHAN

A method for estimating freeway trip tables with volume data by using a gravity-based model is presented. The 1969 San Francisco Bay Area origin-destination survey is used to test the estimates. The results of using the gravity-based model are compared with the estimates obtained by using the SYNODM approach developed at the University of California, Berkeley. The gravity-based approach achieves a closer fit to the actual trip tables than does the current version of SYNODM.

This paper describes a method for estimating freeway trip tables by using ramp volume data and compares the resulting estimates with those obtained by using the SYNODM approach (an estimation procedure currently used by traffic planners). Below is a description of both the general problem of estimating trip tables by using volume data and the freeway-restricted problem. Also included is a summary of currently proposed estimation procedures and their limitations.

GENERAL PROBLEM

For the past few years there has been a surge of interest in developing a technique for estimating a trip matrix or origin-destination (O-D) matrix for urban areas by using street volumes as the primary source of knowledge. This is because the collection of trip origin-destination data is costly, time consuming, and less accurate than the more easily collected traffic volume data. Volume data have been collected in most cities on a regular basis for a number of years through the use of automatic traffic counters. However, current O-D data, which require extensive travel surveys of the urban population, are not available to today's transportation planner and are not likely to be available in the future for the majority of our urban areas.

The state of the art in the general problem area is summarized by Willumsen (1), who divides the estimation methods into three broad groups of models. These include gravity-based models, network-equilibrium models, and entropy-maximizing models. Gravity-based models assume that trips follow a gravity pattern. In the approaches considered so far this leads to linear or nonlinear regression solutions. The linear models (2-7), where tested, are used to forecast link volumes. They give acceptable estimates of link flows (errors of 20 percent or less). The nonlinear approaches (8-10) yield slightly better results. However, accuracy is still based on observed flows. Since a variety of O-D matrices can produce the same pattern of link volumes in a network, this is not a sufficient test of the accuracy of the trip tables.

The second group of general models, network-equilibrium models, is based on Wardrop's first principle. Such models (11-13) yield solutions that depend on the initial solution assumed. The solutions are, therefore, not unique and have not been verified adequately.

The last group of models uses an entropy-maximizing approach to find the most likely trip matrix compatible with observed flows. In one such model the solution depends on an a priori estimation of the O-D matrix. Other approaches (1, 15-17) circumvent this problem but do not consider the impact of distance or travel time on tripmaking behavior. All three approaches to the general problem have not been tested against actual O-D data, although some have been checked by using synthetic networks.

FREEWAY OR CORRIDOR PROBLEM

Unlike the general problem, the freeway or corridor problem lends itself to model verification. Also, in dealing with a restricted network of corridor flows, we avoid some of the problems created by the ubiquitous nature of vehicle travel. The natural constraints on freeway flows reduce the relative number of unknowns, although the problem is still underdetermined.

Since an O-D matrix of the freeway portion of vehicle trips is required for certain traffic-planning models such as the FREQ6PE simulation model (18), these data are of immediate interest to traffic and transportation planners. Yet the methods used to collect these data, although easier than comprehensive travel surveys, are still costly and time consuming. The most widely used method is the license plate survey, where observers are positioned at every ramp for a particular freeway segment and license plates of passing vehicles are recorded. These are then traced to determine points of vehicle entry and exit. Although the surveys are inconvenient, they have been conducted in most major cities and therefore provide a basis for verification of models that attempt to estimate the freeway trip matrix by using ramp volumes as a basis. If sufficiently accurate models can be developed for this problem, they can be used not only to help monitor existing traffic congestion problems but also to illuminate the general problem. Also, since volume data are available for different time intervals (e.g., hours of the day), accurate models could reproduce the changes in trip patterns over time that would be of benefit to transit and traffic planners.

Although some work on generating freeway O-D tables (ramp-to-ramp trip tables) from link volumes has been reported, the results in this area are inconclusive and the theoretical basis of the proposed models is weak. The primary activity in this area comes from the developers of FREQ6PE, a combination traffic simulation and ramp control optimization model. This model requires a ramp-to-ramp trip table for every time interval (e.g., 15-min intervals) for the period of study (e.g., peak-hour period). A computer model called SYNODM (19) has been developed to synthesize the required trip tables. It is a simple proportionality scheme that distributes off-ramp traffic to upstream on-ramps. Specifically, if we let

- M = set of all freeway entrances upstream of exit j,
- V_i^* = total trips originating at i that have not yet been assigned a destination and are upstream from j,
- D_j = total trips exiting at destination j, and
- T_{ij} = total trips originating at i and exiting at j.

Then,

$$T_{ij} = D_j \left(V_i^* / \sum_{\ell=1}^{M} V_\ell^* \right) \quad (1)$$

The trips are assigned beginning with the first upstream off-ramp and continuing to successive downstream off-ramps.

There are several problems with such a simplistic approach. The most noticeable is the implicit assumption that trip distance or travel time is not a factor in travel behavior. Thus, if there are 100 still unassigned vehicles that enter at on-ramp 1 (10 miles upstream from off-ramp j) and 100 vehicles enter at on-ramp 10 (1 mile upstream from the nearest off-ramp j), the number of vehicles from 1 to j and from 10 to j would be equal. Yet, intuitively one would expect that vehicles that enter at 10 would have a relatively low probability of getting off at the next exit. By the same token, vehicles that have already traveled 10 miles would have a relatively high probability of exiting at the next stop. Therefore, one could expect significant errors here in the form of overpredicting the number of very short trips and very long trips. The authors admit that the proportionality assumption is a crude approximation. In assessing the accuracy of their model they state that it "...does tend to distribute correctly 70-80 percent of the traffic in most cases, and in the absence of an O-D study that is probably a reasonable approximation" (20).

The level of error in those cases that are not correctly assigned is not discussed. If, as our intuition indicates, these errors could be substantial, the resulting O-D table is not valid. The developers are currently investigating other methods and have recently revised SYNODM to include known interchanges as inputs to improve accuracy (21,22). The gravity-based method proposed below also has this capability and includes an impedance factor as well.

RESEARCH DESIGN

A description of the proposed estimation procedure, the data base used for testing the trip estimates, and the error measurements used in comparing this procedure with the estimates obtained by using SYNODM is presented in this section.

Estimation Procedure

The procedure assumes a gravity-based model to be applied along a particular section of freeway. Three inverse impedance functions are calculated based on average trip distance along the section. (A separate curve was used for internal-internal or ramp-to-ramp trips, external-internal or mainstream-to-ramp trips, and internal-external or ramp-to-mainstream trips.) Since Voorhees (25) has shown that a gamma function is most appropriate for total trip patterns, a gamma function was assumed with adjustments made for external trip ends. Also assumed was that one could obtain a reasonably good estimate of through trips either by collection that used overpasses or from knowledge of previous O-D percentages.

The gravity-based model has the formulation shown below:

$$T_{ij} = O_i X_j f(d_{ij}) / \sum_j X_j f(d_{ij}) \quad (2)$$

$$\sum_i T_{ij} = D_j \quad (3)$$

$$f(d_{ij}) = [\beta^\alpha/\Gamma(\alpha)]\, d_{ij}^{(\alpha-1)} \exp(-\beta d_{ij}) \quad (4)$$

where

T_{ij} = trips from origin ramp i to destination ramp j,
O_i = trips that originate at ramp i,
D_j = trips that exit at ramp j,
d_{ij} = impedance of travel between i and j (e.g., distance or travel time),
f = inverse impedance function (i.e., travel propensity function), and
X_j = normalization factor for destination point j.

Given the impedance function, an iterative procedure is used to solve for X_j. This procedure has been shown to always converge to a unique solution (24,25).

Although several forms of inverse impedance functions have been tested in the past, Voorhees (25) showed rather conclusively that the gamma function gave the best fit when calibrating models to generate total travel matrices. This would appear to also apply to freeway trip tables because one expects a unimodal function to discourage both very short and very long trips. The shape parameter (α) was found by Voorhees to be approximately 1.5 for total travel for most cities. Since this was obtained for a total inverse impedance function, and since freeway travel can be expected to tolerate longer distances than can other types of trips, one would expect that a distribution less skewed to the left would be appropriate. A preliminary value for α of approximately 3 is suggested. (Note, for a sample experiment described later in this paper, values of the shape parameter were varied from 2 to 4 with minor differences in the resulting trip table.)

The size parameter (β) is equal to $\alpha/\bar{d}$ where $\bar{d}$ is the average impedance for the network. Since we know the total number of trips and the link volumes and impedances, this can be calculated as

$$\bar{d} = \sum_{k=1}^{K} v_k d_k / T \quad (5)$$

where

T = total trips (i.e., total number of origins or destinations),
v_k = volume on freeway subsection ℓ,
K = number of freeway subsections in study section, and
d_k = length of freeway subsection ℓ.

In determining the impact of origins farther upstream or downstream from the freeway section, a simple constant that is equal to the largest possible value for that inverse impedance function was chosen. Thus, for example, for a travel function with α = 3 and $\bar{d}$ = 5.0, a gamma function would be generated as shown in Figure 1. If one end of the trip originated at the mainstream-on point or ended at the mainstream-off point this constant (f*) would be used for short trips (trips that have study section length d* or less). Thus, the external-internal and internal-external functions would resemble the solid line in Figure 1 and the internal-internal function would be a strict gamma function.

Figure 1. Example of inverse impedance curve.

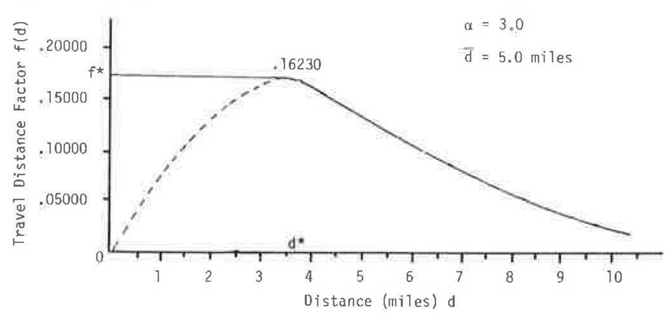


Table 1. Freeway subsection characteristics.

Sub-section No.	No. of Lanes	Length (ft)	Input	Subsection Description: Entry Point	Exit Point
1	5	1630	OD	Mainline origin	Powell off
2	5	1960		Powell off	Powell on
3	5	1550	OD	Powell on	Ashby off
4	4	1960		Ashby off	Ashby on
5	4	500	O	Ashby on	500-ft point
6	4	4790	D	500-ft point	University off
7	4	3030		University off	University on
8	4	2160	OD	University on	Gilman off
9	4	2030		Gilman off	Gilman on
10	5	1250	OD	Gilman on	Buchanan off
11	4	900	D	1 Buchanan off	Hoffman off, left
12	3	1320	D	Hoffman off	Pierce off
13	3	720		Pierce off	Pierce on
14	3	2610	OD	Pierce on	Central off
15	3	1660		Central off	Central on
16	3	1890	OD	Central on	Carlson off
17	3	2310		Carlson off	Carlson on
18	3	1460	OD	Carlson on	Potrero off
19	3	3800		Potrero off	Cutting on
20	4	1100	O	Cutting on	Grade change point
21	4	660	D	Grade change point	Macdonald off
22	4	1480	D	O Macdonald off	San Pablo off
23	3	1480		San Pablo off	San Pablo on
24	4	800	OD	San Pablo on	Solano off
25	3	4690	D	Solano off	San Pablo Dam off
26	3	2190		Dam Road off	Dam Road on
27	3	2320	OD	Dam Road on	Road 20 off
28	3	830		Road 20 off	Grade change point
29	3	1180		Grade change point	Road 20 on
30	3	2560	OD	Road 20 on	Mainline destination

Note: O = origin, D = destination.

In developing the above functions, an initial value of $\beta = \alpha/\bar{d}$ was chosen for all three curves. After each run of the model, new parameters were calculated for the external-internal (β_1), internal-internal (β_2), and internal-external (β_3) functions based on the average distances of these trips. Thus, for a freeway section that has N points of entry or exit,

$$\bar{d}_1 = \sum_{j \neq N} T_{1j} d_{1j} / O_1 \quad (6)$$

$$\bar{d}_2 = \sum_{i \neq 1} \sum_{j \neq N} T_{ij} d_{ij} / \sum_{i \neq 1} \sum_{j \neq N} T_{ij} \quad (7)$$

$$\bar{d}_3 = \sum_{i \neq 1} T_{iN} d_{iN} / D_N \quad (8)$$

where point 1 represents the mainstream-on-node and point N the mainstream-off node.

Assuming that T_{1N} is given, the estimation procedure is as follows:

1. Calculate $\bar{d}$;
2. Remove through trips (e.g., subtract T_{1N} from O_1 and D_N);
3. Letting $\alpha = 3$ and $\Gamma(\alpha) = 2$, calculate initial values of $\beta_1 = \beta_2 = \beta_3 = \alpha/\bar{d}$, where β_1, β_2, and β_3 are the size parameters for the three inverse impedance curves;
4. Run the gravity model;
5. For the run obtained in step 4, calculate $\bar{d}_1$, $\bar{d}_2$, and $\bar{d}_3$ and use these new estimates to calculate β_1, β_2, and β_3; and
6. Repeat steps 4-5 until the $\bar{d}_i$s used in the travel functions agree with the $\bar{d}_i$s calculated (usually one or two iterations).

Step 2 was suggested by Willis of the University of California, Berkeley.

Data Base

The experiment was performed for a subsection of the San Francisco Bay Area freeway network. The study corridor included the northbound portion of the Eastshore Freeway (I-80), beginning upstream from the Powell exit ramp and terminating downstream from the CA-20 entrance ramp. Table 1 describes this freeway section and its points of entry and exit. Volume data for five 15-min time slices starting at 3:45 p.m., October 1969, were used as the O-D inputs for the gravity-based model. An O-D survey performed during the same time periods was used for comparing the accuracy of gravity-based model outputs with those estimated by the SYNODM procedure.

Error Measures

Three types of error measures were considered. These were average absolute trip errors, average total percentage of trip errors, and average individual percentage of trip errors. These are defined below.

$$\text{Avg trip error} = \sum_{ij} |T_{ij}^c - T_{ij}^o| / N \quad (9)$$

$$\text{Avg total percentage trip error} = \sum_{ij} (|T_{ij}^c - T_{ij}^o| / T) \times 100 \quad (10)$$

$$\text{Avg individual percentage trip error} = \sum_{ij} [(|T_{ij}^c - T_{ij}^o| / T_{ij}^o) / N] \times 100 \quad (11)$$

where

T_{ij}^c = trips from i to j calculated,
T_{ij}^o = trips from i to j observed,
N = number of error values, and
T = total trips in cells used.

Cells that have less than five observed trips were ignored in the above calculations to avoid unreasonably high individual percentage errors associated with very few trips.

RESULTS

Observed and calculated trip tables that represent the 15-min slices from 3:45 to 5:00 p.m. were compared. SYNODM trip tables were also calculated. Figure 2 shows an example comparison for time slice 3. As expected, the SYNODM trip tables were more likely to overpredict short trips than was the gravity-based model. A summary of the average trip distances given in Table 2 bears this out. The average distance for internal trips for the SYNODM estimate is consistently shorter than that observed; however, the gravity-based approach is very close in its estimates of all three types of trip lengths.

Figure 3 shows an example comparison of the trip-length distributions for the actual and gravity-based calculated trips. For all five time slices these frequency distributions are very close.

A final comparison of observed, estimated, and SYNODM estimated trips is given in Table 3. In all three error measures, the gravity-based approach is consistently better. However, as total trip volumes increase (time slices 4 and 5) the SYNODM approach becomes competitive. As expected, the individual percentage of error is somewhat high due to relatively low numbers of trips in some cells.

INTERPRETATION OF RESULTS

Although the results of this experiment are by no means conclusive, distance impedance should be considered in estimating freeway trip tables, even for relatively short freeway subsections. If through

Table 2. Average trip distances along freeway section.

	Internal-Internal, $\bar{d}_2$			External-Internal, $\bar{d}_1$			Internal-External, $\bar{d}_3$			
Time Slice	Actual	Gravity-Based Estimated	SYNODM	Actual	Gravity-Based Estimated	SYNODM	Actual	Gravity-Based Estimated	SYNODM	$\bar{d}$
1	3.33	3.61	2.51	3.67	3.65	3.85	4.74	4.48	4.16	4.29
2	3.48	3.65	2.70	3.51	3.56	3.84	5.40	5.10	4.88	4.57
3	3.39	3.56	2.53	3.88	3.94	3.95	5.04	5.04	4.44	4.51
4	3.36	3.49	2.80	3.35	3.32	3.93	5.35	5.24	5.19	4.86
5	3.03	3.21	2.53	3.78	3.66	3.81	4.96	5.08	4.67	4.31

Figure 2. Freeway trip tables for time slice 3.

Observed Trips

190	285	250	59	175	310	10	124	52	45	46	126	32	94	35	201
	2		8		13		8		13	5		3	8	8	29
			3		19		8	4	6	7	2	7	19	10	65
					47		4	6	21	2		8	21	11	70
				15	11		6	6	29	2		7	21	11	72
										2		1	5	3	19
								2	4	2	2	3	9	4	34
									2	4		5	15	7	47
												12	33	16	109
												12	35	17	116
														8	52
															40

Gravity-Based Estimated Trips

190	286	233	63	163	338	8	117	51	82	43	77	39	102	43	200
	1	10	3	8	17		6	3	4	2	4	2	5	2	30
		8	4	12	26	1	10	5	8	4	7	4	9	4	51
			1	6	16	1	11	6	12	7	13	8	20	8	82
				1	3		7	5	11	8	15	9	24	11	88
								1	2	1	3	2	5	2	15
									2	2	5	4	11	5	32
										2	5	5	16	8	44
										1	3	11	36	20	99
												7	33	24	116
														3	57
															40

SYNODM Estimated Trips

190	273	218	54	133	281	7	103	46	73	37	69	41	112	57	356
	14	11	3	7	15		5	2	4	2	3	2	5	3	15
		21	5	13	27	1	10	4	7	3	6	4	10	5	31
			8	19	39	1	14	6	10	5	9	6	15	8	47
				19	39	1	14	6	10	5	9	5	15	7	46
							4	2	2	1	2	1	3	1	9
								4	6	3	5	3	8	4	25
									8	4	7	4	12	6	37
										10	18	11	29	15	91
												14	37	19	116
													14	7	42
															43

trips can be estimated or measured, the gravity-based approach looks reasonably accurate. If this were coupled with knowledge of one or two other interchange values, the results should approach observed values. For example, for some reason the number of observed vehicles getting on at Gilman and directly off at Buchanan is high in some time slices. If this were suspected in advance, trips from Gilman to Buchanan might be measured and used as an additional factor in calibrating the model (adjusting the travel distance factor for that particular interchange).

Most ramp interchanges that are closely spaced, however, do not exhibit this property. Thus, distance appears to be a factor and methods such as SYNODM, which do not take distance into account, will not do as well. The errors that result from this omission are higher than indicated by our error measures because cells that have four or fewer trips were dropped from the analysis. Consider, for example, the observed trip table for time slice three and the estimated tables shown in Figure 2. Because of the short time interval, several of the less frequented interchanges have no observed trips and were dropped before calculating the percentage error terms. Yet SYNODM in several cases estimates significant numbers of trips for these cells, thus the true error differences between the two procedures is even larger. In both cases, the predictions might improve if larger time intervals were used.

Note also that SYNODM overpredicts through trips. These represent very long trips, and this is another indication that distance is a factor. Even if through trips were assumed as known in the SYNODM procedure, other very long trips would probably be overpredicted.

Figure 3. Trip length distribution for time slice 3.

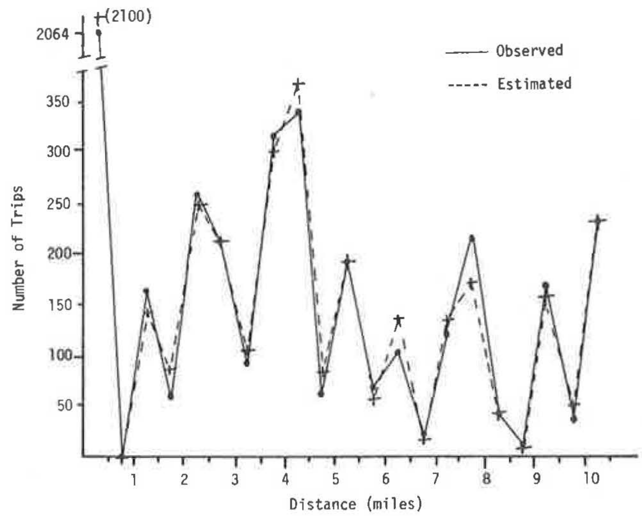


Table 3. Comparison of trip errors.

	Avg No. of Trips in Error		Avg Total Trip Error (%)		Avg Individual Trip Error (%)	
Time Slice	Proposed Method	SYNODM	Proposed Method	SYNODM	Proposed Method	SYNODM
1	5.9	9.1	12.6	15.4	21.1	21.7
2	8.1	10.0	17.0	22.2	33.1	34.9
3	6.6	12.8	12.1	23.9	18.4	31.5
4	11.4	11.5	20.7	21.5	39.2	40.2
5	8.3	8.5	17.1	17.4	27.5	27.6

SUMMARY AND CONCLUSIONS

Knowledge of the number of through trips on a freeway section appears sufficient to calibrate a reasonably accurate gravity-based trip-table estimator. Whether knowledge of the number of trips for any major O-D pair would work as well should be explored because overpasses may not always coincide with the freeway subsection under study. In an upcoming revision, SYNODM may also assume knowledge of through trips to improve its accuracy. However, without incorporation of an impedance factor, it may still have unreasonably high predictions for very short and very long trips.

In any event, further exploration of the accuracy of these techniques with other O-D data bases is needed to determine whether existing models are sufficiently accurate. The possibility of improving accuracy by obtaining data on one or more interchanges in lieu of a complete O-D survey should also be investigated.

ACKNOWLEDGMENT

This material is based on work supported by the National Science Foundation.

REFERENCES

1. L.G. Willumsen. Estimation of an O-D Matrix from Traffic Counts--A Review. Working Paper 99, Institute for Transport Studies, Univ. of Leeds, Leeds, England, Aug. 1978.
2. D. Low. A New Approach to Transportation Systems Modelling. Traffic Quarterly, Vol. 26, 1978, pp. 391-404.
3. R. Smith and W. McFarlane. Examination of a Simplified Travel Demand Model. Transportation Engineering Journal of ASCE, Vol. 104, No. TE1, 1978.
4. Urban Traffic Models: Possibilities for Simplification. OECD Road Research Group, Paris, 1974.
5. M. Gaudry and L. LaMarre. Estimating O-D Matrices from Traffic Counts: A Simple Linear Intercity Model for Quebec. Centre de Recherche, Univ. de Montreal, Publ. 105, 1978.
6. J. Symons, R. Wilson, and J. Paterson. A Model of Inter-City Motor Travel Estimated by Link Volumes. Proc., Australian Road Research Board, Vol. 8, Session 20, 1976.
7. J. Holm, T. Jensen, S. Nielsen, A. Christensen, B. Johnsen, and G. Ronby. Calibrating Traffic Models on Traffic Census Results Only. Traffic Engineering and Control, Vol. 17, April 1976.
8. P. Hogberg. Estimation of Parameters in Models for Traffic Prediction: A Non-Linear Regression Approach. Transportation Research, Vol. 10, 1976, pp. 263-265.
9. P. Robillard. Estimating an O-D Matrix from Observed Link Volumes. Transportation Research, Vol. 9, 1975, pp. 123-128.
10. M. Wills. Non Linear Estimation of Origin-Destination Flows and Gravity Model Parameters from Traffic Counts on Links. Department of Geography, Univ. of British Columbia, Vancouver, 1977.
11. S. Nguyen. Procedures for Equilibrium Traffic Assignment with Elastic Demand. Centre de Recherche sur les Transports, Univ. de Montreal, Publ. No. 39, 1976.
12. S. Nguyen. Estimating an O-D Matrix from Network Data: A Network Equilibrium Approach. Centre de Recherche sur les Transports, Univ. de Montreal, Publ. No. 87, 1977.
13. M. Turnquist and Y. Gur. Estimation of Trip Tables from Observed Link Volumes. TRB, Transportation Research Record 730, 1979, pp. 1-6.
14. H.J. Van Zuylen. A Method to Estimate a Trip Matrix from Traffic Volume Counts. Planning and Transport Research and Computation Company, Ltd., Summer Annual Meeting, Univ. of Warwick, July 1978.
15. E. Hauer and B.T. Shin. Estimation of Origin-Destination Matrices from Observed Flows: Simple Systems. Univ. of Toronto, York Univ., Res. Rept. No. 65, March 1980.
16. A.C. Wilson. The Use of Entropy Maximising Models. Journal of Transportation Economics and Policy, Vol. 3, No. 1, 1969.
17. F.J. Cesario. A Primer on Entropy Modelling. American Institute of Planners, Journal, Jan. 1975.
18. P.P. Jovanis, Y. Wai-Ki, and A.D. May. FREQ6PE--A Freeway Priority Entry Control Simulation Model. Institute of Transportation Studies, Univ. of California, Berkeley, Res. Rept. UCB-ITS-RR-78-9, Nov. 1978.
19. M. Eldor. Program SYNODM. Integrated Real-Time On-Ramp Control Strategies. Univ of California, Berkeley, Res. Rept. UCB-ITS-DS-76-1, Ph.D. dissertation, Sept. 1976.
20. K. Ovaichi, A.D. May, R.F. Teal, and J.R. Ray. Synthetic Origin-Destination Algorithm. Simulation of Freeway Priority Strategies (FREQ3CP). U.S. Department of Transportation; Univ. of California, Berkeley, June 1975, p. 399.
21. V.K. Chan, R.G. Dowling, and A.E. Willis. Deriving Origin-Destination Information from Routinely Collected Traffic Counts: Phase I. Institute of Transportation Studies, Univ. of California, Berkeley, Working Paper UCB-ITS-WP-80-1, 1980.
22. A.E. Willis and V.K. Chan. Deriving Origin-Destination Information from Routinely Collected Traffic Counts: Phase II. Institute of Transportation Studies, Univ. of California, Berkeley, Working Paper UCB-ITS-WP-80-10, 1980.
23. A.W. Evans. Some Properties of Trip Distribution Methods. Transportation Research, Vol. 4, 1970, pp. 19-36.
24. N.L. Nihan. Properties of Gravity Model Solutions and Methods of Solution. Univ. of Washington, Seattle, Urban Transportation Program Res. Rept. 78-10, 1978.
25. S.J. Bellomo, R.B. Dial, and A.M. Voorhees; A.M. Voorhees and Associates, Inc. Factors, Trends, and Guidelines Related to Trip Length. NCHRP, Rept. 89, 1970, 59 pp.
26. D.G. Miller and N.L. Nihan. The Subarea Focusing Concept for Trip Distribution in the Puget Sound Area. TRB, Transportation Research Record 610, 1976, pp. 37-43.
27. N.L. Nihan and J. Lassabliere. Sensitivity of Trip Distribution to Subarea Focusing. Transportation Engineering Journal of ASCE, Aug. 1976, pp. 555-569.

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

Alternate Path Analysis Algorithm for Urban Freeway Corridor Evaluation

WILEY D. CUNAGIN AND RAMEY O. ROGNESS

Quick-response procedures and programmable calculator routines have been of increasing interest in transportation planning. As part of a continuing Highway Planning and Research Program study, the PASSER IV system of quick-response methodologies is being developed for analyzing urban freeway corridor alternatives. This will provide a practical and user-oriented tool to evaluate several classes of transportation system management alternatives. An algorithm is presented for estimating the levels of traffic flow on individual parallel facilities in an urban freeway corridor, based on equilibrium traffic assignments. A quick-response routine for the algorithm has been developed for use with a programmable calculator. The level of detail for the routine is ultramacroscopic and deterministic. The routine was designed to be modular to permit additional scenarios, extensions, and modifications to be easily appended. The routine as described is undergoing continuing revision and evolution.

Increased traffic demand and traffic congestion along freeway corridors in major Texas cities are making the effective management and use of existing facilities, as well as the implementation of minor geometric modifications for improving traffic flow, important functions of the various agencies involved. Existing analytical methods and related computer programs offer proven performance capabilities to address these problems, but most are seriously deficient in addressing analyses that require quick response:

1. They do not permit quick and simple analyses of problem areas to allow evaluation of several alternative improvements in a cost-effective manner,
2. They do not fully treat continuous frontage roads that are almost unique to Texas, and
3. They require a large amount of field data and computational effort to conduct the evaluation.

As a result, the use of quick-response procedures and programmable calculator routines has become of increasing interest and implementation.

Practical and user-oriented methods have been proposed. The Signal Operations Analysis Package (SOAP) programmable calculator routines can be used in the design, evaluation, and analysis of signal operation (1). The routines incorporate several computational techniques for analysis of a single approach to an intersection. Routines are available for calculation, analysis, and evaluation of signal settings and measures of effectiveness. Other procedures have been developed such as evaluation routines based on the PASSER II computer program (2) and critical movement analysis procedures (3).

Quick-response routines have been developed for travel-estimation procedures (4,5) and simplified methods have been developed for transportation analysis (6-8). Analysis techniques, including air quality evaluation (9) and energy impacts on travel (10), have been proposed. The development of simplified methods implementable on a programmable calculator has great interest.

The PASSER IV system of quick-response methodologies is now being developed for analyzing urban freeway corridor alternatives to provide transportation system analysts with useful tools to evaluate several classes of transportation systems management (TSM) feasible alternatives. This paper presents, as a part of the PASSER IV system, an algorithm for estimating the levels of traffic flow on individual parallel facilities in an urban freeway corridor, based on equilibrium traffic assignment. The algorithm can be applied to multiple parallel facilities quickly and efficiently. A quick-response routine for the algorithm has been developed for use with a programmable calculator.

SITUATION

The urban freeway corridors are the existing transportation backbone of every major city in Texas. The potential operational capacity of the freeway frontage roads and adjacent parallel arterial streets are major factors in the urban area. In order to manage and improve these critical transportation facilities, several situations and problems must be addressed.

Several of these problems have already been identified, regarding the effective transportation analysis of urban freeway corridor traffic management strategies and the application of TSM improvements to Texas freeways and parallel facilities. The analysis of these available alternative strategies can be time-consuming, costly, and data intensive.

Simplified methods (quick-response techniques) were needed to permit the transportation engineer or planner to expeditiously evaluate a wide range of TSM-based alternatives by using a minimum of data complexity and effort. As part of the Texas Highway Planning and Research Program (HPR) continuing study on development of freeway corridor evaluation system, PASSER IV, a quick-response analysis methodology for expedient evaluation of several classes of TSM-based feasibility studies from an operational viewpoint has been derived. The PASSER IV concept is to permit the decisionmaker the option of efficiently obtaining credible performance measures for various proposed scenarios. The algorithm presented here is based on equilibrium traffic assignment. It estimates the traffic flow levels (and measures of effectiveness) on parallel facilities in an urban freeway corridor.

The algorithm assumes that

1. Travelers behave in a manner that minimizes their travel time,
2. Travel time versus volume/capacity (v/c) ratio curves that describe the parallel paths may be determined, and
3. Piecewise linear approximations of these curves may be computed.

The algorithm is limited by the accuracy of origin-destination estimates, corridor volume estimates, and the travel time versus v/c curves.

The level of detail for the calculator procedure is ultramacroscopic and deterministic in design. Simplicity and user-oriented operation were emphasized. The routine was designed to be modular in design to permit additional TSM alternative scenarios to be addressed by subsequent additions and subroutines.

Algorithm Background

The algorithmic approach to the three alternate-path traffic-assignment problems is based on Wardrop's

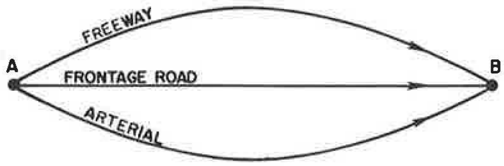


Figure 1. Alternate urban freeway corridor paths.

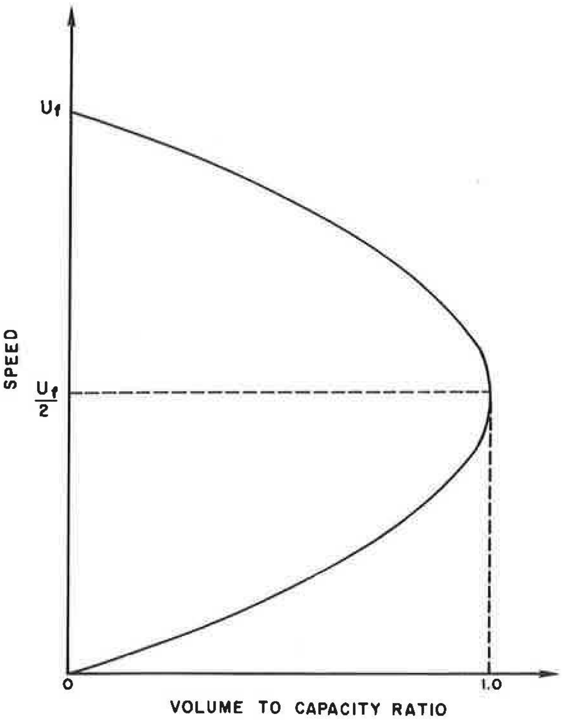


Figure 2. Urban freeway speed versus v/c.

first principle (user optimization) of equilibrium flows (11). The original corridor scenario for the three alternate paths was freeway, frontage road, and a parallel arterial street. The algorithm uses travel time relations for allocating the traffic to the three paths.

The freeway travel time is based on the relation between average freeway speed and v/c ratio as developed by Texas Transportation Institute (TTI). The frontage road travel time and the arterial street travel time is based on speed, volume, capacity, and signal density. These relations are developed as a piecewise linear function of travel time to v/c ratio for each alternate path.

The procedure allocates corridor travel demand to the facilities based on travel times. As these volumes are added to each facility, the travel time on the facility is increased. The procedure iteratively determines the allocation of the demand to provide equal travel times for all facilities by using piecewise linear representations of the travel time curves.

Algorithm Development

Traffic flows on three parallel paths are illustrated in Figure 1. Travelers wish to go from point A to B. Point A might be a suburban community and point B could be a central business district. These travelers may choose among paths 1, 2, and 3 for their trip. Each path has its own distances, speed, and capacity attributes. For a typical urban freeway corridor in Texas, path 1 is the freeway main lanes, path 2 is the frontage road, and path 3 is a parallel arterial street.

The solution approach presented here for allocating traffic among these competing paths is based on Wardrop's first principle of equilibrium flows in a transportation network (11). This principle states that each individual traveler will choose a path that gives him or her minimum travel time under the perceived operating condition. This assumption is known as user optimization and is in general agreement with observed behavior. The driver perceives (or anticipates) the operating conditions on each path and then chooses the path that he or she thinks will minimize travel time from point A to point B.

Traditional nonequilibrium traffic assignment techniques have not addressed allocation of traffic explicitly so that this condition is met. For example, in an all-or-nothing assignment, the technique finds the minimum travel time between two zones under specific conditions. All traffic is then assigned to the path that has that minimum time. The presence of this traffic causes the resulting travel time on that path to become much greater than the calculated value and, if minimum travel times were again computed, another path between the two zones would probably be chosen. This diversion of traffic is addressed in capacity-restraint assignment, yet travelers may still not be on a path that gives them minimum travel time. A number of methods are now used to redistribute assigned traffic more realistically in a corridor following a traffic assignment for the urban area. Many of these methods, however, require substantial effort and time to use and are not amenable to quick and simple analysis to evaluate several alternatives for TSM strategies in the corridor.

The algorithm presented in this paper explicitly treats the perceptions of path choice of the individual traveler and is sensitive to TSM actions that may be applied in the corridor.

TRAVEL TIME FUNCTIONS

In modeling the path choices of individual drivers, it is first necessary to model the variation of travel time on a path with increasing traffic on that path.

For a typical urban freeway corridor in Texas, as depicted in Figure 1, path 1 is the freeway main lanes, path 2 is the frontage road, and path 3 is a parallel arterial street. In order to compare travel times along each of these paths to satisfy the equal travel time condition (user optimization), travel times along each path must be determined as a function of the volume and capacity on that path. For freeways, speed has been related to v/c ratio by the relation shown in Figure 2 (12). The quantity u_f is the free speed for the facility.

Creighton, Hamburg, Inc., in work for the Federal Highway Administration (FHWA), propose modification of the relation shown in Figure 2 to that shown in Figure 3 to model reduction in speed due to congestion for the FHWA micro assignment model (13). For v/c values in the range (0, 0.8), this curve is the same as the Highway Capacity Manual curves shown in Figure 2 (12). For values of v/c greater than 0.8, the curve drops linearly to a value of 0 when v/c = 1.0, as shown in Figure 3.

The monotonically decreasing form of the function in Figure 3 agrees with the observed condition that average speed decreases as the v/c ratio increases. One logical difficulty, however, is that the speed in Figure 3 decreases to zero at a volume equal to

Figure 3. FHWA freeway speed versus v/c for freeway arterial vehicle mile per hour splitter.

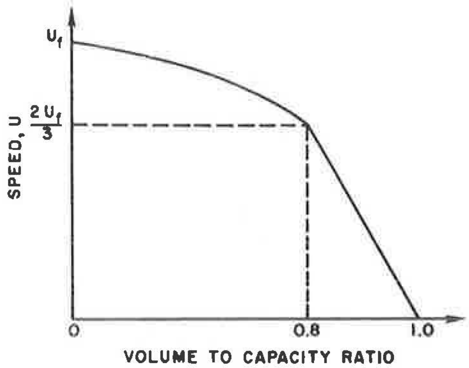


Figure 4. TTI urban freeway speed versus v/c.

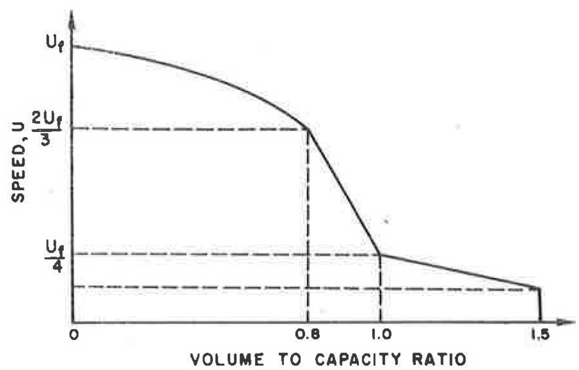


Figure 5. TTI urban freeway travel time versus v/c.

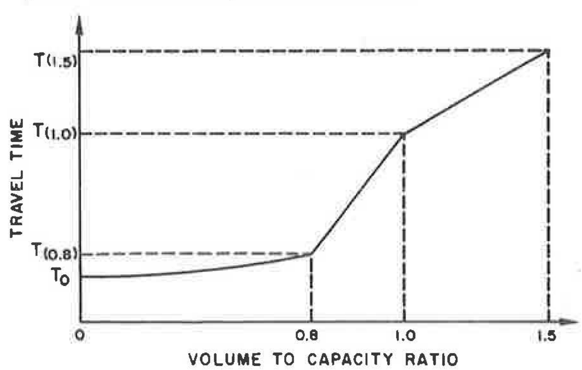
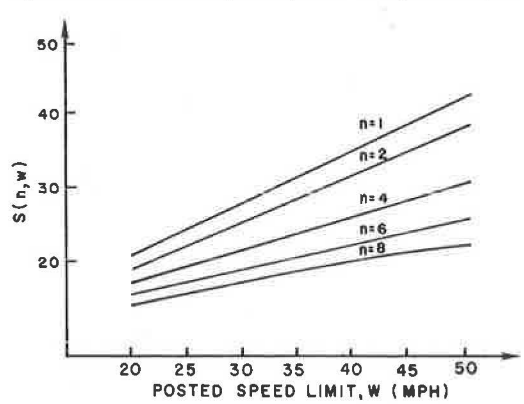


Figure 6. Free flow versus posted speed w and signal density n.



capacity, especially since Figure 2 shows a speed of $u_f/2$ when volume is equal to capacity.

For the freeway speed model used in this algorithm, speed at capacity was set at $u_f/4$ to approximate average actual speed. In addition, because volumes greater than estimated capacity are sometimes observed (e.g., when level of service C volumes and capacities are used), the freeway speed curve was extended in this research to a speed value of 10 mph when v/c = 1.5. The freeway speed curve developed by TTI is shown in Figure 4.

The relation shown in Figure 4 is piecewise linear for v/c > 0.8, so that mathematically the relation can be expressed as follows:

$$S_{fwy} = \begin{cases} 0.5\,S_0 + (S_0^2 - 2v)^{1/2} & v/c \leq 0.8 \\ S_1 + [(S_2 - S_1)/0.2]\,(v/c) - 0.8 & 0.8 < v/c < 1.0 \\ S_2 + [(10 - S_2)/0.5]\,(v/c) - 1.0 & 1.0 < v/c < 1.5 \\ 10 \quad v/c > 1.5 \end{cases} \tag{1}$$

where

S_{fwy} = speed on freeway at volume v per lane (mph),
v = freeway volume per lane (vehicles/h),
c = capacity per lane (vehicles/h),
S_0 = free flow speed on freeway (mph),
S_1 = speed on freeway when v/c = 0.8 (mph), and
S_2 = speed on freeway when v/c = 1.0 (mph).

This model provides a determinable relation between speed and volume for the freeway situation.

From the speed versus v/c relation shown in Figure 4, a travel time relation may be constructed by using

$$T(v/c) = \text{Travel time} = \text{Distance/Speed} \tag{2}$$

for each continuous interval. The resulting travel time relation is shown in Figure 5. This relation shows that, as the volume (or v/c) on the freeway increases, the travel time increases. This developed relation agrees with expected results. The piecewise linear nature of the travel time curves makes possible the evaluation of successive critical points on the curves for parallel facilities rather than the solution of a set of mathematical equations. Although modification of the FHWA's freeway-surface arterial VMT splitter speed versus v/c curves were used here to derive travel time curves, other curves, such as those of Davidson (14), or FHWA (15), may be used as long as they are modified to a piecewise linear form.

For signalized roadways, the relation between speed and capacity is complicated by the presence of the signals along the roadway, which provide a further component of delay. The effect of this delay can be correlated to the signal density and signal timings. The relation developed is a modified version of that in FHWA's micro assignment model (13). This relation provides for travel time to be dependent on volume and signal density. For signalized roadways the equations are

$$S = \begin{cases} S_0(n,w) + (v/c)f(n) & v/c < 0.8 \\ S_1 + [(S_2 - S_1)/0.2]\,(v/c) - 0.8 & 0.8 < v/c \leq 1.0 \\ S_2 + [(5 - S_2)/0.5]\,(v/c) - 1.0 & 1.0 < v/c \leq 1.5 \\ 5 \quad v/c > 1.5 \end{cases} \tag{3}$$

where

S = speed on signalized roadway at volume v per lane (mph),
v = roadway volume per lane (vehicles/h),
c = capacity per lane (vehicles/h),

n = signal density (signals/mile),
w = posted speed (mph),
$f(n)$ = speed reduction with unit increase in v/c,
$S_0(n,w)$ = free-flow speed for signalized roadway with signal density n and posted speed w,
S_1 = speed when v/c = 0.8,
S_2 = speed when v/c = 1.0, and
$S_0(n,w) = 3600/(3600/w) + 12.5n$.

$$f(n) = -0.0672n^3 + 0.781n^2 - 3.2232n \quad n < 5.5 \quad (4a)$$

$$f(n) = 0.138n - 6.028 \quad n \geq 5.5 \quad (4b)$$

A family of curves that relate free-flow speed to posted speed and signal density is shown in Figure 6. A family of curves that show average speed for varying values of signal density (n), posted speed, and values of v/c is illustrated in Figure 7.

Travel time curves may be constructed by using the speed curves shown in Figure 7 and Equation 2. The travel time curves developed are illustrated in Figure 8.

Figure 8 shows that, although the effect of signal density is somewhat masked, the travel time relation behaves as would be expected.

ALGORITHM

Once the travel time functions have been defined for each of the three paths in Figure 1, the problem remains to determine how the travel demand from A to B will be distributed among paths 1, 2, and 3. Obviously, if all that is considered is the free-flow travel time, all of the drivers will choose path 1 (the freeway path) as in an all-or-nothing assignment. But, the actual travel time increases as each motorist enters the facility, so that eventually the path 1 travel time for an additional motorist is increased such that, if that motorist enters path 1, then path 2 or path 3 will have lower free-flow travel times than the travel time on path 1 under the existing conditions. Since this is contrary to Wardrop's first principle, the motorist selects the minimum travel time path.

The travel time curve for each path, due to its piecewise linear nature, contains a series of inflection points (discontinuities). The successive evaluation of these critical points is the basis for this algorithm. The free-flow travel time of each path (the intercept with the travel time axis) is considered to be a critical point. The remaining critical points (discontinuities) on the curves project onto the travel time axis to define the intervals of travel time for which the slopes of all of the travel time curves are simultaneously constant. The total assigned volume is computed at the upper limit of each of these travel time intervals. When this assigned value exceeds the total demand, the volumes on each path are backed off simultaneously, proportional to the slopes of the piecewise linear travel time curves on that interval.

Figure 7. FHWA signalized roadway speed versus v/c.

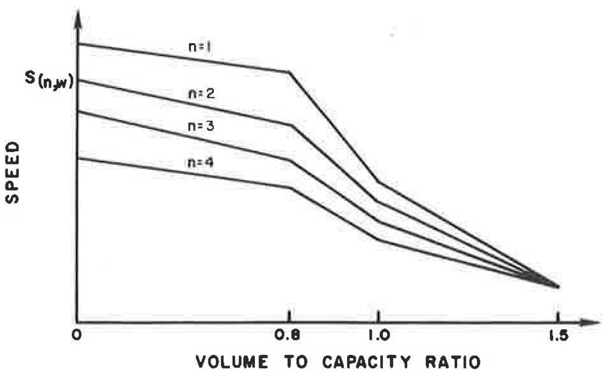
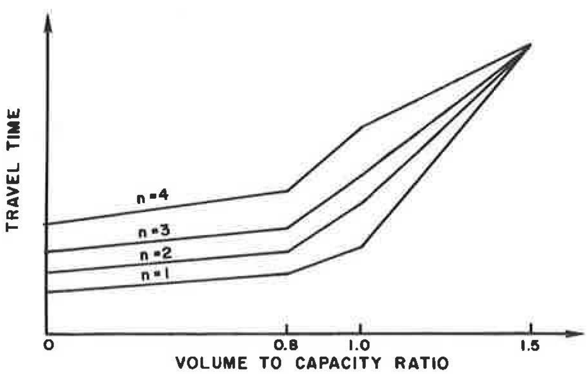


Figure 8. TTI signalized roadway travel time versus v/c.



CALCULATOR ROUTINE FEATURES AND CAPABILITY

The routine has undergone several revisions in its development. The addition of enhancements and modifications to the original routine is an evolutionary process. Improvements in run time, program structure, and number of steps and memories have been accomplished to increase the efficiency and applicability of the procedure.

Original Procedure

The original procedure was developed for the algorithm just described to consider a typical urban freeway corridor in Texas. The three parallel paths available were established as the freeway main lanes, frontage roads, and a parallel arterial street. The input data are given in the table below.

Input Data	Freeway	Frontage Road	Arterial
No. of lane	X	X	X
Distance	X	X	X
Speed	X	X	X
Capacity	X	X	X
Signal density		X	X
Total demand			

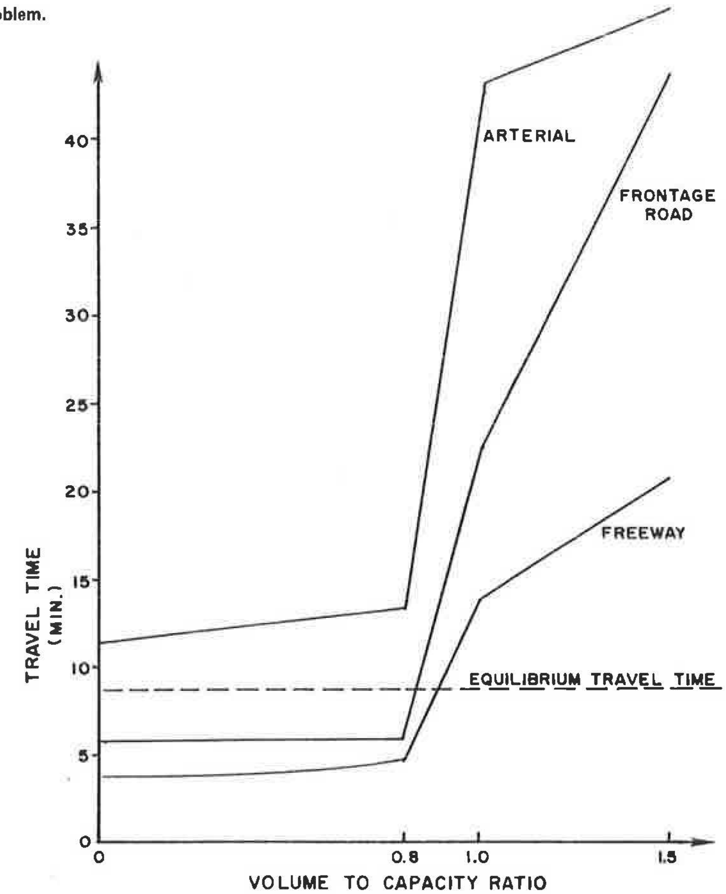
The input data along with embedded data in the routine provide the characteristics of the facility and demand volume. The piecewise linear segment of each travel time curve is established at v/c of 0, 0.8, 1.0, and 1.5. A representative series of travel time curves are illustrated in Figure 9. The free-flow speed is the only variable (and number of signals for nonfreeway paths) that the user can input to describe the curve. The corresponding speeds for v/c of 0.8, 1.0, and 1.5 are fixed internally. The output for the original routine are system travel time (at equilibrium), traffic volumes on each path, and v/c for each path.

The original routine satisfied the objectives of the study. During its development, certain structural and design limitations were recognized. Also, several enhancements, modifications, and variables input were recognized as desirable for incorporation in the procedure.

Procedure Extensions

The original procedure was revised. The revised

Figure 9. Travel time functions for example problem.



procedure still retained the basic three alternate path algorithms. It used fewer program steps, less input cards, and fewer memory locations. The use of indirect addressing and skip on zero routines improved computational efficiency.

One improvement in the program structure is that any number of freeways, frontage roads, or arterial streets may be used to a maximum of three total paths. Input of these path data may be in any order.

To provide greater flexibility and utility of the procedure, three variables were added--quality of progression factor, variable overcapacity limit, and variable overcapacity speed. The quality of progression factor (range of 0-1) provides a means to model the progressive quality along an arterial street or the frontage road to match existing or future operational characteristics more closely. The default value is one.

The overcapacity limit and the overcapacity speed are related. The overcapacity limit is the v/c for the final reference point on the travel time curve. The default value is 1.5. The input is 0 for the default value or a value greater than 1.0. The overcapacity speed is the corresponding speed at the overcapacity limit used. The default value is 10 mph for freeways and 5 mph for signalized facilities. These three new variable inputs provide great flexibility for the user to model the problem to be analyzed. This flexibility provides increased ability to model real-world conditions in a corridor. However, this flexibility requires that additional user instructions be provided to aid in proper selection of the variable values.

To increase the capability of the procedure to better model greater complexity and provide additional path alternative analysis, two features were added to the program. The first addition is the ability to handle more than one speed along a path. This corresponds to be the ability to analyze different travel times on segments of a path. An example would be different posted speed limits along an arterial street. The second feature is the capability to analyze more than three alternate paths or a combination of parallel streets. This feature is directed toward providing the capability of analyzing freeway main lanes, frontage roads, and three parallel arterial streets as alternate paths. The multiple alternate paths for the facility type are preprocessed to provide a composite representation of the facilities before input to the main algorithm. The algorithm output for those facilities is then fed into a postprocessor to provide the estimated traffic assignment to those paths. Extensions to multiple paths or three representations of corridor facilities could be analyzed similarly by the procedure.

SUMMARY

The application of programmable calculator routines and simplified methodologies to analyze TSM alternative improvements in a freeway corridor is shown. The ability of the calculator routine to analyze more than the basic three-path situation is indicated for corridor traffic assignment.

The routine is part of a continuing HPR research study. The procedure is undergoing continuing revision and evolution to increase the efficiency and widen the applicability of the procedure to corridor evaluations.

ACKNOWLEDGMENT

This paper presents the quick-response technique portion of a continuing Texas HPR study, Development of Freeway Corridor Evaluation System--PASSER IV, sponsored by the Texas State Department of Highways and Public Transportation and FHWA and conducted by TTI. Carroll J. Messer is acknowledged for his assistance in the development of the basic algorithm. Herman E. Haenel, Blair G. Marsden, and Pete Osburn of the Texas State Department of Highways and Public Transportation are acknowledged for their guidance and assistance in all phases of the study. The contents of this paper reflect our views, and we are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Texas State Department of Highways and Public Transportation or FHWA. This report does not constitute a standard, specification, or regulation. As part of a continuing HPR study, this paper has not been subject to a final review process or acceptance by the sponsor.

REFERENCES

1. Signal Operations Analysis Package, Volume 4--Programmable Calculator Routines. Federal Highway Administration, Implementation Package 79-9, 1979.
2. Signal Timing Optimization to Maximize Traffic Flow, A 3-Day Workshop. Texas Transportation Institute, Texas A&M Univ., College Station, 1981.
3. Critical Movement Analysis. Univ. of Florida, Gainesville, Sept. 1980.
4. A.B. Sosslau, A.B. Hassam, M.M. Carter, and G.V. Wickstrom. Travel Estimation Procedures for Quick Response to Urban Policy Issues. NCHRP, Rept. 186, 1978, 70 pp.
5. A.B. Sosslau, A.B. Hassam, M.M. Carter, and G.V. Wickstrom. Quick-Response Urban Travel Estimation Techniques and Transferable Parameters, User's Guide. NCHRP, Rept. 187, 1978, 229 pp.
6. I. Salomon. Application of Simplified Analysis Methods: A Case Study of Boston's Southeast Expressway Carpool and Bus Lane. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, 1979.
7. K.H. Karash, A. Baver, and M.L. Manheim. Workshops in Simplified Methods, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Aug. 1979.
8. M.L. Manheim, P. Furth, and I. Salomon. Examples of Transportation Analysis Using Pocket Calculators. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, 1979.
9. K.K. Karash and E. Hollingshead. A Case Study of the Use of Pocket Calculator and Workshop Methods for Analyzing Air Quality Related Transportation Control Strategies. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Working Paper 80-5, 1980.
10. T.J. Tardiff and J.L. Benham. Quick-Response Methodology for Analyzing the Travel Impacts of Fuel-Supply Limitations. Paper presented at 60th Annual Meeting, TRB, 1981.

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

Consideration of Alternative Access, Egress, and Line-Haul Travel Choices Within UTPS Framework

ASHOK KUMAR AND YEHUDA GUR

In many large metropolitan areas more than one line-haul transit service is often available in some travel corridors. Examples include express bus and rail rapid transit, commuter rail and rail rapid transit, private suburban bus lines and competing service provided by regional transit operator. This is especially true as one moves away from the core area and corridors become wider. Coupled with the choice of line-haul modes are several choices of accessing these modes such as walk, feeder bus, park-and-ride, and kiss-and-ride. This paper addresses these issues and describes a systematic procedure for analyzing such mode choices. It is argued that straightforward use of urban transportation planning system (UTPS) programs prevents meaningful analysis of important policy issues due to their all-or-nothing assignment principle, when real access-egress and line-haul choices have to be considered.

Much progress has been made in the last two decades in quantitative aspects of long-range planning of highway and mass transportation facilities. The forecasting of travel demand along highway links and transit lines that comprise the transportation network of a metropolitan area has been greatly facilitated by the availability of two software packages, PLANPAC (1) and Urban Transportation Planning System (UTPS) (2), developed by the U.S. Department of Transportation. Several publications (3,4) describe the sequence of trip generation, trip distribution, mode choice, and route-assignment models used to simulate the traffic flow by using these packages. This paper addresses the problems associated with application of computer programs UNET, UPATH, UPSUM, UMODEL, and ULOAD (2) if alternative access, egress, and line-haul choices are available between an origin-destination (O-D) pair. Briefly, UNET is used to prepare the computerized description of a transit system that serves the study area. UPATH finds the minimum impedance (travel time, travel cost, or both) path between any O-D pair in the system and zone-to-zone fare matrix. UPSUM computes the travel time along minimum impedance path and can store the time spent walking, waiting, transferring, and in-motion along various travel modes (walk, automobile, bus, or rail) used between an O-D pair. UMODEL computes the share of transit trips (mode split) given the transit level-of-service data prepared by UPATH and UPSUM, highway level-of-service data prepared by either PLANPAC programs BUILDHR (1) and BUILDVN (1), or UTPS programs HR (2) and UROAD (2). It computes total person trips between an O-D pair by using a user-specified mode split model. Finally, ULOAD loads the resulting transit trips along the minimum impedance path to produce a transit assignment.

PROBLEM DEFINITION

The process described above is satisfactory for synthesizing transit travel patterns if only one transit path is overwhelmingly used by an O-D pair. However, in large metropolitan areas, as one moves away from the core of the region, several options for commuting exist. For example, rail rapid transit with walk, feeder bus, park-and-ride and kiss-and-ride access; express bus with walk, feeder bus, park-and-ride, and kiss-and-ride access. Even if a very fine zone system and detailed network description are used in demand analysis, the problem of all-or-nothing assignment cannot be easily overcome. Assessment of travel demand along competing line-haul and access-egress service is essential in project-level planning and design of transit facilities. The following sections describe the computerized network analysis and mode choice estimation process developed for the Northeast Ohio Areawide Coordinating Agency (NOACA) as part of its alternatives analysis work program. It is described in detail elsewhere (5). The procedure follows, in its general structure, the mode split procedures of the Chicago Area Transportation Study (CATS) (6) and the North Central Texas Council of Governments (7).

Modeling Procedures

The procedure provides

1. Explicit estimation of use of one or more available line-haul transit modes;
2. Explicit representation and estimation of access-egress modes and their impact on use of main line-haul mode;
3. Method to represent parking fee-walking distance trade-off faced by automobile users in high-density areas, such as the central business district (CBD); and
4. Method to represent and analyze impacts of alternative transportation system management strategies (such as parking costs, toll pricing, fare changes, and fuel price changes) on transit use.

Modeling Structure

The heart of the procedure is a disaggregate logit mode-choice model that estimates modal use for individual trips. The procedure provides a modeling structure, including sampling and aggregation procedures based on the principles of Monte Carlo simulation (5), that links the mode-choice model to available aggregate descriptors of level-of-service and socioeconomic attributes of the travelers.

The mode choice model is a modified nested binary choice logit model. The derivation of logit choice models and the justification of their use for the analysis of mode choice are thoroughly discussed in the literature [for example, see Lisco and Stopher (8,9), and McFadden and Domencich (10)]. The analysis of transit starts with an estimation of the disutility of each access and egress mode of each valid transit path. The submode that has the least disutility is assumed to represent the resultant disutility of the access and egress portions of the path. Given the access and egress resultant disutilities and the line-haul service attributes, the composite disutility of each valid transit path is determined.

The transit path that has the least disutility is assumed to represent the resultant disutility of transit for analysis of the automobile versus transit mode choice. A binary logit mode choice formulation is used in computing the probability of choosing transit $[P_r(t)]$ and automobile $[P_r(a)]$. The automobile option reflects both drive alone and passenger modes. In order to develop the vehicle trip table for highway assignment, the expected automobile occupancy is estimated as a function of (a) trip purpose, (b) trip length, (c) trip orientation, and (d) zonal income of the tripmaker's place of residence. The development of the automobile-occupancy model and associated look-up curves are described elsewhere (5,11). The expected automobile occupancy is also used in computing the disutility for automobile travel. Automobile operating cost and parking fee are divided by expected automobile occupancy to reflect shared cost of automobile travel among the occupants.

After the probability of transit use has been computed, the probability of using alternative transit line-haul modes is computed by using a binary choice logit model. The results are then weighted by the probability of using transit. In its present form, the modeling process assumes that the individual tripmaker will use the best access and egress mode associated with each line-haul mode with the probability one. Therefore, absolute probability of the best access and egress mode is taken to be equal to the probability of line-haul mode computed above.

Note that simulated trips that originate from the same zone will have different access-egress modes that have the least disutility depending on the distance from the line-haul facility. Therefore, at the zonal level, distribution over various access-egress modes is achieved. This contrasts sharply with the conventional use of UNET and UPATH programs where all trips get assigned to the centroid connector by providing access to the transit network. Details of disutility calculations and estimation of mode-choice probability follow. Details of the model structure can be found elsewhere (5).

DISUTILITY CALCULATIONS

The variables in disutility calculations are defined in Table 1.

For transit access, the disutility of walk access to line-haul facility i is computed as

$$U_{wa}(i) = \text{VWALK} \times \text{WKTL}(i,a) \quad (1)$$

$$\begin{aligned} U_{ba}(i) = {} & \text{VWALK} \times \text{WKTB}(i,a) + \text{VWAIT} \times \text{WTTB}(i,a) \\ & + \text{VIVT} \times \text{BIVT}(I,a) + \text{VCOST} \times \text{BAFAR}(i,a) \\ & + \text{VTFER} \times 1 + \text{VBIASA}(B) \end{aligned} \quad (2)$$

where i is the transit path and a is access.

Disutility of park-and-ride access is computed as

$$\begin{aligned} U_{pa}(i) = {} & \text{VWALK} \times \text{PRWK}(i) + \text{VIVT} \times \text{PRIVT}(i) \\ & + \text{VCOST} \times [\text{PROPC}(i) + 0.5 \times \text{PRPCST}(i)] + \text{VBIASA}(P) \end{aligned} \quad (3)$$

Disutility of kiss-and-ride access is computed as

$$U_{pa}(i) = \text{VIVT} \times 2 \times \text{PRIVT}(i) + 2 \times \text{VCOST} \times \text{PROPC}(i) + \text{VBIASA}(k) \quad (4)$$

Resultant access disutility is computed as

$$U_a(i) = \text{Min}[U_{wa}^{(i)}, U_{ba}^{(i)}, U_{pa}^{(i)}, U_{ka}^{(i)}] \quad (5)$$

For transit egress, the disutility of walk egress from transit line-haul is computed as

$$U_{we}(i) = \text{VWALK} \times \text{WKTL}(i,e) \quad (6)$$

where e is egress.

Disutility of feeder bus egress is computed as

$$U_{be}(i) = VWALK \times WKTB(i, e) + VWAIT \times WTTB(i, e) + VIVT \times BIVT(i, e) + VCOST \times BAFAR(i, e) + VTFER \times 1 + VBIASA(B) \quad (7)$$

Resultant egress disutility is computed as

$$U_e(i) = Min[U_{we}(i), U_{be}(i)] \quad (8)$$

For transit line-haul, the disutility of transit line-haul path is calculated as follows:

$$U_{tm}(i) = VWAIT \times TOVT(i) + VIVT \times TIVT(i) + VTFER \times NTFER(i) + VBIASTM(i) + U_a(i) + U_e(i) \quad (9)$$

where i is 1 for a path that contains feeder bus and express bus only and i is 2 for a path that contains feeder bus and rail only.

Resultant transit utility is calculated as

$$U_t = Min[U_{tm}(1), U_{tm}(2)] + VBIAST \quad (10)$$

For automobile, the disutility of automobile travel is calculated as

$$U_a = VIVT \times AIVT + VWALK \times AWTIME(D) + \{VCOST[0.5 \times APCOST(D) + AOPC]/OCC\} + VAA \times AA \quad (11)$$

The probability of choosing transit is computed as

$$P_r(t) = 1/[1 + exp(U_a - U_t)] \quad (12)$$

The probability of choosing automobile is computed as

$$P_r(a) = 1 - P_r(t) \quad (13)$$

The probability of choosing specific transit path is calculated as

$$P_r[tm(1)] = \left(1/\{1 + exp[U_{tm}(2) - U_{tm}(1)] \times P_r(t)\}\right) \quad (14)$$

$$P_r[tm(2)] = 1 - P_r[tm(1)] \quad (15)$$

The values of calibration parameters used in disutility calculations are given in Tables 2 and 3 for trips destined to the CBD and to the non-CBD, respectively. These values were obtained by researching the disaggregate mode choice literature and fine-tuning them to replicate observed ridership patterns in the Cleveland metropolitan area. The details of the model calibration and validation procedures can be found elsewhere (5).

SYNTHESIS OF A PSEUDOOBSERVATION

As mentioned earlier, the modeling process described in this paper uses pseudosample enumeration technique to provide zonal level aggregate mode-split

Table 1. Variables used in modal disutility calculations for trips to CBD and non-CBD destinations.

Notation	Description	Notation	Description
WKTL	Walk time to or from line-haul facility	AIVT	In-vehicle time spent driving automobile if automobile is line-haul mode
WKTB	Walk time to or from feeder bus that serves line-haul facility		
WTTB	Wait time to board feeder bus	AOPC	Operating cost of driving automobile if automobile is line-haul mode
BIVT	In-vehicle time spent riding feeder bus	AA	Automobile availability estimated as AA = 0, if no. of automobiles owned by household is 0, AA = 0.8 + 0.2/(no. of persons in the household), if no. of automobiles owned by household is 1, and AA = 1, if no. of automobiles owned by household is 2 or more
BAFRA	Fare for feeder bus		
PRWK	Walk time from park-and-ride lot to line-haul facility		
PRIVT	In-vehicle time spent driving automobile to park-and-ride lot	U_{wa}	Disutility of walk access
PROPC	Operating cost of driving automobile to park-and-ride lot	U_{ba}	Disutility of feeder bus access
PRPCST	Parking fee for leaving car at park-and-ride lot	U_{pa}	Disutility of park-and-ride access
TOVT	Total wait time to board first line-haul and subsequent line-haul facilities	U_{ka}	Disutility of kiss-and-ride access
		U_a	Resultant access disutility
TIVT	Total in-vehicle time spent riding first line-haul and subsequent line-haul facilities	U_{we}	Disutility of walk egress
		U_{be}	Disutility of feeder bus egress
NTFER	No. of line-haul transfers	U_e	Resultant egress
AWTIME(D)	Walk time between parking lot and final destination if automobile is line-haul mode	U_{tm}	Disutility of transit line-haul path
		U_t	Resultant transit disutility
APCOST(D)	Parking fee paid if automobile is line-haul mode	U_a	Disutility of automobile line-haul travel

Table 2. Calibrated values of parameters used in modal disutility calculations for trips to CBD.

Description	Notation	Home-Based Work Trips	Home-Based Nonwork Trips	Home-Based School Trips	Non-Home-Based Trips
Value of in-vehicle time (min)	VIVT	-0.025	-0.012	-0.025	-0.016
Value of out-of-pocket cost (¢)	VCOST	-0.012	-0.01	-0.012	-0.008
Value of out-of-vehicle walk time (min)	VWALK	-0.058	-0.04	-0.058	-0.024
Value of out-of-vehicle wait time (min)	VWAIT	-0.09	-0.03	-0.09	-0.048
Value of no. of transit transfers	VTFER	0	0	0	0
Value of automobile availability	VAA	0	+4.12	0	NA
Value of bias coefficient for feeder bus mode of access-egress	VBIASA(B)	0.25	0	+0.25	+0.35
Value of bias coefficient for park-and-ride mode of access	VBIASA(P)	-0.36	-0.22	-0.36	-0.125
Value of bias coefficient for kiss-and-ride mode of access	VBIASA(K)	-0.40	-0.26	-0.40	-0.1
Value of bias coefficient for express bus in line-haul operation	VBIASTM(E)	+0.31	-0.52	+0.31	-0.67
Value of bias coefficient for rail transit in line-haul operation	VBIASTM(R)	+0.31	-0.45	+0.31	-0.43
Value of transit bias	VBIAST	+0.14	+2.48	+0.14	+0.06

Table 3. Calibrated values of parameters used in modal disutility calculations for trips to non-CBD.

Description	Notation	Home-Based Work Trips	Home-Based Nonwork Trips	Home-Based School Trips	Non-Home-Based Trips
Value of in-vehicle time (min)	VIVT	-0.01	-0.001	-0.001	-0.008
Value of out-of-pocket cost (¢)	VCOST	-0.01	-0.01	-0.010	-0.009
Value of out-of-vehicle walk time (min)	VWALK	-0.03	-0.03	-0.021	-0.05
Value of out-of vehicle wait time (min)	VWAIT	-0.06	-0.04	-0.05	-0.05
Value of no. of transit transfers	VTFER	-0.23	-0.1	-0.10	0
Value of automobile availability	VAA	+3.3	+5.0	+3.5	NA
Value of bias coefficient for feeder bus	VBIASA(B)	+0.42	+0.23	+0.34	+0.55
Value of bias coefficient for park-and-ride mode of access	VBIASA(P)	-0.55	-0.10	-0.73	-0.60
Value of bias coefficient for kiss-and-ride mode of access	VBIASA(K)	-0.25	-0.10	-0.40	-0.40
Value of bias coefficient for express bus in line-haul operation	VBIASTM(E)	-0.65	-1.75	-0.87	-1.50
Value of bias coefficient for rail transit in line-haul operation	VBIASTM(R)	-0.70	-1.35	+0.13	-1.45
Value of transit bias	VBIAST	1.67	+1.32	+2.65	-1.56

forecasts for planning purposes. This is one of the most satisfactory procedures to develop aggregate mode-split rates by using disaggregate mode choice models when computer resources are available. Details and discussion of aggregate forecasting from disaggregate choice models can be found elsewhere (5,12). A computer program, MSPLIT (13), was written that performs the necessary Monte Carlo simulation by sampling frequency distributions of zonal socioeconomic attributes and level-of-service data. The process is described briefly here and details can be found elsewhere (5).

Assigning of Automobile-Related Level-of-Service and Socioeconomic Attributes

The components of disutility associated with automobile travel are shown in Equation 11. Automobile in-vehicle time and operating costs on the line-haul portion of the journey are taken from input zone-to-zone highway travel time and distance skim matrices. The intrazonal variability of these components is assumed to be small and so can be ignored. For non-CBD destinations, walk time at destination and parking fee are also assumed to have minimal intrazonal variability and are estimated by using input zonal level data. For high-density areas such as the CBD, where considerable variation in parking fee choice and associated walking distance to reach the final destination exists, a simulation approach is used to assign these attributes to the sampled observation. Cumulative probabilities of walking certain distances between the parking lot and final destination in the Cleveland CBD are shown elsewhere (5). These values were derived from a special parking lot survey conducted in the downtown area. In addition, for each traffic analysis zone within the CBD area, a curve was derived to show the minimum parking fee that must be paid if one wishes to park within 0.1, 0.2, 0.3,...,1.0 mile from the zone centroid. MSPLIT generates a walking distance randomly by using the aforementioned probability distribution and computes a parking fee for the associated distance by using input parking fee-walking distance curve. Methodology for constructing these curves is also described elsewhere (5).

Automobile availability (AA) for the pseudoobservation is estimated by random sampling from the joint distribution of automobile ownership and household size available for the zone of trip production. The methodology for developing joint automobile ownership and household size distributions as a function of zonal mean automobile ownership and zonal mean household size is described elsewhere (14). These distributions have been hardcoded inside the MSPLIT computer program. Automobile availability is defined by using the sampled number of automobiles and number of persons in the household, by using the relation given in Table 1. Since households that have a greater number of automobiles make more trips than households that have no car or fewer cars, the joint probability distribution of automobile ownership and household size is weighted prior to sampling by the relative tripmaking frequencies of households of varying size and automobile ownership characteristics. These tripmaking frequencies are input to MSPLIT as control cards; their derivation is described elsewhere (15).

Determining Choice Set of Public Transportation

The number of transit paths that are assumed to be available to a pseudoobservation and their characteristics depend on the transit network. The available access-egress submodes vary for each path, depending on the priority mode of the path and automobile ownership levels. The following paragraphs explain the process of determining the choice set and its attributes.

Four distinct types of transit service are available in the Cleveland metropolitan area. In the densely developed central city area, frequent radial and crosstown bus service is provided. These routes are designated as local bus routes and are coded as mode 4 in the UNET (2) transit network description. The service to inner and outer suburbs within Cuyahoga County and adjacent developed communities is provided by using radial bus routes that operate more frequently during rush hours. These routes are designated as express bus routes and are coded as mode 6 in the UNET transit network description. In addition to the local and express bus service, a heavy rail line service is available within Cleveland and two light rail lines operate between Shaker Heights and the Cleveland CBD. Heavy rail service is coded as mode 7 and light rail service is coded as mode 8 in the UNET network description. Local crosstown buses interface with the express bus and rail service and, therefore, provide feeder service as well. Modes 4, 6, 7, and 8 are operated by the regional transit authority. Limited intercity bus service between the Cleveland CBD and some of the outlying communities in Lorain and Lake Counties is also provided by private operators. This service is designated as mode 5 in the network description.

Nontransit modes 1, 2, and 3 are used to designate CBD sidewalk links, centroid automobile connectors, and centroid walk connectors, respectively.

The automobile connectors are used only if express bus and rail service cannot be accessed by using local bus from any given zone. As will be shown later, level-of-service provided by alternate access modes is estimated by using a zonal transit service (ZTS) description file (13) prepared exogenously.

The modeling process developed for the Cleveland area is capable of analyzing mode split between competing rail and express bus service (if available) between any O-D pair. In order to accomplish this split, it is necessary to develop three sets of transit paths by using computer program UPATH (2). These paths are developed as follows:

1. Set 1 includes paths developed by using complete transit network description.
2. Set 2 includes paths developed by using transit modes 4, 5, and 6 only; that is, rail service is excluded from the network description by using no transfer allowed (NOX) option of UPATH (2).
3. Set 3 includes paths developed by using transit modes 4, 5, 7, and 8 only; that is, express bus service is excluded from the network description by using the NOX option.

Each set is then analyzed to determine for each O-D pair:

1. Priority mode, that is, the highest numbered mode used in the path;
2. First line-haul mode used;
3. Last line-haul mode used;
4. Wait for first line-haul and subsequent line-haul modes used in the path;
5. Total line-haul in-vehicle time;
6. Number of line-haul transfers;
7. Automobile in-vehicle time, if access to the transit system is using mode 2 (automobile connector); and
8. Wait time for the feeder bus at the origin end, if feeder bus is not part of the line-haul.

A computer program TPATH (5,13) has been written to facilitate the transit path analysis mentioned above. This program is similar in concept to the UTPS program UPSUM (2). Both programs read the path description produced by UPATH and prepare zone-to-zone transit skim trees. TPATH, however, distinguishes between the line-haul and access-egress portions of the paths. It trims the access-egress portions of the path and summarizes in the skim trees only the line-haul attributes.

Legs that have mode 3 are never considered part of the line-haul. The only exception is in the high-density area, such as the CBD, where the detailed sidewalk network (mode 1) and fine zone system are used. If program (TPATH) encounters a 3-1 or 1-3 mode sequence, then those legs are also considered part of the line-haul. This preserves the user-coded travel impedance, which is considered sufficiently accurate. Intrazonal variability in such areas is assumed to be small and can be ignored.

Legs that have modes 6, 7, or 8 (express bus and rail) are always considered part of the line-haul and are never removed from the network. If the highest numbered mode used in the path is 5 or less (local or suburban bus as the priority mode), legs that have modes 4 or 5 are considered part of the line-haul.

If mode 6, 7, or 8 is the priority mode in the path, legs that have mode 4 at the beginning or end of the line-haul portion are considered either approach (access-egress) or line-haul legs, depending on the user-specified criteria for either end. The decision to remove modes 4 and 5 from the line-haul portion depends on input values of two parameters:

1. CRTIME--If the in-vehicle time on modes 4 or 5 is greater than the criterion CRITIME, then they are considered part of the line-haul or
2. CRATIO--If the ratio of the in-vehicle time on modes 4 or 5 to the total path's in-vehicle time is greater than criterion CRATIO, then they are considered part of the line-haul.

By using these two parameters it is possible to preserve transit legs of certain length at either end as part of the line-haul.

If a leg that has mode 2 (automobile connector) is the first leg in the path, then it is considered part of the line-haul. Recall that automobile connectors are coded only if access at the origin end is not possible by using walk and feeder bus modes. Automobile egress is never permitted in the path building.

Determination of Choice of Access Modes

The choice set generated for a pseudoobservation depends on the use of the first line-haul mode as determined by TPATH. If the first line-haul mode is 4 or 5, then only walk access is considered. If the first line-haul mode is 6, 7, or 8, then walk, feeder bus, kiss-and-ride, and park-and-ride options are considered for automobile-owning households and walk and feeder bus for non-automobile-owning households.

Determination of Choice of Egress Modes

The choice set for pseudoobservation for egress is based on the use of the last line-haul mode. If the last line-haul mode is 4 or 5, then only walk egress is considered. If the last line-haul mode is 6, 7, or 8, then walk and feeder bus, if available, are considered for egress.

Determination of Choice of Transit Line-Haul Paths

The determination of choice for line-haul travel depends on the priority mode along three sets of paths (full network, network excluding rail service, and network excluding express bus service) as analyzed by using TPATH. The identification of choice set is done as follows. If the priority mode is 5 or less on full network (network 1), then the only alternative to automobile is the local or suburban bus. Estimated transit trips are assigned to network 1. If the priority mode is 6 on network 1 and the path in network 1 is identical to the path in network 2 (no rail present), two cases are possible, namely:

1. No connection in network 3 (no express bus present) is found; therefore, no rail alternative to express bus exists between the O-D pair in question. No line-haul choice is analyzed and trips are assigned to network 2.
2. Connection in network 3 is present. If the connection in network 3 has a priority mode of 7 or 8, then the transit split is estimated and trips are further split between network 2 and 3. If the priority mode in network 3 has a priority mode of 5 or less, no line-haul split is estimated and all trips are assigned to the express bus path.

If the priority mode is 7 or 8 on network 1, three cases are possible, namely:

1. No priority mode 6 connection in network 2; in this case no competing express bus service is identified and no line-haul split is performed. All transit trips are assigned to network 3.
2. Paths on network 1 and network 3 are not

identical; in this case, the network 1 path includes both rail and express bus legs, and thus is better than the rail-only path on network 3. No line-haul split is performed. All transit trips are assigned to network 1.

3. Paths on network 1 and network 3 are identical and priority mode on network 2 path is mode 6; in this case there is a choice between express bus and rail. The transit trips are split between two line-haul choices and are assigned to networks 2 and 3.

The determination of number of line-haul paths to be analyzed is done by MSPLIT by using zone-to-zone skim tree matrices produced by computer program TPATH.

Simulation of Access-Egress Impedance

TPATH trims those legs from the transit path that are considered approach links (access-egress). For pseudoobservations generated between an O-D pair, it is assumed that line-haul disutility components are identical for observations and the variation exists mainly in access-egress components. It is further assumed that the bulk of the intrazonal variability can be described by the variability of distance between trip ends and transit stops or stations. Thus, the input to MSPLIT (13) includes a description of the frequency distributions of distance to transit. Separate distributions can be specified for each zone, transit mode, and residential and nonresidential trip end. The distributions are specified in terms of the type of the distribution function and its parameters. Any of the five distributions can be used--linear, bilinear, step, bounded normal, or bounded exponential.

In the process of generating a pseudoobservation the program samples the distribution that corresponds to the first line-haul mode and the origin zone. The resulting distance is converted into access submodes service measures such as walk time, in-vehicle feeder bus time, and automobile-in-vehicle time by using user-specified zonal speeds. The total access impedance by each of the modes is determined by considering all the other standard components, such as feeder bus waiting time and automobile parking costs, as listed in Table 1. These elements are specified at the zonal level. A similar procedure is used to determine the egress impedance at the destination.

Determining the Frequency Distributions of Distance to Transit

The Northeast Ohio Areawide Coordinating Agency mode split procedure provides a default method to estimate the distributions based on available or easily obtainable data. The default procedures are sufficiently accurate for most standard cases. They were formulated based on a simulation analysis and validated by comparison to manually derived distributions (16,17). The default procedure is driven by a separate computer program, DFREQ (13). In standard applications the default distributions are used in the majority of cases, with user-determined distributions for areas that have odd-shaped service or are of a special interest.

Two types of transit service are considered--continuous and discrete. Continuous service is characteristic of local buses and express buses that operate in the collector-distributor phase, when they stop frequently to serve passengers. Discrete service is characteristic of rail lines and express buses that operate in the line-haul phase, when they stop only at a few designated locations. Gur showed (5) that frequency distribution of distance to continuous service can best be described as a linear function. The distribution to discrete service can best be described by a bounded normal distribution. DFREQ determines the distributions' parameters.

Parameters of the Linear Distribution Function and Their Estimation

The parameters of the linear function are as follows:

XMIN--minimum distance to the continuous bus service,
XMAX--maximum distance to the continuous bus service, and
Slope R--ratio of probability to walk distance XMAX to XMIN.

The parameters are estimated as a function of density of service (route miles of service operating per square mile of the zonal area) and activity concentration (that is, the extent to which trip ends are concentrated near the transit service).

The activity concentration factor can either be specified by the user or determined inside the program as a function of the percentage of the developed area in the zone. Another option is to specify different concentration factors for residential and nonresidential trip ends in order to account for the higher propensity of commercial areas to locate in accessible locations.

Parameters for Bounded Normal Distribution and Their Estimation

The parameters of the bounded normal distribution are as follows:

DMIN--minimum distance to discrete service,
DMAX--maximum distance to discrete service, and
SIGMA--standard deviation of parent complete normal distribution.

The discrete transit service (express bus, rail-rapid transit, or both) available for each traffic-analysis zone is described by specifying up to three nodes for each mode on the transit network that serves the zone in question. A separate station data (SDATA) file is coded, which gives X and Y coordinates of each transit node and zone centroid. The program DFREQ estimates the parameters of the distribution by assuming a square zone and calculating the distribution of distance from the zone's area to the closest transit station. For further details see elsewhere (5).

Depending on the nature of analysis, MSPLIT can be used to estimate either zonal transit trip ends or zone-to-zone transit trips (transit trip table) for assignment purposes. The sampling logic used to generate pseudoobservations is fully described elsewhere (5).

MSPLIT also saves the attributes of automobile and transit modes simulated for a pseudoobservation in a sample file. By manipulating attributes such as travel time or travel cost in the sample file, MSPLIT can be rerun rather inexpensively to assess the impact on mode split.

TRANSIT TRIP ASSIGNMENT

Since the modeling procedure described uses three sets of paths (full network, network with no rail, and network with no express bus service) and produces three path-specific transit trip tables, a special trip-loading sequence is necessary by using computer program ULOAD (2). The loading job consists of three steps. In the first step the trip table created by using full network is loaded to the paths created by using full network. The resulting partly loaded legs are saved and serve as an input to the next step. In the second step the trip table created by using network without rail is assigned to the partly loaded network from step 1 on the paths described by network with no rail service included. The resulting leg file is used as an input to the last stage of the process, where the trip table created by using network without express bus is assigned, by using paths described by network with no express bus service included. This multiple loading option is facilitated by the LEGS2 option of the UTPS program ULOAD.

CONCLUSIONS

The modeling process described provides a flexible framework for analyzing multiple options of access-egress modes available to a tripmaker. When present, the process provides a mechanism to split transit travel between two line-haul modes that serve an O-D pair. The disutility expressions used in the nested logit model use explanatory variables that are commonly used in network and trip generation analysis for a metropolitan area and are suited for long-range policy planning as well.

There are a number of obvious advantages to using this modeling process. First, it permits the analysis of policy issues that relate to selecting the service attributes of different transit modes that serve the same areas. Second, the inputs to the program describe easily measurable attributes of the transit system. The procedure relieves much of the weight of the approach link coding, which in standard models is a major determinant of the transit network loading.

ACKNOWLEDGMENT

We would like to thank Frank Koppelman of Northwestern University, who contributed to the model's formulation in the early steps of the project, and Jende Augustine Hsu of NOACA staff, who assisted in the model calibration.

REFERENCES

1. Program Documentation, Urban Transportation Planning. March 1972.
2. UTPS Users Guide. Urban Mass Transportation Administration; Federal Highway Administration, Feb. 1977.
3. Urban Transportation Planning, General Information and Introduction to System 360. Federal Highway Administration, March 1972.
4. Cambridge Systematics, Inc. Introduction to Urban Travel Demand Forecasting: Vol. 1, Demand Modeling, Software Systems Development Program. Urban Mass Transportation Administration, March 1974. NTIS: PB-236848.
5. John Hamburg and Associates; Urban Systems, Inc. NOACA Mode Split Modeling Study--Final Draft Report. Northeast Ohio Areawide Coordinating Agency, Cleveland, July 1981.
6. Y. Gur, E. Ryan, A. Vyas, and E. Lowe. Estimation of Demand for Public Transportation. TRB, Transportation Research Record 728, 1979, pp. 76-79.
7. Y. Gur; Urban Systems, Inc. The NCTCOG Modal Split Model. North Central Texas Council of Governments, Arlington, June 1980.
8. T.E. Lisco. The Value of Commuters' Travel Time--A Study in Urban Transportation. Univ. of Chicago, Ph.D. dissertation, 1967.
9. P. Stopher and T. Lisco. Modeling Travel Demand: A Disaggregate Behavioral Approach-Issues and Applications. Proc., 11th Annual Meeting of Transportation Research Forum, 1970, pp. 195-214.
10. T.A. Domencich and D. McFadden. Urban Travel Demand. North-Holland-Elsevier, New York, 1975.
11. F. Speilberg. Auto Occupancy Projections Using a Modal Split Model. Seven County Transportation Study, Cleveland, 1968.
12. F.S. Koppelman and M.E. Ben-Akiva. Aggregate Forecasting with Disaggregate Travel Demand Models using Normally Available Data. Paper presented at the World Conference on Transport Research, Rotterdam, Netherlands, April 26-28, 1977.
13. Urban Systems, Inc. Program Writeups for Computer Programs MSPLIT, TPATH, and DFREQ. Northeast Ohio Areawide Coordinating Agency, Cleveland, May 1981.
14. Trip Generation Analysis--Volume 2, Methodology for Estimating Joint Distribution of Household Size and Auto Ownership. Northeast Ohio Areawide Coordinating Agency, Cleveland, Oct. 1980.
15. Trip Generation Analysis--Volume 1, Home Based Person Trip Production Models. Northeast Ohio Areawide Coordinating Agency, Cleveland, May 1981.
16. Y. Gur; Creighton, Hamburg, and Associates, Inc. The Distribution of Distance to a Linear Service from a Square Zone. Chicago Area Transportation Study, 1975.
17. S. Howe. Dallas North Central Subarea Transportation Study--Travel Model Development Report. North Central Texas Council of Governments, Arlington, (forthcoming).

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

Transferability Analysis of Disaggregate Choice Models

FRANK S. KOPPELMAN AND CHESTER G. WILMOT

The transferability of disaggregate choice models is widely assumed in travel demand analysis and forecasting. However, research results are mixed in their assessment of transferability. This paper considers transferability from the perspective of the usefulness of information provided by a model that predicts in a context different from that in which it is estimated. Issues that influence transferability are discussed and methods are formulated to evaluate transferability. These methods are demonstrated by application to spatial transfer between pairs of geographic sectors within a single urban area by using identically specified models. This application provides useful insights into the transferability of choice models. We observe an inconsistency between general measures of error that indicate that transferability in this context is appropriate and statistical analyses that reject hypotheses that support transferability. Model transferability is a property of the estimation and application contexts as well as the specification of the model. Transferability is substantially improved by adjustment of alternative-specific constants. These results indicate the need for additional research to identify the conditions of model specification and context characteristics for which model transfer is effective. Directions for such research are identified.

The transfer of a model is the application of a model formulated and estimated in one context to another context. Transferability implies that the model transferred can provide useful information about the behavior or phenomenon of interest in the application context. The transferability property is commonly invoked implicitly when models are estimated on historic data and used to predict into the future. The transferability property is invoked explicitly when models are estimated in one area and used to predict in another area.

The purpose of this study is to develop an approach to examine the transferability of disaggregate travel choice models. We describe this approach and demonstrate it by application to the intraregional transfer of a disaggregate model of mode choice to work. This application demonstrates the usefulness of the proposed approach when studying the transfer of a specific model specification between different spatial contexts.

ISSUES IN MODEL TRANSFERABILITY

Model transferability has been discussed extensively in the travel demand analysis and prediction literature. Yet, there is little agreement on the definition of transferability or on the circumstances in which it is appropriate. A set of issues that need to be addressed includes the following.

First, we define transfer as the application of a model, information, or theory about behavior developed in one context to describe the corresponding behavior in another context. We further define transferability as the usefulness of the transferred model, information, or theory in the new context.

Second, we identify two general conditions, one theoretical and one practical, for effective model transferability. The theoretical condition for model transferability is that the underlying behavioral process described by the model is the same in the application context as in the context in which the model was estimated. If this condition does not hold (if, for example, people in one context are utility maximizers and people in another context are satisficers), models will not be transferable between the contexts. However, if this condition does hold, a further practical condition for model transferability is that the model be well-specified and that the data used to estimate it are such that the model describes the underlying behavioral process.

Third, transferability is not satisfactorily described as a dichotomous property. Rather, it is appropriate to consider the degree of transferability of a model, theory, or information from one context to another (1). Thus, we must develop measures that describe transferability in continuous terms.

Fourth, different portions of a model, theory, or data may be more or less transferable than other portions. Thus, it is appropriate to consider the notion of partial transferability (2). That is, it is appropriate to evaluate separable components or portions of a model for transferability to a new context.

Fifth, we distinguish between prior and posterior analysis of transferability. Posterior analysis of transferability is the determination that a model, theory, or data would have been transferable to an application context after observing the formulation and estimation of an analogous model, theory, or data in the application context. Prior analysis of transferability is the determination that a model, theory, or information is transferable before undertaking a similar development effort in the application context. Posterior analysis of transferability may be used as a basis to make prior inference of transferability in some new context. All empirical studies undertaken to date have been posterior studies. The study reported in this paper is also a posterior study; however, it is undertaken as a part of an overall effort to develop a capability to make prior inferences about transferability.

Sixth, we consider two distinct classes of tests for model transferability--both applied to posterior studies of transferability. Tests of model parameters are designed to evaluate, either subjectively or statistically, the degree to which the transferred model describes the behavioral process in the application context. Tests of model predictions are designed to evaluate, either subjectively or statistically, the accuracy with which predictions of a transferred model describe travel behavior in an application context.

The work reported here addresses each of the issues. We develop and apply methods to evaluate the transferability of disaggregate travel choice models. We assume the existence of behavior equivalence (i.e., equivalent behavioral processes in each context) and evaluate the transferability of a model that has demonstrated its ability to reproduce travel behavior in other studies. We consider transferability of portions of the model parameters. We examine posterior transferability based on tests of parameter equivalence and prediction usefulness.

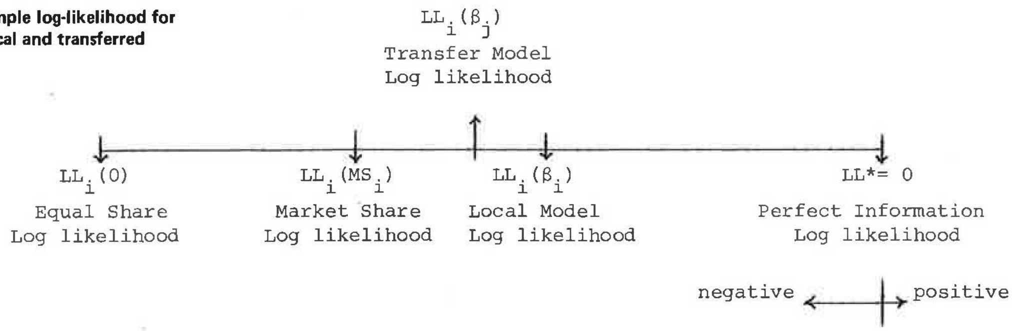
METHODS FOR EVALUATING MODEL TRANSFERABILITY

A set of measures that may be used to evaluate model transferability includes the following classes:

1. Tests of model parameter equality,
2. Tests of disaggregate prediction, and
3. Tests of aggregate--zonal level--prediction.

The first set of measures is based on conventional tests of equality between model parameters. These tests are symmetric between contexts. The second and third sets of measures describe the predictive accuracy of the transferred model and can be formulated as absolute tests or relative tests. The

Figure 1. Sample log-likelihood for alternative local and transferred models.



absolute measures are based on some comparison between transfer model prediction and observed behavior. The relative measures adjust these comparisons by the accuracy of prediction that would be obtained by a similarly specified model calibrated in the application context. Aggregate transfer measures differ from disaggregate measures because they are influenced by the distribution of exogenous variables in the aggregate group.

Model Parameter Equality

We hypothesize that the underlying choice process in two or more contexts can be described by a common model, the model equality test statistic (METS). This is defined by

$$METS_{ij} = -2[LL_{i\cup j}(\beta_{i\cup j}) - LL_i(\beta_i) - LL_j(\beta_j)] \quad (1)$$

where $LL_i(\beta_j)$ is the log likelihood that the behavior observed in context i was generated by the model estimated in context j and $i\cup j$ describes the union of contexts i and j.

This test is analogous to the commonly used test of equality of models between market segments (3). In this case the market segments of interest are the pair of contexts between which transferability is being considered. The resultant statistic is chi-square distributed with degrees of freedom equal to the number of model parameters.

The METS statistic can be used to test the equality of the entire set of model parameters or a subset of model parameters. In particular, we may consider the case where some of the parameters of the model are assumed to be context dependent and others are assumed to be equal across contexts. Atherton and Ben-Akiva (2), McFadden (4), and Ben-Akiva (5) discuss the case of context-specific alternative-specific constants. In this case the hypothesis to be tested is that the underlying model is equal across contexts with respect to a selected subset of parameters. The resultant test is chi-square distributed with degrees of freedom equal to the number of parameters tested for transfer.

These tests have been used in earlier studies of transferability (6-9). Unfortunately, these tests have an important deficiency in transferability analysis. This deficiency is the inherent symmetry of the tests, whereas transferability is a directional property. To observe this point, consider two contexts; one is a large urban region with a wide range of population groups and the other is a suburban area of an urban area that has relatively little diversity. Although it might be appropriate to use a model estimated in the first context for prediction in the second, it is unlikely that the reverse transfer will be useful.

Disaggregate Measures of Transferability

We formulate a set of transferability measures based on the ability of a transferred model to describe individual observed choices in the application context. These measures are based on the generally used log-likelihood measure. Specifically, we define the log of the likelihood that the observed data in application context i were generated by the transferred model estimated in context j $[LL_i(\beta_j)]$. We examine this log likelihood of the transferred model relative to the log likelihood for a null (equal shares) model $[LL_i(0)]$; the log likelihood for a market share model $[LL_i(MS_i)]$; the log likelihood for a model estimated in the application context $[LL_i(\beta_i)]$; and the log likelihood of a perfect model $[LL_i^*]$, which is equal to zero. The relation between these measures is shown in Figure 1.

A natural measure of the transferability of the model estimated in context j for application in context i is the difference in likelihood between this model and a corresponding model estimated in context i, $-[LL_i(\beta_j) - LL_i(\beta_i)]$. We use this measure to formulate three specific indices of transferability.

First, we define the transferability test statistic (TTS) as twice the difference in log likelihoods identified above:

$$TTS_i(\beta_j) = -2[LL_i(\beta_j) - LL_i(\beta_i)] \quad (2)$$

This statistic is chi-square distributed with degrees of freedom equal to the number of model parameters under the assumption that the parameter vector of the transferred model is fixed. This test is used by McFadden and others (10) in their conditional choice set test of the IIA property and by Atherton and Ben-Akiva (2) in their tests of transferability between Washington, D.C., and New Bedford, Massachusetts.

This statistic tests the hypothesis that the underlying parameter values in context i are equal to the estimated values in context j. It is equivalent to the model equality test statistic when there is no error in the transferred parameter estimates. Otherwise, it will have larger values than the model equality test statistic and thus will be more likely to reject the equality hypothesis. The transferability test statistic for model j applied to context i is, in general, not equivalent to the corresponding statistic for model i applied to context j. Thus, it is possible and reasonable to accept transferability in one direction between a pair of contexts but reject it in the other.

Second, the transfer index (TI) describes the degree to which the log likelihood of the transferred model exceeds some base or reference model (we use the market shares model) relative to the improvement provided by a model developed in the application context. We define TI by

$$TI_i(\beta_j) = [LL_i(\beta_j) - LL_i(MS_i)]/[LL_i(\beta_i) - LL_i(MS_i)] \quad (3)$$

This index measures the predictive accuracy of the transferred model relative to a locally estimated model. TI has an upper bound of one, which it obtains when the transferred model is as accurate as a locally estimated model. This index does not have any lower bound. Negative values imply that the transferred model is worse than the local base (market shares) model.

Third, the transfer rho-square (ρ_T^2) describes the degree to which the log likelihood of the transferred model exceeds that of the base model relative to the degree of improvement in log likelihood achieved with a perfect (predicts all choice correctly) local model. This measure is analogous to the commonly used rho-square measure (3). We define the transfer rho-square measure by

$$\rho_i^2(\beta_j) = [LL_i(\beta_j) - LL_i(MS_i)]/[L_i^* - LL_i(MS_i)]$$
$$= 1 - [LL_i(\beta_j)/LL_i(MS_i)] \quad (4)$$

This measure is related to TI by

$$\rho_i^2(\beta_j) = TI_i(\beta_j)\,\rho_i^2(\beta_i) \quad (5)$$

Accordingly, it is upper bounded by the local rho-square measure, has no lower bound, and negative values are interpreted as for the transfer index.

The three measures defined above are interrelated by their dependence on the difference in log likelihood between the transferred and local models. However, they provide different perspectives on model transferability. The transfer rho-square provides an absolute measure of disaggregate transferability, the transfer index provides a relative measure, and the transfer test statistic provides a statistical test measure.

Each of these measures may be applied to tests of partial model transferability by substitution of the log likelihood for the partly transferred model (a model with some transferred parameters and some locally estimated parameters) in place of the log likelihood for the transferred model in Equations 2, 3, and 4. The partial transfer log likelihood will always lie between the transfer model log likelihood and the local model log likelihood in Figure 1.

Aggregate Measures of Transferability

The planning process is primarily concerned with the prediction of aggregate rather than disaggregate travel flows. Thus, it is appropriate to consider transferability in terms of the accuracy of aggregate predictions. We define the error in aggregate prediction and examine ways in which these errors can be summarized across alternatives and aggregate groups.

We choose to examine the following relative error measure for prediction of alternative choice frequency in some aggregate group:

$$REM_{mg} = (\hat{N}_{mg} - N_{mg})/\hat{N}_{mg} \quad (6)$$

where

REM_{mg} = relative error measure in prediction of alternative m for group g,

$\hat{N}_{mg}$ = number of persons predicted to choose alternative m from group g, and

N_{mg} = number of persons observed to choose alternative m from group g.

In order to evaluate the aggregate predictive accuracy of a choice model we summarize this measure over alternatives and groups by means of the weighted root mean square error (RMSE) measure, defined by

$$RMSE = \left(\sum_{m,g} \hat{N}_{mg} REM_{mg}^2 / \sum_{m,g} \hat{N}_{mg}\right)^{1/2} \quad (7)$$

This measure is an index of the average relative error in prediction weighted by the size of the prediction element and structured to place emphasis on large relative errors. RMSE can be disaggregated into alternative-specific error measures and into average and variational components to aid error analysis. These properties and their use in transportation error analysis are described by Koppelman (11-13).

An alternative measure of the accuracy of aggregate prediction tests the hypothesis that the observed frequencies of choice in each group are, collectively, generated by the prediction model. We formulate the aggregate prediction statistic (APS) as

$$APS = \sum_{m,g} \hat{N}_{mg} REM_{mg}^2$$
$$= \sum_{m,g} (\hat{N}_{mg} - N_{mg})^2/\hat{N}_{mg} \quad (8)$$

This statistic, which is equivalent to the chi-square one sample test (14), is chi-square distributed under the assumption the $\hat{N}_{mg}$ is predicted without sampling error. This is equivalent to the assumption adopted in formulating the transferability test statistic.

APS is more likely to reject the hypothesis that all frequencies come from the candidate model than would a statistic that takes account of sampling variation. The degrees of freedom for the APS for full model transfer are (number of alternatives - 1) x (number of groups). However, when applied for local prediction or with locally adjusted alternative-specific constants, the degrees of freedom need to be reduced by the number of alternatives less one to (number of alternatives - 1) x (number of groups - 1).

A relative measure of aggregate prediction accuracy is useful. We define the relative aggregate transfer error (RATE) measure as the ratio between the transfer RMSE and local RMSE measures,

$$RATE = RMSE_i(\beta_j)/RMSE_i(\beta_i) \quad (9)$$

These measures are interrelated by their dependence on the relative error measure defined in Equation 6. However, they offer different perspectives on model transferability at the aggregate level. RMSE provides an absolute measure of aggregate transferability, RATE provides a relative measure, and APS provides a statistical test measure.

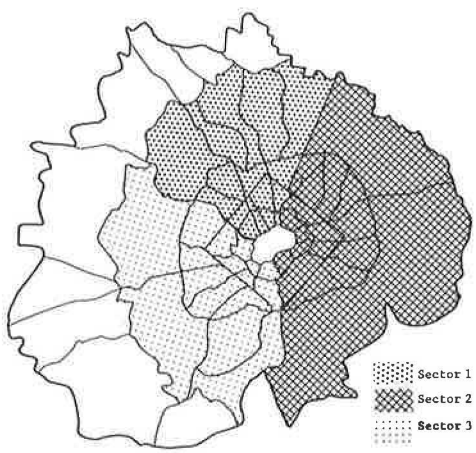
APPLICATION OF METHODOLOGY

We demonstrate the use of the transferability measures by their application to the transfer of mode choice to work models for the Washington, D.C., area. These models describe the choice among drive alone, shared ride, and transit alternatives for breadwinners who work in the central business district (CBD). There are a total of 2654 such breadwinners, 2088 of whom have all three alternatives available and 566 of whom do not have drive alone available due to lack of a driver's license or lack of available cars in the household. We divide the population into three groups by geographic sector as shown in Figure 2.

Table 1. Estimation results for sectors and the region.

	Sector 1 (n = 944)		Sector 2 (n = 964)		Sector 3 (n = 746)		Region (n = 2654)	
Item	Parameter	t-Value	Parameter	t-Value	Parameter	t-Value	Parameter	t-Value
DAD	-3.39	-9.2	-1.84	-5.3	-2.73	-6.8	-2.59	-12.4
SRD	-2.46	-8.8	-2.10	-8.5	-2.52	-7.3	-2.35	-14.8
CPDDA	4.13	11.1	3.07	8.8	3.58	9.0	3.59	17.0
CPDSR	2.05	7.3	1.77	8.2	1.59	5.0	1.83	12.4
OPTCINC	-0.0069	-0.5	-0.0242	-2.1	-0.0280	-1.7	-0.0232	-3.2
TVTT	-0.0418	-6.9	-0.0151	-3.2	-0.0223	4.6	-0.0243	-8.3
OVTTD	-0.0258	-0.4	-0.105	-1.8	-0.0421	-0.5	-0.0667	-1.85
GWSR	0.746	4.8	0.526	3.6	0.680	4.2	0.653	7.4
NWORKSR	0.096	0.9	0.264	2.7	0.502	4.1	0.268	4.4
Log likelihood								
At zero	-962.5		-933.8		-790.0		-2686	
At market shares	-904.4		-898.5		-771.6		-2587	
At convergence	-754.4		-803.4		-688.3		-2266	
Likelihood ratio statistic	416.2		260.7		203.4		840.8	
Likelihood ratio index								
ρ^2 (0)	0.216		0.140		0.129		0.157	
ρ^2 (market share)	0.166		0.106		0.108		0.124	

Figure 2. Estimation sectors in Washington, D.C., region.



The demographics of these CBD-based breadwinners in sectors one and three are similar. Those in sector two are generally younger and come from households that have more persons, more workers, lower household income, fewer cars, and fewer licensed drivers. Average travel service characteristics are similar across sectors for drive alone and shared ride. Transit service is most expensive in sectors one and three and most time consuming in sector three.

Method of Analysis, Model Specification, and Parameter Estimates

We study the various transferability measures and tests in application to the transferability of work-mode-choice-to-CBD models between pairs of sectors depicted in Figure 2. We estimate models by using the specification employed by Koppelman (7,12,13) in his analysis of aggregation error in prediction with disaggregate choice models.

Variable Name	Description
DAD, SRD	Dummy variable specific to drive-alone and shared-ride alternative; measures average bias between pairs of alternatives other than that represented by the included variables
CPDDA, CPDSR	Cars per driver included separately as alternative-specific variables for the drive-alone and shared-ride modes; measures change in bias among modes due to changes in automobile availability within the household
OPTCINC	Round-trip out-of-pocket travel cost divided by income [cents/($1000/year)]; measures effect of travel cost on mode utility with cost effect modified by household income level
TVTT	Round-trip total travel time (min); measures linear effect of combined in- and out-of-vehicle travel time in mode utility
OVTTD	Round-trip out-of-vehicle travel time divided by trip distance (min/mile); measures additional effect of out-of-vehicle travel time in utility in addition to the effect represented in TVTT; this added effect is structured to decline with increasing trip distance
GWSR	Dummy variable that indicates if the breadwinner is a government worker specific to the shared-ride alternative; measures effect on shared-ride utility of shared-ride incentives for government workers
NWORKSR	Number of workers in the household specific to the shared-ride alternative; measures change in utility of shared ride when there is an opportunity to share ride with a household member

The estimated models for each sector and for the region are reported in Table 1. All models are highly significant overall and the parameters for all models are significant at the 0.01 level, except for OPTCINC and OVTTD for all sectors and NWORKSR for sector one.

The overall goodness of fit as measured by the rho-square evaluated at zero or market share is low for disaggregate choice models. However, studies by various researchers have found this general specification to be satisfactory for analysis of this data set and similar specifications have been employed for many of the disaggregate work mode choice studies reported in recent years (2,9,11).

Model Parameter Equality

The model equality test statistics for the complete model and for the partial model (all parameters except alternative-specific constants) between pairs of sectors and for all sectors jointly reject the equality hypothesis at the 0.05 level for two of three sector pairs (except sectors 2 and 3) and for the set of three sectors jointly. Based on this result we would reject the hypothesis of model equality. By implication, this would suggest the rejection of model transferability.

Disaggregate Measures of Transferability

The transferability test statistics (Equation 3) for full and partial model transfer are reported in Table 2. The reported values for the transfer of the full and partial models are all significant at the 0.01 level except for the partial model transfer from sector 2 to sector 3. These results reject strongly the hypothesis of intraurban transferability for the model used in this application.

There are other interesting observations to be made from the results in Table 2. First, as expected, the transferability test results are not symmetric. In fact, the transfer directionality is quite large. Second, improvements in the TTS value when moving from transfer of the total to partial model transfer are large in almost every case.

TI (Equation 5) for full and partial transfer is reported in Table 3. The TI varies from 0.61 to 0.89 for full model transfer and from 0.76 to 0.94 for partial model transfer. The index values increase dramatically when moving from transfer of the total model to transfer of the partial model for all transfers except sector two to sector one, which already had a high value. These results indicate that, when alternative-specific constants are adjusted to match application choice shares, the transfer model can provide a high proportion of the information that would be obtained by estimation of a model in the application context.

The transfer rho-square measures are not reported here. They can be obtained from the model rho-square market share measure and TI (Table 3) by Equation 5.

Table 2. Transferability test statistic.

	Prediction Based on					
	Sector 1		Sector 2		Sector 3	
Estimate Based on	Full Model	Partial Model	Full Model	Partial Model	Full Model	Partial Model
Sector 1			51.2[a]	45.4[a]	65.0[a]	31.6[a]
Sector 2	34.8[a]	34.4[a]			27.8[a]	11.4
Sector 3	54.8[a]	31.2[a]	34.0[a]	15.2		

[a]Significant at the 0.01 level.

Table 3. Transfer index.

	Prediction Based on					
	Sector 1		Sector 2		Sector 3	
Estimate Based on	Full Model	Partial Model	Full Model	Partial Model	Full Model	Partial Model
Sector 1	1.0	1.0	0.73	0.76	0.61	0.82
Sector 2	0.89	0.89	1.0	1.0	0.83	0.94
Sector 3	0.82	0.90	0.82	0.92	1.0	1.0

Table 4. RMSE.

	Prediction Based on					
	Sector 1		Sector 2		Sector 3	
Estimate Based on	Full Model	Partial Model	Full Model	Partial Model	Full Model	Partial Model
Sector 1	0.201	0.201	0.244	0.238	0.308	0.245
Sector 2	0.251	0.250	0.236	0.236	0.299	0.264
Sector 3	0.280	0.243	0.270	0.242	0.257	0.257

Table 5. APS for estimation-prediction sector pairs.

	Prediction Based on					
	Sector 1		Sector 2		Sector 3	
Estimated Based on	Full Model	Partial Model	Full Model	Partial Model	Full Model	Partial Model
Sector 1	38.1	38.1	57.3	54.7	70.8[a]	44.8
Sector 2	59.4[a]	59.0[a]	53.7	53.7	66.9[a]	52.0[a]
Sector 3	74.0[a]	55.7[a]	70.4[a]	56.5	49.1	49.1

[a]Significant at 0.01 level.

Aggregate Measures of Transferability

We now examine the aggregate prediction capability of transferred models. The sectors described in Figure 1 are subdivided into residence zones (16 in sector 1, 19 in sector 2, and 16 in sector 3). Aggregate predictions for mode choice in each residence zone are obtained by summing individual prediction probabilities for each alternative (13). These predictions are compared with the observed travel mode choices to compute the relative error measure (REM) defined in Equation 6 for each mode and zone. These error measures are combined over modes and zones to obtain the aggregate error measures.

The aggregate error in applying each sector model (with and without alternative-specific constants) to each sector by using RMSE is reported in Table 4. Based on RMSE, all of the predictions give reasonably accurate estimates of aggregate mode share. The errors for transfer models are not dramatically greater than those for local models, especially when alternative-specific constants are adjusted to match local data.

Next, we use APS, defined in Equation 8, to test the hypothesis that the aggregate choice frequencies by mode for the zones in each sector are generated by the models estimated in each sector (Table 5). The hypothesis that the observed choice frequencies are generated by the models tested is rejected at the 0.01 level for model transfer in five of six cases for the full model and three of six cases for the partial model. Thus, the APS analysis rejects the hypothesis of model transferability.

We do not report RATE. These measures can be obtained from RMSE for transfer prediction and local model prediction by Equation 9.

SUMMARY AND CONCLUSIONS

This paper develops a methodology and related measures to be used in the analysis of transferability effectiveness. The measures developed include both indices and statistical tests applicable at either the disaggregate or aggregate level. The measures developed are applied to the analysis of intraregional transferability of disaggregate CBD-work mode choice models estimated on geographic portions of a common data set by using identical specifications. The application illustrates the differences in substantive interpretations and conclusions that can be obtained by use of different measures. It also identifies the variability in transfer effectiveness that exists even within this relatively narrow range of estimation and application contexts.

Conclusions from Relative Transfer Measures

The relative transfer measures indicate that transferred models have relatively small error compared with that incurred by use of local models. TI indicates that transferred models provide at least 80 percent of the information provided by local models in four of six full model transfers and five of six partial model transfers. RATE (ratio of transfer and local RMSE) indicates that the use of transferred models increases aggregate error by 20 percent or less for four of six full model transfers and 10 percent or less for four of six partial model transfers. These results suggest that it is reasonable to conclude that models are transferable between these geographic sectors.

Conclusions from Statistical Tests

The statistical tests generally reject hypotheses that are consistent with transferability. The METS test rejects model equality at the 0.05 level for two of three sector pairs. The TTS test rejects model transfer at the 0.01 level for all full model transfers and five of six partial model transfers. The APS test rejects model transfer at the 0.01 level for five of six full model transfers and three of six partial model transfers. These results suggest that transferability between these geographic sectors should be rejected.

Transfer Error Importance Versus Significance

The results of magnitude of transfer error tests and those of transfer significance tests lead to different conclusions about transfer effectiveness. That is, the transfer errors are deemed to be unimportant in magnitude; however, hypotheses that support transferability are significantly rejected. This apparent inconsistency results from confusion in the literature between the observed magnitude of differences between transfer and local models and the statistical significance of such differences that reflect both the magnitude and the precision of the estimates and predictions obtained. Although statistical tests can be used to alert the planner or analyst to differences between models, they must be considered with reference to the magnitude of errors that are acceptable in each application context. Although the magnitude of prediction error attributable to either a local or transferred model depends on the distribution of explanatory variables in the application context, our experience suggests that the apparent inconsistency between statistical rejection and practically small differences is commonly observed.

Transfer Measures Sensitivity

The transfer measures formulated are highly sensitive to differences in sector pair transfer effectiveness. For example, the TI that measures transferability relative to the local model ranges from 0.61 to 0.89 for full model transfer and from 0.76 to 0.92 for partial model transfer. Thus, these measures appear to be able to discriminate among levels of transferability.

Asymmetry of Transferability Measures

The measures of transferability developed are consistently different, in some cases dramatically different, between the same pairs of sectors, depending on the direction of transfer. This indicates that transferability is not determined solely by differences between sectors but also by the identity of the estimation and application contexts.

Adjustment of Alternative-Specific Constants

The transfer effectiveness measures improve substantially when alternative-specific constants are updated to match choice shares in the application context. These results emphasize the importance of updating alternative-specific constants to take account of differences in the average effect of excluded variables between estimation and application contexts.

Model Goodness-of-Fit and Transfer Effectiveness

The order of models with respect to goodness-of-fit is sector one, sector two, and sector three. However, the sector-one model is not consistently the most-effective model for transfer. Specifically, the sector-two model transfers to sector three better than does the sector-one model by most of the measures employed.

Although it is generally recognized that models that have high goodness-of-fit are not necessarily well-specified and thus are not necessarily effective in transfer, goodness-of-fit measures are commonly used to guide the selection of model specification and the selection of models for application. In this case, all the models have identical specification. Our results indicate that selection of the context from which to draw an identically specified model for transfer application cannot be based on estimation context goodness-of-fit. This result motivates the need to identify characteristics of estimation and application contexts between which models may be effectively transferred.

Contextual Determinants of Transfer Effectiveness

Characteristics of the estimation and application contexts have an important influence on transferability. Research is needed to identify the degree to which contextual characteristics determine transfer effectiveness and the specific contextual characteristics that are important.

Specification and Transfer Effectiveness

It is argued in the literature that transferability improves with improved model specification (1,5,15). Although this view is reasonable, it has not been validated empirically. The understanding of transfer effectiveness will be enhanced by research into the relation between model specification and transfer effectiveness. Additional research is being undertaken to explore this relation.

Prior Prediction of Transfer Effectiveness

The transfer analysis reported here and in all previous research on the transferability of travel models is based on posterior analysis. That is, these studies examine the question, "Would it have been appropriate to transfer a specific model from a specific context to another specific context?", by comparison of transfer results to results obtained by application of a local model. The objective for the future is to use an understanding of the relation between model specification and characteristics of both estimation and application contexts to provide prior guidance about the probable transferability of different models estimated in different contexts for use in the application context of interest. This will be the focus of future research.

ACKNOWLEDGMENT

The work reported in this paper was supported by the Office of University Research, U.S. Department of Transportation. James Ryan of the Urban Mass Transportation Administration, who is contract technical monitor, provided technical advice in the development of this work. We also acknowledge the contributions of Eric I. Pas to the refinement of ideas in this paper. Comments by Joel Horowitz and Moshe Ben-Akiva contributed to refinements of certain statistical tests.

REFERENCES

1. S.R. Lerman. Interspatial, Intraspatial, and Temporal Transferability. In New Horizons in Travel Behavior Research (P.R. Stopher, A.J. Meyburg, and W. Brög, eds.), Lexington Books, D.C. Heath and Co., Lexington, MA, 1981.
2. T.J. Atherton and M.E. Ben-Akiva. Transferability and Updating of Disaggregate Travel Demand Models. TRB, Transportation Research Record 610, 1976, pp. 12-18.
3. D. McFadden. Conditional Logit Analysis of Qualitative Choice Behavior. In Frontiers in Econometrics (P. Zarembka, ed.), Academic Press, New York, 1973, Chapter 4.
4. D. McFadden. Properties of the Multinomial Logit (MNL) Model. Urban Travel Demand Forecasting Project, Institute for Transportation Studies, Univ. of California, Berkeley, Working Paper 7617, Sept. 1976.
5. M.E. Ben-Akiva. Issues in Transferring and Updating Travel Behavior Models. In New Horizons in Travel Behavior Research (P.R. Stopher, A.H. Meyburg, and W. Brög, eds.), Lexington Books, D.C. Heath and Co., Lexington, MA, 1981.
6. R.A. Galbraith and D.A. Hensher. Intra-Metropolitan Transferability of Mode Choice Models. R. Travers Morgan; Macquarie Univ., Australia, 1980.
7. F.S. Koppelman. Intra-Urban Transferability of Disaggregate Choice Models. Transportation Center, Northwestern Univ., Evanston, IL, Aug. 1977.
8. P.S. McCarthy. Further Evidence on the Temporal Stability of Disaggregate Travel Demand Models. Department of Economics, Purdue Univ., Lafayette, IN, March 1981.
9. A. Talvitie and D. Kirshner. Specification, Transferability, and the Effect of Data Outliers in Modelling the Choice of Mode in Urban Travel. Transportation, Vol. 7, 1978, pp. 311-331.
10. D. McFadden, K. Train, and W.B. Tye. An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model. TRB, Transportation Research Record 637, 1977, pp. 39-46.
11. F.S. Koppelman. Travel Prediction with Models of Individual Choice Behavior. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Ph.D. dissertation, 1975.
12. F.S. Koppelman. Methodology for Analyzing Errors in Prediction with Disaggregate Choice Models. TRB, Transportation Research Record 592, 1976, pp. 17-23.
13. F.S. Koppelman. Guidelines for Aggregate Travel Prediction Using Disaggregate Choice Models. TRB, Transportation Research Record 610, 1976, pp. 19-24.
14. S. Siegel. Non-Parametric Statistics. McGraw-Hill, New York, 1956, p. 42.
15. J.J. Louviere. Some Comments on Premature Expectations Regarding Spatial, Temporal, and Cultural Transferability of Travel Choice Models. In New Horizons in Travel Behavior Research (P.R. Stopher, A.H. Meyburg, and W. Brög, eds.), Lexington Books, D.C. Heath and Co., Lexington, MA, 1981.

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

Wisconsin Work Mode-Choice Models Based on Functional Measurement and Disaggregate Behavioral Data

GEORGE KOCUR, WILLIAM HYMAN, AND BRUCE AUNET

This paper describes a series of mode-choice models developed by the Wisconsin Department of Transportation to assess transportation policy issues consistently across four sets of urban areas in the state. The models were developed by using a combination of functional measurement (or by asking respondents their likely mode choice in a series of situations) and disaggregate demand modeling (to calibrate the models and provide a test of the correspondence between stated and actual behavior). Bus, walk, bicycle, ridesharing, and drive-alone modes are included. Key variables include gasoline availability, gasoline price, queuing time to purchase gasoline, bicycle lanes, ridesharing programs, and transit service improvements. The models are being used in statewide policy analysis, for local planning, and for quick-response analysis. They represent an approach to demand analysis and may be an efficient and effective tool for examining other demand issues.

In a single statewide modeling study, the Wisconsin Department of Transportation (WisDOT) has developed work trip mode-choice models for four sets of urban areas of different character: one large city, one medium city, and two sets of small cities. These models permit WisDOT to address key policy issues by incorporating the effects of gasoline availability, gasoline price, queues for gasoline purchases, ridesharing programs, transit service improvements, bicycle lanes, and other factors. The models are estimated by a combination of functional measurement (also called conjoint analysis) and disaggregate demand modeling. Functional measurement models (1,2) are based on asking respondents their likely mode choice in a series of situations constructed from an experimental design. One or more situations closely resemble current conditions. We use a logit model to compare stated behavior under current conditions with actual behavior and adjust the models derived from the functional measurement task if there is a difference. The models are further refined by using sensitivity analysis.

The department undertook this statewide effort to enhance its ability to plan in a multimodal context. By administering similar surveys in all the urban areas of the state, it gained the ability to examine a broad range of urban transportation policies in a consistent manner. The department can now determine the absolute and comparative impacts of many policy proposals on driving alone, sharing a ride, walking, bicycling, and riding a bus.

Not only are the models useful for statewide policy analysis, they also enhance WisDOT's ability to provide technical assistance to urban areas in preparing transportation plans. Also, the pivot point and elasticity formulations of these models are being used for quick policy analysis. Finally, these urban work trip models complement a set of intercity mode and trip-frequency models developed earlier by using functional measurement (3,4). Ultimately, the department will have a comprehensive set of models for statewide policy analysis and system planning.

The functional measurement and disaggregate modeling methodology was devised to address WisDOT's forecasting requirements within a moderate budget level and relatively short time frame. Functional measurement was chosen because most key policy issues that face WisDOT cannot be captured readily in disaggregate models. Top administrators were specifically interested in learning the effects of gasoline rationing, long lines at gasoline stations, large increases in gasoline price and parking costs, improved bicycle facilities, and other issues not customarily addressed by demand models. Fuel price and availability exhibit no variability in the usual cross-sectional data sets because all individuals face the same conditions at a given point in time. In small cities bus fares are constant and virtually no parking fees are charged for work trips. Several modes of interest, such as bicycle facilities and commuter rail, are nonexistent in most areas. Finally, the data-collection effort for a statewide disaggregate model would be extensive.

Model validity was a strong concern, so we performed a second stage of analysis by using a logit model to further calibrate the original models. In this stage the forecasts derived from the functional measurement model for the status quo are compared with actual behavior, and the stated behavioral model is adjusted if there is a discrepancy. The calibration procedure can require fewer data than a traditional disaggregate model.

Two staff members completed the analysis in six months. An additional six months was needed to prepare reports and documentation, and some programming, keypunching, and consultant assistance were required.

FOCUS GROUPS

To begin the analysis four focus group interviews were held. The discussions of the focus group either verified the factors believed to be important a priori or suggested others to be treated in the qualitative analysis of mode split. The focus group consisted of 8-12 individuals who were convened for 1.5 h in a structured session. We obtained several interesting qualitative results. For example, individuals said their travel behavior was more sensitive to the change in the pump price of gasoline than to gasoline price per mile, which suggests that fuel efficiency was a consideration only when buying a vehicle. Also, participants of the focus group regarding bicycle travel said condition of the riding surface was a major concern, which was an unanticipated factor. In addition, many women said that under no circumstances would they stop driving alone to work because they had to carry groceries or children on the way to or from work. This suggested that sex and the number of children should be included in the final models to explain travel choice (5).

DESIGN OF EXPERIMENTS AND SURVEY

Six experiments were prepared to meet the objectives of the study. The four that pertain to ridesharing, walking, bicycling, and local bus service are reported in this paper. Two other experiments for express bus and commuter rail were also administered, but these modes are available to few travelers in Wisconsin. The experiment for ridesharing is illustrated in Figure 1, and other experiments are very similar. All surveys used drive-alone as the base mode.

A typical multivariable experimental model involves a series of independent variables that affect some dependent variable, such as mode choice. Each independent variable is considered at two or more values or levels, as designated by the experimental plan. In the ridesharing experiment gasoline price has four levels ($1.30, $1.70, $2.00, and $2.60), and the four other factors have two levels. The experiment is thus a $4^1 \times 2^4$ design.

The experimental results are analyzed to evaluate the statistical significance of the independent variables, estimate their effects, and establish functional relations. In conducting such analyses, one is interested in the main effect of each variable, that is, the effect on experimental response of going from one level of the variable to the next, all other variables being at their average values. In many situations the effect of two independent variables is not additive, and the variables are said to interact (i.e., the effect of one variable on the response depends on the value of some other variable).

A common multivariable experimental plan is the full factorial experiment, which consists of all possible combinations of levels for each of the variables. In our case, this would require $4^1 \times 2^4$, or 64 situations. A full factorial experiment permits one to obtain estimates of the effects of all possible interactions.

Many higher-order interactions can be assumed to be negligible, which leads, however, to a substantial reduction in the number of situations required. Such designs are called fractional factorial plans. In Figure 1 we use a one-eighth fraction, or only eight situations; this assumes that all interactions are negligible. This plan allows approximate estimates of the effects of a large set of policy variables in a relatively simple mailout survey, although it is at the expense of assuming a linear, additive model without interactions. This trade-off between survey complexity and model richness was made to ensure as high a response rate to the survey as possible, and to allow high

Figure 1. Ridesharing experiment.

Under what situations would you drive alone or share a ride (carpool or vanpool) to work?

Consider that you are going to work and that driving alone or sharing a ride in a car pool or van pool are your only choices.

Below are a number of factors describing eight different situations where you are faced with choosing whether to drive alone or share a ride to work.

Look at each situation across the entire line and please answer in the last column to the right how likely you are to drive alone or share a ride to work.

	AUTO FACTORS			CAR POOL/VAN POOL FACTORS		PLEASE– ANSWER IN THIS COLUMN: HOW LIKELY ARE YOU TO DRIVE ALONE OR SHARE A RIDE? (CIRCLE A NUMBER)				
	Gas Availability	Gas Price	Parking Cost to Drive Alone	People You Share A Ride With	Employee Work Schedule	Always Drive Alone	Probably Drive Alone	Indifferent	Probably Share A Ride	Always Share A Ride
SITUATION 1	Ample Supply	$1.30/gallon	Free	Co-Worker/ Neighbor	Flexi-time (hours can vary daily)	1	2	3	4	5
SITUATION 2	Ration of 10 gallons/week*	$2.60/gallon	Free	General Public (Carpool Matching)	Flexi-time (hours can vary daily)	1	2	3	4	5
SITUATION 3	Ration of 10 gallons/week*	$2.00/gallon	$30/month	Co-Worker/ Neighbor	Flexi-time (hours can vary daily)	1	2	3	4	5
SITUATION 4	Ample Supply	$2.60/gallon	$30/month	Co-Worker/ Neighbor	Fixed 8 hour day	1	2	3	4	5
SITUATION 5	Ration of 10 gallons/week*	$1.70/gallon	Free	Co-Worker/ Neighbor	Fixed 8 hour day	1	2	3	4	5
SITUATION 6	Ample Supply	$2.00/gallon	Free	General Public (Carpool Matching)	Fixed 8 hour day	1	2	3	4	5
SITUATION 7	Ample Supply	$1.70/gallon	$30/month	General Public (Carpool Matching)	Flexi-time (hours can vary daily)	1	2	3	4	5
SITUATION 8	Ration of 10 gallons/week*	$1.30/gallon	$30/month	General Public (Carpool Matching)	Fixed 8 hour day	1	2	3	4	5

*If your car gets 15 miles per gallon, you can travel 150 miles per week.

confidence in the responses received--both crucial considerations for statewide policy planning.

Catalogs of experimental designs are available in the literature (6,7). We developed our own simple designs. In addition to the experiment, each survey instrument contained background questions of two types. Some were questions concerning socioeconomic characteristics of respondents and thus were suitable for checking representativeness of the samples and measuring the sensitivity of mode choice to socioeconomic variables. The remainder gathered data on actual travel choices of individuals and the attributes of competing modes.

SURVEY ADMINISTRATION

The sizes of the survey sample were determined based on desired levels of sampling error and expected response rates. The sampling error was set at ±5 percent, with 95 percent confidence for categorical variables, particularly the 1-5 response scale in the experiments. A conservative 20 percent usable response rate was assumed. These considerations, applied to the number of cities and separate modes for which models were desired, resulted in the mailing of about 17 000 questionnaires.

WisDOT mailed the surveys to residents who renewed their drivers' licenses in August and September 1980. The gross response rate was 57 percent (9208 surveys), but some surveys had incomplete information. The usable response rate was 46 percent. Because we received more than double the expected response rate, we were able to exclude respondents who did not travel to work, so we could compare each person's stated responses with actual travel choices. Respondents sorted out at this stage were retired people, other individuals who do not work, individuals who work at home, and students. Also, some respondents who filled out the walk or bicycle experiments were dropped because they lived too far from work to consider walking (more than 3 miles) or bicycling (more than 7 miles) as practical choices. We retained 3185 surveys for model development; 1791 of them pertain to the four models reported in this paper. Between 273 and 679 surveys were used in the four sets of urban areas.

We checked the samples for representativeness by comparing the frequency distribution of selected socioeconomic characteristics of respondents with 1970 census data. The proportions of individuals in any one-way tabulation by sex, age, household size, and income (adjusted for inflation) were within ±10 percent of the census. The only exceptions were that, in some cities, the 15-24 age category, one-person households, and incomes under $5000 annually were underrepresented. Exclusion of students, retired, and other unemployed respondents explains the difference.

As a further check of representativeness, we compared the actual mode choices reported by respondents with the results of a strict probability sample conducted a year earlier by the Wisconsin Survey Research Laboratory (8). The comparison was satisfactory.

ANALYSIS OF SURVEY RESPONSES

The first stage in building the actual models was to fit linear additive models on the experimental responses obtained in the survey. The functional form and variables were already set in the design step so that model estimation is a simple task at this stage. The only flexibility in model estimation is in the socioeconomic variables and their functional form because they are not part of the experimental design. Multiple linear regression is used to estimate the models. The dependent variable is the response on the 1-5 scale, assuming that the stated likelihood of choosing a nonautomobile mode is proportional to utility. This is equivalent to using a linear approximation to a logit function. The independent variables are the experimental variables (level of service) and the background responses (socioeconomic characteristics).

The automobile-related variables appear in each survey form because automobile was the base mode against which each competing mode is set. Restrictions that the coefficients of the automobile variables be equal across all experiments are required for consistency in the multimodal model developed in the next step; the easiest way to apply these restrictions is to estimate a multiple linear regression across all the surveys jointly. The results of this are given in Table 1.

Formally, the equations in the table are as follows:

$$U_{ai} = -\sum_k c_{ak}X_{ak} + \sum_l c_{sl}X_{sl} + \sum_m c_{wm}X_{wm} + \sum_n c_{bn}X_{bn} + \sum_p c_{tp}X_{tp}$$

$$= -U_a + U_s + U_w + U_b + U_t \quad (1)$$

where

U_{ai} = utility of mode relative to driving alone (i.e., the response to a situation on the 1-5 scale from any experiment i); i = s (shared ride), w (walk), b (bicycle), or t (transit);
c = vector of coefficients;
X = vector of variables in experiment s, w, b, or t; variables for mode a appear in all experiments;
k = index that corresponds to drive-alone and socioeconomic variables;
l = index that corresponds to shared-ride variables;
m = index that corresponds to walk variables;
n = index that corresponds to bicycle variables; and
p = index that corresponds to local bus transit variables.

The utilities U_a, U_s, U_w, U_b, and U_t are the absolute utilities of each mode (not relative to drive alone), which are used in the calibration step. The Xs are dummy variables; for example, all X_{ls} = 0 except when i = s. Thus, Equation 1 encompasses each binary experiment but allows a multimodal treatment by incorporating the restriction that the automobile utility coefficients are the same in all binary comparisons.

Table 1 gives the results of analyzing the experimental responses for each city. Most of the coefficients show relatively little variation across cities, which suggests that transferability of these coefficients among urban areas is a possibility.

Table 1. Variables, coefficients, and goodness-of-fit statistics for regressions on experimental responses.

Variable Name	Definition	Madison (n = 305) Coefficient	t-Value	Milwaukee County (n = 273) Coefficient	t-Value	Fox River Valley Cities (n = 534) Coefficient	t-Value	Other Cities (n = 679) Coefficient	t-Value
Automobile Utility (U_a)									
CA	Automobile constant	-5.271		-4.697		-4.448		-5.051	
GA	Gasoline availability, 0 if ample supply, 1 if rationing	-0.320	-6.30	-0.377	-6.57	-0.318	-7.93	-0.315	-8.99
GP	Gasoline price ($/gal)	-0.234	-5.48	-0.320	-6.62	-0.284	-8.41	-0.284	-9.59
PK	Parking costs ($/month)	-0.016	-6.93	-0.017	-6.91	-0.017	-8.77	-0.016	-9.82
WT	Wait time to buy gasoline (min)	-0.008	-0.89	-0.004	-0.38	-0.013	-2.30	-0.007	-1.29
IN	Annual household income ($000s in 1980)	+0.012	6.02	+0.010	3.73	+0.001	0.59	+0.008	5.09
VP	Vehicles per person 16 years old and over in household	+0.178	3.12	+0.078	1.19	+0.096	2.48	+0.004	0.13
TT	Travel time (min)	-0.030	-2.77	-0.025	-2.27	-0.019	-1.89	-0.33	-3.70
Shared-Ride Utility (U_s)									
CR	Shared-ride constant	0.216	3.08	-0.090	-1.21	0.360	5.91	0.085	1.61
RD	Ridesharing partner, 0 if general public matching, 1 if coworker or neighbor	+0.222	2.58	+0.216	2.21	+0.138	2.00	+0.081	1.41
WS	Work schedule, 0 if flexitime, 1 if fixed 8-h day	+0.401	4.66	+0.384	3.94	+0.581	8.46	+0.399	6.93
TT	Travel time (min)	-0.030	-2.77	-0.025	-2.27	-0.019	-1.89	-0.033	-3.70
Walk Utility (U_w)									
CW	Walk constant	0.386	4.46	0.268	2.820	0.151	2.30	0.119	2.01
WD	Walk distance to work (miles)	-0.897	-3.36	-0.936	-3.08	-0.925	-5.48	-0.784	-5.03
SW	Sidewalks, 0 if all the way, 1 if part of the way	0	-[a]	0	-[a]	0	-[a]	-0.053	-0.68
SN	Season, 0 if summer, 1 if winter	-0.756	-5.66	-0.750	-4.93	-0.868	-10.29	-0.848	-10.83
Bicycle Utility (U_b)									
CB	Bicycle constant	-0.275	-3.81	-0.130	-1.610	-0.225	-3.56	-0.418	-7.49
BD	Bicycle distance to work (miles)	-0.245	-5.24	-0.213	-3.67	-0.254	-6.69	-0.276	-8.19
BL	Bicycle lane, 0 if marked lane in street, 1 if no lane	-0.356	-3.81	-0.216	-1.87	-0.330	-4.27	-0.296	-4.40
SS	Street surface, 0 if smooth, 1 if rough	-0.383	-4.11	-0.470	-4.05	-0.431	-5.57	-0.400	-5.93
TR	Traffic, 0 if quiet, 1 if busy	-0.517	-5.53	-0.500	-4.31	-0.417	-5.39	-0.378	-5.61
Bus Utility (U_t)									
BT	Bus transfer time (min)	-0.044	-2.00	-0.035	-1.58	-0.019	-0.96	0	-[a]
BF	Bus fare ($)	-0.221	-0.81	-0.443	-1.58	-0.240	-0.96	-0.195	-0.88
HW	Bus headway (min)	0	-[a]	0	-[a]	-0.006	-0.84	-0.007	-1.14
TT	Travel time (min)	-0.030	-2.77	-0.025	-2.27	-0.019	-1.89	-0.033	-3.70
R^2		0.151		0.116		0.139		0.131	
F		21.44		14.24		32.56		38.73	
Data points		2440		2184		4272		5432	

[a]Coefficient was set to zero because the t-value was less than 0.3 and the wrong sign occurred.

Gasoline availability, gasoline price, and parking cost all have a significant effect on mode choice. A wait in line of between 5 and 20 min to purchase gasoline is less significant but has a stronger impact in small cities, where currently it may be more convenient to purchase gasoline and where there has been no previous experience with long queues to buy gasoline. Income and vehicles per person are generally significant also. This use of socioeconomic variables as additive terms in the automobile utility was chosen for simplicity and consistency across urban areas. The use of different socioeconomic specifications could improve the model goodness-of-fit somewhat but at the price of added complexity.

The travel time coefficients for drive alone, shared ride, and transit were constrained to be equal for consistency. Work schedule and ridesharing partner were both significant variables in the ridesharing utility.

The walk utility is strongly dependent on distance and season, but sidewalk availability was not perceived as a major factor, except by some respondents in the small cities, which have less extensive sidewalk systems. Bicycle utility also depends strongly on distance, but it also depends on the presence of a bicycle lane, street surface, and traffic levels. (Season was not included in the bicycle-automobile experiment, but the season coefficient from the walk model is used in the bicycle utility function for policy analyses.)

The bus utility equation (Equation 7) contains surprising results over the ranges of variables tested, which show strong sensitivity to overall travel time but relatively little to headway (15- to 30-min range) and fare (40- to 80-cent range). Transfer times of 0-5 min had a modest affect. Respondents may have had difficulty in assessing individual time components for a bus trip and, therefore, used the total time variable to determine their choice.

The city-to-city variations in the constants are as anticipated. Madison shows the highest propensity to use non-drive-alone modes, and other cities have lower constants in those cases. The R^2 of the regressions ranges from 0.116 to 0.151, which is expected given the lack of market segmentation, the inclusion of invariant respondents who indicated all 1s or all 5s on the survey, and the simple socioeconomic descriptions used. The F-statistics are all significant.

Calibration

In the calibration step of the analysis, the models built from stated behavior in the experiment are tested against actual, current behavior as a check. We substitute levels of independent variables that represent current conditions into the experimentally derived utility functions to obtain values of $\bar{U}_a$, $\bar{U}_s$, $\bar{U}_w$, $\bar{U}_b$, and $\bar{U}_t$ for each respondent. These values are then substituted into a logit formulation to test how well they explain current choice:

$$p_i = \exp(a_i+b_i\bar{U}_i)/\sum_{\text{all } j} \exp(a_j+b_j\bar{U}_j) \quad (2)$$

where

- p_i = probability of a respondent choosing mode i (equal to 0 or 1 in actual data);
- $\bar{U}_i$ = a respondent's computed utility value for mode i under current conditions, calculated from regression results; and
- a_i, b_i = coefficients to be determined in logit estimation.

The equations below represent the regression results for Madison as five separate utility equations, as required for the validation. These separate equations sum to the original equation, with a negative sign for drive alone.

The linear utility equations for Madison from regressions on experimental responses are as follows: For automobile,

$$\begin{aligned} U_a = {} & -5.271 - 0.320GA - 0.234GP - 0.016PK - 0.008WT \\ & \quad (-6.30 \quad -5.48 \quad -6.93 \quad -0.089 \\ & + 0.012IN + 0.178VP - 0.030TT \\ & \quad 6.02 \quad 3.12 \quad -2.77) \end{aligned} \quad (3)$$

For shared ride,

$$\begin{aligned} U_s = {} & 0.216 + 0.222RD + 0.401WS - 0.030TT \\ & (3.08 \quad 2.58 \quad 4.66 \quad -2.77) \end{aligned} \quad (4)$$

For walk,

$$\begin{aligned} U_w = {} & 0.386 - 0.897WD - 0.756SN \\ & (4.46 \quad -3.36 \quad -5.66) \end{aligned} \quad (5)$$

For bicycle,

$$\begin{aligned} U_b = {} & -0.275 - 0.245BD - 0.356BL - 0.383SS - 0.517TR \\ & (-3.81 \quad -5.24 \quad -3.81 \quad -4.11 \quad -5.53) \end{aligned} \quad (6)$$

For local bus transit,

$$\begin{aligned} U_t = {} & -0.044BT - 0.221BF - 0.030TT \\ & (-2.00 \quad -0.81 \quad -2.77) \end{aligned} \quad (7)$$

where

- U_a = automobile utility,
- GA = gasoline availability,
- GP = gasoline price ($/gal),
- PK = parking costs ($/month),
- WT = wait time to buy gasoline (min),
- IN = annual household income ($000s in 1980),
- VP = vehicles per person $\geq$ 16 years old in household,
- TT = travel time (min),
- U_s = shared-ride utility,
- RD = ridesharing partner,
- WS = work schedule,
- U_w = walk utility,
- WD = walk distance to work (miles),
- SN = season,
- U_b = bicycle utility,
- BD = bicycle distance to work (miles),
- BL = bicycle lane,
- SS = street surface,
- TR = traffic,
- U_t = bus utility,
- BT = bus transfer time (min), and
- BF = bus fare ($).

In order to gain some understanding of the values of a_i and b_i that indicated satisfactory correspondence between the experimental model and actual behavior, a simple analysis was performed. We know immediately, of course, that we wish all $b_i > 0$ and all a_i to be small in some sense. Figure 2 shows the hypothesized relation in a binary case between linear regression results and the binary logit equation. If stated behavior (linear model) corresponds to actual behavior (logit model), then we expect the linear utility equations to perform well in the logit model. A linear approximation tangent to the logit function at p = 0.5 (as drawn) has a slope of 0.25 and thus intersects the p = 0 and p = 1 axis at U = -2 and U = +2, respectively. This scale, from -2 to +2, is our 1-5 response scale shifted downward three units. We can expect b_j to

Figure 2. Comparison of linear and logit model forms.

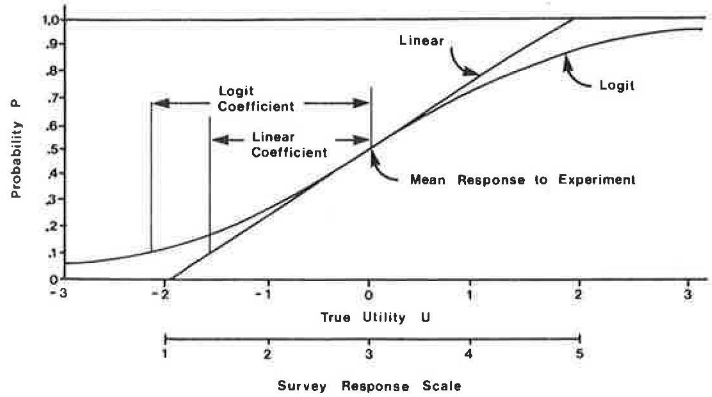


Table 2. Multinomial logit calibration results.

	Madison (n = 312)				Milwaukee County (n = 282)				Fox River Valley Cities (n = 661)				Other Cities (n = 873)			
	a		b		a		b		a		b		a		b	
Mode	Coefficient	t-Value	Coefficient	t-Value	Coefficient	t-Value	Coefficient	t-Value	Coefficient	t-Value	Coefficient	t-Value	Coefficient	t-Value	Coefficient	t-Value
Drive alone	+13.221	3.13	+2.558	2.57	+5.166	1.17	+2.716	1.54	+12.437	1.94	+2.496	1.50	+7.942	2.21	+1.398	0.93
Rideshare	-1.510	-1.47	-0.404	-0.58	-17.173	-0.09	+3.419	0.06	-1.332	-1.02	+0.228	1.05	-1.308	-1.49	+1.600	0.62
Walk	+0.813	1.49	+2.39	3.29	+2.812	3.88	+2.758	2.78	+1.646	1.47	+2.211	4.15	+2.000	2.66	+3.108	6.21
Bicycle	-0.390	-0.60	+0.740	0.90	+1.525	1.10	+2.119	1.12	+0.347	0.30	+1.602	1.66	+1.403	1.64	+2.090	2.80
Bus	0.0	-[a]	+1.331	0.98	0.0	-[a]	+0.575	1.06	0.0	-[a]	+4.550	1.99	0.0	-[a]	+1.665	1.12

Note: The b coefficients are tested against the null hypothesis that b = 1, and the a coefficients are tested against the null hypothesis that a = 0, except for drive alone, where the null hypothesis is a = 3. The -2* log-likelihood ratio was 319.07 for Madison, 507.71 for Milwaukee County, 1032.53 for Fox River Valley cities, and 1246.10 for other cities.

[a]Coefficient was set to zero because the t-value was less than 0.3 and the wrong sign occurred.

approximately equal 1 and a_j to equal 0. The use of p = 0.5 as the point at which the approximation is made is justified by the experimental design, which can create sets of situations in which the alternatives are well matched.

In the multinomial case, the approximation will necessarily be centered at $p < 0.5$ for most modes; this implies that $b_j > 1$ because the lower slope of the logit curve at $p \neq 0.5$ produces a linear scale longer than four units between the p = 0 and p = 1 axes. We still expect all a_j to be 0 if there are no systematic biases across experiments, with one exception. (The a_j for automobile is expected to be +3 because automobile's position on the survey response scale is the reverse of the other modes.) One a_j must be set arbitrarily, so we set the bus a_j equal to zero; thus, the bicycle, walk, and shared ride a_j are also expected to be zero.

These arguments are intended only to give an approximate sense of the values of a_j and b_j to expect from the logit-estimation step. Furthermore, this calibration is approximate for the same reasons that limit our ability to estimate a revealed preference model for the study--lack of variability in several major variables, unavailability or low use of alternatives, multicollinearity, and other problems. Even so, it is important to attempt to calibrate the models to test their accuracy. Because we are estimating only two coefficients per mode in the validation (a_j and b_j), we may succeed in establishing them when trying to estimate all coefficients would fail.

Most data required for calibration were self-reported, although a few items were gathered from transportation planning data bases. Self-reported data were checked against planning data where possible, but the comparison was inconclusive because of the aggregation errors in the planning data (e.g., multiple bus lines in a zone, varying parking charges).

The calibration results appear in Table 2. We describe the calibration results for Madison in detail and briefly compare them with those of the other areas. (The number of respondents is higher than in the regression step because responses with incomplete experimental data could be used in this step.) The results show a very strong relation between the experimentally derived utilities and actual behavior, so we turn to an examination of the adjustment coefficients a_j and b_j. The coefficients a_j are tested against a null hypothesis of zero (+3 for drive alone), and b_j is tested against a null hypothesis of one.

The Madison drive-alone utility derived from the experiment apparently understates the sensitivity of actual behavior to the variable set, because the coefficient b_j is 2.558 and is significantly different from 1 at the 95 percent level of confidence. The adjustment in the constant a_j is not as large as it appears: The new constant is 13.221 + 2.558 (-5.271) or -0.262, as compared with the original value of -5.271. However, an adjustment of +3 was expected a priori. The adjustment in a is also statistically significant at a 95 percent level of confidence.

The ridesharing calibration is inconclusive. The only variables in its utility equation are the ridesharing partner (invariant in the sample, all being coworkers or neighbors), the work schedule (taking only two values), and the time (a fixed difference from automobile). Thus, there is little variability on which to relate the utility values to actual behavior. This mode is an extreme example of the difficulties in validating models. Neither a_j nor b_j is statistically different from the null hypothesis, which we fail to reject.

The walk mode has a coefficient b_j that is significantly different from 1; the calibrated constant is 0.813 + 2.39 (0.386) or 1.735. The bicycle mode's coefficients are quite close to their presupposed values, and the adjustments are not significant. The same holds for the bus mode.

In general, the coefficients for the other models follow the same pattern as the Madison coefficients. The calibration coefficients are larger than we would ideally like to see, but they indicate a relatively good correspondence between the experimental models and actual behavior. Coefficients that are different from the a priori values may also occur for a variety of reasons not related to the correspondence between stated and actual behavior--errors introduced by the linear approximation, errors in self-reported data, aggregation errors in planning data (believed to be significant in this case), and the simplicity of the socioeconomic description.

An examination of the results for the coefficient b_j suggests that we should use the calibration coefficients to revise all the walk utility functions, the Madison drive-alone utility, and the other cities' bicycle utility.

Table 3. Final models.

Variable	Milwaukee County	Madison	Fox River Valley Cities	Other Cities
Automobile utility				
Gasoline availability	-0.377	-0.320[a]	-0.318	-0.315
Gasoline price	-0.320	-0.234	-0.284	-0.284
Parking cost	-0.017	-0.016	-0.017	-0.016
Wait time to buy gasoline	-0.004	-0.008	-0.013	-0.007
Annual household income	0.010	0.012	0.001	0.008
Vehicle per person ≥ 16 years old in household	0.078	0.178	0.096	0.004
Travel time	-0.025	-0.030	-0.019	-0.033
Shared-ride utility				
Ridesharing partner	0.222	0.216	0.138	0.081
Work schedule	0.401	0.384	0.581	0.399
Travel time	-0.030	-0.025	-0.019	-0.033
Walk utility				
Walk distance to work	-2.581[b]	-2.144[b]	-2.045[b]	-2.437[b]
Sidewalks	0.0	0.0	0.0	-0.165
Season	-2.069	-1.807	-1.919	-2.636
Bicycle utility				
Bicycle distance to work	-0.213	-0.245	-0.259	-0.577[b]
Bicycle lane	-0.216	-0.356	-0.330	-0.619
Street surface	-0.470	-0.383	-0.431	-0.836
Traffic	-0.500	-0.517	-0.417	-0.790
Season	-2.069	-1.807	-1.919	-2.636
Bus utility				
Bus transfer time	-0.035	-0.044	-0.019	0.0
Bus fare	-0.443	-0.221	-0.240	-0.195
Bus headway	0.0	0.0	-0.006	-0.007
Travel time	-0.025	-0.030	-0.019	-0.033

[a]Indicates b_j different from one, but original coefficients used based on sensitivity analysis.
[b]Indicates group of coefficients multiplied by b_j significantly different from one.

Sensitivity Analysis

Before selecting the final model coefficients, we used the incremental form of the logit model to perform sensitivity analysis:

$$p_i^1 = p_i \exp(\Delta U_i^*) / \sum_{\text{all } j} p_j \exp(\Delta U_j^*) \qquad (8)$$

where

p_i^1 = revised share of mode i;
p_i = base share of mode i;
U_i^* = validated utility of mode $i = a_i + b_i U_i$, where a_i is significantly different from 0, and b_i is significantly different from 1; and
ΔU_i^* = change in the validated utility of mode j due to a change in a variable from the base case, ΔX.

The sensitivity analysis indicated that most of the validated models provided reasonable results. However, if predictions are made with the validated Madison drive-alone utility function, we find that a $0.60 increase in gasoline price from $1.30/gal causes the mode share for driving alone to decline from 56 to 45 percent, a reduction equal to 9 percent of all work trips. These results are outside the range expected on the basis of gasoline price elasticities reported in the literature. When we used the calibrated automobile utility function to predict the effect of changes in fuel availability and parking costs, we also obtained changes in market shares too large to be believable. Because of the possible confounding factors that could have produced a coefficient b_j different from one, we chose to retain the original experimental utility equation for Madison drive alone.

FINAL MODELS

The final models appear in Table 3. Only the walk models and the other cities bicycle model have been adjusted through the calibration step, as described above; the other models are in their original form based on the experiment. Only the Madison drive-alone model had a significant b_j but was not changed due to sensitivity results. All other models have also been tested in sensitivity analysis and produce reasonable results. Adjusted constants are not shown, as they are dependent on the level of aggregation used; a simple procedure is used to find base values of the constants when the models are applied for forecasting.

Table 4 gives the elasticities and values of time that emerge from the final models. The values generally agree with the previous literature, although the range of variation is outside that of past data and creates some differences.

The results of this effort highlight some key

Table 4. Selected elasticities and values of time.

Urban Area	Direct Elasticities: Gasoline Price	Parking Cost	Bus Fare	Bus Travel Time	Cross Elasticities[a]: Gasoline Price	Parking Cost	Marginal Value of Time[b] ($/h)
Milwaukee County	-0.166	-0.059	-0.247	-0.349	+0.448	+0.186	4.64
Madison	-0.196	-0.106	-0.117	-0.396	+0.249	+0.134	7.69
Fox River Valley	-0.152	-0.071	-0.141	-0.279	+0.387	+0.183	4.01
Other cities	-0.183	-0.082	-0.189	-0.480	+0.356	+0.158	7.15

Note: All elasticities are point elasticities and were calculated at the mean value of the independent variables in the experimental data sets: gasoline price = $1.90/gal, parking cost = $15/month, bus fare = $0.60, travel time = 15 min.

[a]Logit models have constant cross elasticities (i.e., for a 1 percent change in gasoline price, for example; all other modes have the same change in demand).
[b]Marginal values of time calculated by using the travel time and the gasoline price coefficients.

issues in integrating functional measurement and disaggregate models. When using functional measurement to address issues not well captured in data on actual behavior, testing of the correspondence between stated and actual behavior is difficult. The standard validation approach of simple prediction of mode shares with the functional measurement model and comparison to aggregate actual shares is sensitive to the values of the independent variables assumed (and about which there is some latitude) and generally does not yield statistical measures of the closeness of correspondence (9). This study attempted to assess whether functional measurement models could be used in a logit framework without adjustment and whether sufficient variability existed to check the performance of the model. The results are encouraging, although more work is clearly needed.

MODEL APPLICATIONS

These models are currently in use in several functions at WisDOT. First, they are being used in their incremental logit form for statewide policy-level analysis of key issues that face the department. By comparing the impacts of state policies in a consistent fashion across Wisconsin urban areas, the department can target its programs where their effect is largest. A policy report has been prepared based on the models (10) and concludes, for instance, that transit assistance should be targeted at larger urban areas where its effect is significant, that ridesharing should be promoted in all areas, with an emphasis on employer and neighborhood matching programs versus less-effective general public matching programs, and that bicycle lanes may be cost-effective investments for diverting travelers from driving alone, even though their impact is only seasonal. In many cases bicycle lanes have greater impacts than transit improvements and lower cost. In Madison, for example, if bicycle lanes were marked on the streets in a corridor where the percentage of people that use each mode to work equaled each mode's share for the city as a whole, drive alone's share of the work trips would decrease by almost 3 percent. In contrast, a 5-min reduction in bus transfer time would divert less than 2 percent of the total trips from drive alone and a 10-min reduction in bus travel time would decrease drive alone's share by only 1 percent. The direction for transit improvements, when considered alone, will involve decreases in travel time and fare increases, as service level generally appears more important than fare to the public over the ranges examined.

Some of the more interesting conclusions and policy implications of the study include the following. Approximately 112 000 of 1.5 million one-way daily home-bound work trips would switch from driving alone to other modes if gasoline were rationed (10 gal/registered vehicle each week). A wait of 30 min to buy gasoline at a service station would cause 70 000 of the 1.5 million daily drive-alone trips to shift to other modes.

The models reported here indicate that a general public carpool matching program is not as effective as an employee or neighborhood-based ridesharing programs in Wisconsin cities larger than 50 000 people. However, a similar set of models for long-distance commuter travel between Madison and its satellite communities indicate that residents of villages and small communities in rural areas are nearly as willing to share rides with strangers as with neighbors or coworkers. Fear of strangers seems to be more prevalent in larger cities than in small rural communities, as expected. Thus, a general public carpool matching program might work well for commuters who live in small communities outside Wisconsin's larger cities. Universal flexitime for workers in the urban areas studied would cause 58 000 fewer home-based work trips by ridesharing to occur daily than if everyone worked fixed 8-h shifts.

The addition of marked bicycle lanes to all streets throughout each of the cities studied would encourage an additional 26 000 bicycle work trips in good weather months, a 39 percent increase in total summertime bicycle trips. Bicycle lanes would impact strongly on bicycle ridership in the medium and smaller cities of Wisconsin but would have little effect in the state's largest city, Milwaukee. The allowing of pavements throughout 10 cities to deteriorate from smooth to rough riding surfaces would cause a reduction of 38 000 bicycle work trips on nice days--a 42 percent reduction in total bicycling in the summertime. Thus, local street maintenance practices should pay particular attention to keep pavements on popular bicycle routes in good condition to avoid loss in bicycle ridership.

The models are also being made available to urban areas for use in their planning process. They can be implemented in the urban transportation planning system (UTPS) as part of WisDOT's technical assistance role to local areas. These models will lead to more detailed, yet consistent, evaluations of policies already assessed at a statewide level by incremental logit.

Finally, the models have a quick-response capability through the use of incremental logit and are available to respond to requests by planning and other agencies for quick analyses of proposed services and policies. A major staff capability exists at WisDOT to use these models in this manner.

ACKNOWLEDGMENT

The analysis, results, conclusions, and recommendations are solely ours and do not necessarily represent the views or policies of WisDOT.

REFERENCES

1. P. Green and V. Srinavasan. Conjoint Analysis in Consumer Research: Issues and Outlook. Journal of Consumer Research, Vol. 5, No. 1, Sept. 1978, pp. 103-123.
2. R.J. Meyer, I.P. Levin, and J.J. Louviere. Functional Analysis of Mode Choice. TRB, Transportation Research Record 673, 1978, pp. 1-7.
3. Passenger Travel Demand Model and Analysis. Wisconsin Department of Transportation, Madison, State Highway Plan, Special Rept., 1980.
4. G. Kocur. An Intercity Travel Demand Model System Estimated from Data on Behavioral Intentions. In Proc., Conference on New, Practical Methods of Transportation Demand Analysis, Institute of Urban and Regional Research, Univ. of Iowa, Iowa City, 1981.
5. Development of Wisconsin Urban Work Trip Models for Forecasting Modal Choice. Wisconsin Department of Transportation, Madison, Tech. Rept., 1981, pp. 3-13 and 3-14.
6. W.G. Cochran and G.M. Cox. Experimental Design, 2nd ed. Wiley, New York, 1957.
7. G.J. Hahn and S.S. Shapiro. A Catalog and Computer Program for the Design and Analysis of Orthogonal Symmetric and Asymmetric Fractional Factorial Experiments. General Electric, Schenectady, NY, Tech. Information Series, 1966.
8. Transportation Issues and Answers: A Survey of Public Opinion in Wisconsin. Wisconsin Department of Transportation, Madison, Summary Rept., 1979.

9. J.J. Louviere and others. Laboratory-Simulation Versus Revealed-Preference Methods for Estimating Travel Demand Models. TRB, Transportation Research Record 794, 1981, pp. 42-51.
10. Transportation Issues and Answers: Choice of Travel to Work. Wisconsin Department of Transportation, Madison, Summary Rept., 1981.

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

Elasticity-Based Method for Forecasting Travel on Current Urban Transportation Alternatives

DANIEL BRAND AND JOY L. BENHAM

This paper presents a quick-response incremental travel demand forecasting method that uses travel demand elasticities and readily available ground count travel and land use data. Elasticities are defined and criteria for selecting elasticities are identified. The steps for calculating each component of travel affected by a transportation improvement are described. Personnel and computational requirements for this method are greatly reduced relative to those necessary for forecasting with the conventional four-step sequential process (trip generation, distribution, modal split, and trip assignment). The basic travel behavior assumptions of the method are similar to those inherent in conventional models although, in contrast to sequential derivation and application of these models, internally consistent causal relations are maintained. A range of outputs of interest to policymakers is generated, including changes in total travel, changes in mode-specific travel, and changes in travel on a given route or link. The elasticity-based method has recently been used to forecast patronage on the four major transit alternatives included in the Baltimore North Corridor alternatives analysis. This application is described in the paper and compared with forecasts made in a particular application of the conventional four-step sequential travel demand forecasting system for the same alternatives under the same conditions. This direct comparison of the two forecasting methods provides a unique opportunity to assess the effects on forecast patronage of many assumptions inherent in typical applications of each method.

Much of the concern over urban travel demand forecasting involves the turnaround time and expense of applying existing conventional sequential travel demand models. Also, application of these conventional models often involves a series of restrictive assumptions that can reduce severely their ability to distinguish travel impacts between alternatives (1). These models synthesize travel patterns from scratch based on a long list of land use, socioeconomic, and level-of-service variables, which themselves must be forecast (thus propagating errors) (2). One way to cut significantly the large costs currently associated with urban travel forecasting is to use elasticities with respect to those limited numbers of variables related to the policy option of interest. Also, since elasticities can be behavioral, the spatial extent of the forecasts can be limited to those areas of the region affected by the system change being tested. The most easily available travel data, namely ground count data, can be factored incrementally at some useful and informative level of aggregation. Such an approach saves the time, expense, and uncertainty involved in forecasting and calculating entire sets of independent variables.

The elasticity-based approach described here has recently been used to forecast patronage on four major transit alternatives considered in the Baltimore North Corridor alternatives analysis. In addition to the elasticity-based forecasts, patronage estimates were developed by the Baltimore Regional Planning Council by using the existing four-step, sequential forecasting system estimated with urban transportation planning system (UTPS) software. Hence, the opportunity to compare and evaluate the two methods was provided.

ELASTICITIES

A travel demand elasticity is defined as the percentage change in ridership or traffic volume (depending on what is measured) that results from a 1 percent change in a given independent variable (e.g., travel time or cost) (3). Elasticities are measures of the partial effect on travel of changes, taken singly, in the travel environment that confront travelers. They allow shifts in travel patterns to be estimated at the margin in response to changes in the travel environment and, therefore, existing observed travel unaffected by changes is preserved. Existing synthetic (UTPS) procedures can only duplicate existing travel with some difficulty.

Elasticity-Based Forecasting Method

The elasticity-based forecasting procedure is based on the concept that travel on a new or improved transit facility is composed of four components, each of which results from one mutually exclusive cause or behavior and each of which can be calculated separately and sequentially to include the results of the previous change. The four components are as follows:

1. Transit travel that does not exist today due to growth in numbers of people and jobs; these are changes in travel due to so-called long-run demand, or land use changes;
2. Transit travel that is diverted from (or to) the automobile mode due to changes in automobile-operating costs (e.g., increases in gasoline price) and other automobile level-of-service changes (e.g., reductions in travel time due to highway construction);
3. Transit travel diverted to the improved transit facility from transit facilities for which the new or improved transit facility is a superior substitute; this is diverted travel from facilities of the same mode; and
4. Induced transit travel, or travel that is induced in the corridor and specifically on the transit alternative being evaluated as a result of the new or improved transit facility; induced transit travel includes travel that results from increased rates of choice of destinations served by the improved facility and increased transit trip

frequency (including automobile trips diverted to the transit improvement).

To calculate the first travel component due to land use changes, the origin-destination (O-D) superzone transit trip table is factored to account for growths (or declines) in population and employment. Simple proportional factors on numbers of households and jobs are used to account for growths and declines in transit travel. This assumes a long-run equilibrium between the preferred residence and employment and other activity locations of people, and the travel choices available to them. To account for the fact that the population mix, for example, in a residential neighborhood will change to reflect the (long-run) behavior of people to locate in accordance with their transportation preferences, long-run elasticities must, for the sake of consistency, be used to calculate the second and fourth components of travel.

Because transit demand is a function of both automobile and transit level of service, the second component of travel includes only the change in transit use that results from changes in automobile level of service. Transit trips are factored by using cross-elasticities of transit demand with respect to automobile level-of-service characteristics. No assumptions need to be made that transit is directly substituted for all automobile trips foregone as a result of gasoline price increases, for example, even for work trips. The transit demand cross-elasticities, empirically derived, provide the proportion of automobile trips foregone that use transit in the given situation. In particular, the cross-elasticities provide the percentage of change in transit use that results from each 1-percent change in each automobile level-of-service characteristic.

To calculate the third travel component, diverted transit travel, the amount of transit travel between each superzonal pair on each affected transit route that is diverted to the transit improvement or alternative is calculated. This calculation is based on level-of-service differences between the existing routes that serve the O-D pair and the alternative being tested. The resulting diverted transit trip table will already have been factored appropriately to account for growth in transit travel due to land-use changes and travel from automobile due to changes in the automobile system (travel components 1 and 2).

The alternative-specific superzonal transit trip table (from component 3) is factored by using direct transit elasticities applied to the transit level-of-service differences between the new alternative and the existing bus routes from which travel is diverted to calculate induced travel from the transit improvement.

Calculation of the first two components of travel results in the forecast year transit trip table that reflects the future year population and highway level of service on the base year transit network. Hence, the stage is set for introducing the transit alternatives. With the introduction of new or improved transit lines, existing transit trips will be diverted to the new routes (component 3). This diverted travel represents the base transit ridership on the new routes, which is then factored to reflect the increase in travel (component 4) induced as a result of the improvement in level of service.

Assumptions

The approach outlined above is based on certain behavioral assumptions that should be made explicit. Certain basic assumptions are no different from the assumptions inherent in the conventional sequential series of steps in urban travel forecasting. However, calculation of the four components of travel uses internally consistent relations that account explicitly and appropriately in each step for changes in trip frequency (trip generation), destination choice (trip distribution), modal choice (modal split), and trip diversion (assignment). Changes are calculated in all of these travel choices for every change in the transit and highway system. The lack of feedback to these choices in the usual UTPS process is avoided. Double counting of changes in travel choices is also avoided. That is, in the traditional sequential four-step modeling process, changes in travel behavior in more than one travel choice are contained in the data used to model or explain a single travel choice. When single travel choices are forecast sequentially by using models derived in this manner, the effect is to count changes in these choices several times and thereby inflate the impact of these changes on travel behavior.

The following mapping of the conventional sequence of travel choices on the explicitly and uniquely calculated travel components is helpful.

Long- or Short-Run Travel Choice	Travel Component
Population and employment growth, decline, or redistribution	1
Transit trip frequency, destination choice, and modal choice due to changes in automobile level of service	2
Transit trip frequency, destination choice, and modal choice due to change in transit level of service	4
Transit path choice	3

The time of day travel choice is omitted here for ease of presentation. It is addressed in the Baltimore study through the development of alternative-specific peaking factors that reflect how this travel choice varies with the transportation improvement. Because it goes back to land use changes (component 1), the method assumes a long-run equilibrium between the preferred residence and employment (and other) locations of people and the travel choices available to them. This generally requires that elasticities should be used that have been derived from models estimated by using only a certain kind of data, namely cross-sectional origin-destination data (1). This is not a constraint because most models are estimated by using such data collected at one point in time. The distinction between short- and long-run elasticities is important because it has been found that elasticity estimates based on models calibrated with cross-sectional data are consistently larger than short-run elasticities based on before and after studies (4).

Criteria for Selecting Elasticities

Although the application of elasticities is a relatively simple procedure, the elasticities selected for use in the forecasting approach described above must be consistent with the travel demand changes being measured. For example, models based on cross-section data are often estimated for a given trip purpose and involve a single travel decision such as modal choice. Other models, known as direct-demand or simultaneous-choice models, include a range of travel decisions--trip frequency, modal choice, and destination choice. Note that this set of travel decisions is the behavior being modeled in step 4 above. Still other models calibrated with before and after data typically measure the difference in aggregate demand on a facility or system from a given improvement in level of service. However, if the before and after data are only for a specific facility, they include travel diverted to the facility from competing facilities as well as demand induced as a result of the facility improvement. Elasticities estimated by using such data do not distinguish between diverted and induced travel. The elasticity-based forecasting procedure described in this paper calls for separate calculation of induced and diverted travel components. That is, the transit trips factored to reflect induced travel consist only of trips already diverted to, or confronting, the benefits of the proposed transit improvement. Therefore, facility-specific data from before and after studies are generally inappropriate as sources of elasticities. In addition, elasticities that result from many before and after studies fail to fully account for the effects of changes in level-of-service that are exogenous to the improvement but that influence demand. In general, elasticities derived from models are preferred to before and after studies because they control for more factors that affect travel demand.

Elasticities are transferable contingent on certain conditions. Therefore, the elasticities to be used in patronage forecasting should be selected with several criteria in mind.

1. Elasticities should be derived from travel models that are consistent with travel behavior theory so that the elasticities will be behavioral.
2. Long-run elasticities should be used when future year travel forecasts are required. As discussed above, long-run elasticities can be estimated from cross-sectional (or some time-series) models that include (control for) a large set of relevant variables. Direct demand models (3,5) are preferred, especially for deriving nonwork trip elasticities because they measure at one time the impact of changes in all travel choices on ridership.
3. Elasticities should reflect the travel patterns of the study population to the extent possible by developing composite elasticities estimated for specific trip types or transit users. For example, the observed trip purpose distribution can be used to combine work and nonwork trip elasticities to develop the appropriate peak-period or all-day elasticity for the study area.
4. Socioeconomic characteristics of the population and the base level of service can have an effect on the value of elasticities. Therefore, elasticities appropriate for the study population and level of service should be used.

ELASTICITY-BASED PATRONAGE FORECASTS FOR BALTIMORE NORTH CORRIDOR

The Baltimore North Corridor alternatives analysis considered four basic transit alternatives: light rail, commuter rail, busway, and express bus. The light rail transit alternative consists of a new two-track rail transit system that would extend about 16.5 miles from the northern point of the corridor (Hunt Valley) through MetroCenter. The commuter rail alternative involves a shorter alignment that begins at Timonium (3.5 miles south of Hunt Valley) and ends near the northern border of MetroCenter. A timed transfer shuttle bus service provides collection and distribution service in MetroCenter. The busway consists of an exclusive right-of-way for buses used by two types of routes. A spine service is provided that originates at Hunt Valley, stops at intermediate on-line stations, and circulates on local streets in MetroCenter. Express buses, which provide park-and-ride and collection and distribution service in the North Corridor and circulate in MetroCenter, also use the busway. The express bus alternative consists of a network of park-and-ride lots and express bus services by using the existing roadway system in the North Corridor and circulating in MetroCenter.

Data Preparation

The elasticity-based method is predicated on the ability to identify and work (manually) with a relatively small number of existing routes and links from which travelers might be diverted to the new and improved facility. This is not usually possible when analyzing a new, high-speed expressway that profoundly affects travel on a large number of links in multiple corridors of a region. For the expressway example, detailed computerized conventional network analysis seems inescapable. However, such projects, which have such far-reaching facility interactions, are no longer the focus of most planning exercises. The Baltimore North Corridor transit alternatives are typical of current major transportation improvement proposals in even the largest urban areas. These consist of express transit lines whose travel impacts affect relatively few (albeit large) transportation links in one corridor.

Data preparation for the Baltimore North Corridor alternatives analysis included identification of the bus routes and links currently used that might be diverted to the new and improved facility. Volumes on these links are obtained from observed bus counts and represent the relevant travel universe that might be affected by the proposed alternatives. For purposes of growth factoring (step 1), it is necessary to define the area served by the affected transit links and to delineate analysis zones within the service area. The maximum service area was defined by examining the existing and proposed transit alternatives, their access characteristics, and relevant existing travel data such as data on distance between travelers' origins and transit lines and level of transferring. Because elasticities are applied incrementally, only travel affected by the alternatives needs to be considered. Therefore, data requirements are small relative to forecasting methods that simulate all travel in a region.

Trips on the affected transit links are then assigned to the origin and destination superzones served by those links. This assignment is done on the usual basis of shortest path (i.e., which bus routes serve which superzones), and information on average trip length or from on-board transit surveys if available. UTPS-selected link output, if available from an earlier study, is of course very helpful in this regard for obtaining the existing O-D distribution of observed trips on any transit link.

The actual travel diverted to each alternative is calculated in step 3 by using a proportional assignment procedure. The assignment procedure is based on the concept that the route choice travel decision can be represented as a function of the relative utilities or impedances on the alternate routes. The utilities are a function of the various service attributes, weighted by traveler's preferences for these attributes. Hence, the proportion of trips between two points attracted to each route is proportional to the relative impedances of the routes that connect these points, such that

$$P_i = (1/I_i)/\sum_j (1/I_j) \quad (1)$$

where

P_i = proportion of trips attracted to route i,
I_i = impedance of route i, and
i,j = route alternatives.

The impedance term includes level-of-service attributes such as in-vehicle time, walk time, wait time, and fare. The weights for attributes are derived from travel model coefficients estimated for populations comparable with the study corridor. The product of this step for each alternative (in the Baltimore application) is diverted travel by access mode to each station or express bus route segment by O-D pair and previous transit path. This allows the exact calculation of changes in most level-of-service characteristics faced by transit users to calculate diverted and induced travel. That is, the use of zonal average travel times or waiting times for multiple routes is avoided.

Calculation of the fourth travel component, induced travel, involves two steps. First, the percentage change in level of service faced by existing submodal travel markets is used to calculate increases in transit trips by these markets induced as a result of the improvement. Transit demand elasticities are applied to the service improvement obtained by users who travel between two zones for each base (previous) transit path and access mode. The separating of submodel travel markets avoids the need to aggregate access level-of-service over submodes (e.g., by taking weighted averages).

Aggregation introduces paradoxes and illogical change measures. For example, the bus paradox occurs when improved feeder bus to a trunk transit mode is provided in an improved alternative as a service improvement over park-and-ride and kiss-and-ride. Simple computation of a weighted average in-vehicle access time actually increases travel time with the service improvement since a higher percentage of transit users use the slower feeder bus relative to automobile access. This lowers overall demand for that route, despite the transit service improvement. Hence, the paradox, which is avoided by analyzing the behavior response of existing submodal travel markets separately.

The second step in the calculation of the fourth travel component is calculation of induced travel for new submodal travel markets. For example, with the provision of a park-and-ride station, travel by a new submodal market--automobile access--may be expected. In this case, if diverted and induced trips by walk and feeder bus at the new station are estimated to total 200 and the equilibrium submodal split at the station is 50 percent walk and feeder bus and 50 percent automobile, the station will attract 200 additional trips by automobile access for a total of 400 trips. Future equilibrium station assignment and access mode split depend on riders' origin distance from stations, available feeder bus, roads that connect origin zones and stations, parking availability, household income and automobile availability, and characteristics of the travelers' destination (e.g., parking availability). Access mode split is also heavily affected by the fact that transit travel between suburban areas where automobile level-of-service is good is dominated by transit captives, although travel to the downtown attracts choice riders as well as captives. Therefore, the origin-zone-specific access mode splits for travel to suburban destinations were significantly different from those assumed for travel to MetroCenter. The product of this step is total peak-period travel on the alternative. Note that because induced travel and diverted travel are calculated separately, a direct output of the method is the number of new transit trips associated with the transportation improvement.

Elasticities Selected for Baltimore

For the Baltimore alternatives analysis, elasticities derived from cross-section models (3,5-8) were used to develop constant peak-period transit elasticities. Elasticities with respect to the following transit level-of-service variables were developed: fare, in-vehicle time, out-of-vehicle time, and frequency. The frequency elasticity was used to measure the impact of changes in trip frequency where headways were greater than 10 min. The waiting time (out-of-vehicle time) elasticity alone is inadequate to measure the full effect on patronage of headways greater than 10 min because the conventional definition of wait time as one-half the headway up to a maximum of 5 min was used. Cross-elasticities with respect to the following automobile level-of-service variables were also developed: automobile operating cost and automobile in-vehicle time. The selected values for these elasticities are given in the table below.

Elasticity	Selected Value of Elasticity
Direct	
Transit fare	-0.15
Transit in-vehicle time	-0.37
Transit out-of-vehicle time	-0.65
Transit frequency	+0.26
Cross	
Automobile operating cost	+0.18
Automobile in-vehicle time	+0.20

Results

Total Baltimore North Corridor and MetroCenter peak-period (7:00-9:00 a.m.) transit trips are summarized for each alternative in Table 1. This table gives boardings on each alternative as well as all transit destinations in the North Corridor or origins in MetroCenter. The comparison of total transit trips reveals that the highest level of transit tripmaking occurs with the rail transit and busway alternatives, followed by express bus and commuter rail. The differences in the number of all corridor transit trips between alternative and base (1978 transit network) trips are new trips induced on each alternative.

COMPARISON OF ELASTICITY-BASED FORECASTS WITH UTPS FORECASTS FOR BALTIMORE NORTH CORRIDOR

Patronage forecasts for the alternatives were developed by using both the elasticity-based method and

Table 1. Daily morning peak period Baltimore North Corridor and MetroCenter transit trips by alternative, 1995.

Mode	Total Boardings	All Corridor and MetroCenter Transit Trips[a]	Difference in All Transit Trips Relative to Base[b] (%)
Base[c]		41 575	
Rail transit	14 147	46 560	+11.99
Commuter rail	4 197	42 332	+1.82
Busway	14 172	46 333	+11.44
Express bus	6 801	43 369	+4.32

[a] Includes all trips that have an origin or destination in the North Corridor or an origin in MetroCenter.
[b] Equals the percentage of new trips induced on each alternative.
[c] Refers to 1995 land use and highway system on 1978 transit network.

the local set of sequential travel demand models estimated by using UTPS software. Critical points of difference between the methods are described below.

STRUCTURAL AND CALIBRATION DIFFERENCES

The elasticity-based forecasting system is an incremental method in that changes in observed transit ridership are estimated as a function of changes in level-of-service. The four-step sequential forecasting procedure used to forecast patronage for the Baltimore alternatives, in contrast, is a synthetic method by which total regionwide transit travel is estimated from scratch for each alternative. Transit level-of-service and assignment are based on the minimum single transit path available (determined by the simple, unweighted sum of in-vehicle and out-of-vehicle time), including the minimum time access mode. Similarly, automobile level-of-service is measured on the single shortest path. Interzonal level-of-service variables included in the mode split model are in-vehicle travel time, out-of-vehicle travel time, and user cost. The resulting mode split is then applied to a fixed 1995 trip table. Transit trips are assigned to the network by using an all-or-nothing assignment procedure.

From the above descriptions, several critical differences with respect to the application of the two methods in Baltimore can be identified. First, the existing sequential models assume a fixed person trip table, but the elasticity method relaxes this assumption. Relaxation of the fixed trip table resulted in approximately 1000 additional trips in the case of rail transit. Second, transit level-of-service measures in the existing Baltimore mode split model are based on single minimum path level-of-service and, therefore, may present an optimistic measure of actual transit service used by all members of the public. The elasticity method, on the other hand, uses actual level-of-service faced by travelers on each transit path between a given O-D pair. The existing Baltimore assignment procedure involves all-or-nothing choice and is based only on travel time. The elasticity-based method diverts transit travelers by using a proportional assignment procedure based on several level-of-service variables. Note that the bias imposed by the use of minimum path level-of-service measures may be mitigated in that the coefficients of the existing mode split model were also estimated based on minimum path service measures. However, in many cases, the new facilities tested in this study provide significant service improvements, which leads to a greater difference between the minimum path and average path. Hence, this procedure results in upwardly biased estimates of transit travel.

The elasticity-based method also identifies distinct travel markets based on submodal choice, thereby avoiding the need to average level-of-service across submodes, which often leads to paradoxical results. In addition, this approach recognizes that automobile access to transit represents a distinct mode from walk or (feeder) bus access to transit and serves a different travel market segment. In the existing sequential models used, all transit modes are defined as a single mode that serves one travel market.

Finally, the two methods differ with respect to the level of calibration detail. Although the sequential method is applied at the transportation analysis zone level, the elasticity-based method employs sketch-planning zones. Therefore, the former method has the potential for measuring level-of-service with greater accuracy. The elasticity-based method, however, measures the changes in level-of-service exactly, based on travelers' submode and path for a given interzonal movement. The interzonal measures are used only as the large denominators in the calculations of percentage changes.

The impact of the structural and calibration differences identified above is that the sequential method is expected to result in larger diversions of automobile trips from the fixed trip table to the alternatives relative to the elasticity-based method. This is because of several optimistic assumptions regarding transit service employed in the sequential models, which are compounded in each step of the estimation procedure. In the first step, the minimum transit path is built. This pathbuilding results in an underestimate of actual transit travel time in three ways. First, not all transit users choose the path that has the minimum travel time. For example, automobile access may represent minimum access time, but not all users have an automobile available. Second, because the minimum pathbuilding method does not reflect that travelers weigh out-of-vehicle time more heavily than in-vehicle time, the model loads up new line-haul routes that minimize in-vehicle time relative to headways (wait time) and coverage (walk time). In addition, because the minimum path is both built and skimmed by using the unweighted sum of in-vehicle and out-of-vehicle time, the impact of a transfer between transit vehicles is underestimated, since a transfer imposes a higher proportion of out-of-vehicle time relative to total travel time. Similarly, cost affects travelers' route choice but is excluded in the building of the minimum path. Third, service frequency is excluded from the level-of-service measures. Because differences in frequency are important to travelers, the exclusion of frequency biases the patronage forecasts in favor of low frequency routes.

The above represent several of the major differences associated with the structural assumptions and calibration procedures of the two forecasting methods. Although the Baltimore application of UTPS is a very careful and elaborate procedure, a number of the assumptions reflect local practice rather than constraints imposed by UTPS software. For instance, some UTPS model sets build the minimum path based on a weighted sum of in-vehicle and out-of-vehicle time and cost. This definition of minimum path would reduce the error in the resulting patronage forecasts.

Total North Corridor boardings, inbound boardings, and the percentage of new trips estimated for each alternative by the two forecasting methods are compared in Table 2. (The boardings in Table 2 are lower than those in Table 1 because intra-Metro-Center trips are excluded for comparability with the available UTPS output.) Table 2 shows that the sequential models estimated with UTPS forecast a larger number of total boardings for all alternatives. The average difference between forecasts of total North Corridor boardings shown in Table 2 for the two methods is 41.4 percent. The average difference between forecasts of inbound boardings, however, is only 17.2 percent, which indicates that the greatest difference lies in outbound trips. The share of morning peak-period outbound trips forecast by the sequential method ranges from 45.1 percent for rail transit to 51.7 percent for commuter rail. The elasticity-based estimates of outbound trips ranges from 25.4 percent for commuter rail to 32.1 percent for rail transit.

The 1995 base transit trip table developed by using the elasticity-based method revealed about 25 percent outbound trips and the UTPS trip table revealed about 35 percent outbound trips. A higher

Table 2. Comparison of elasticity-based and UTPS Baltimore North Corridor patronage results by alternative, 1995.

Alternative	Elasticity-Based Method Boardings		UTPS Boardings	
	Total	Inbound	Total	Inbound
Rail transit	13 638	9262	18 508	10 168
Commuter rail	4 106	3063	8 828	4 268
Busway	11 182	8033	12 998	6 383
Express bus	5 587	3888	7 502	3 931

Note: Patronage refers to 1995 daily peak-period (7:00-9:00 a.m.) trips, excluding intraMetroCenter trips for comparability with available UTPS output.

proportion of outbound trips on the new alternatives is reasonable since the improvement in level-of-service relative to the base transit network is greater in the outbound direction; however, the share of outbound trips forecast by the sequential method is exaggerated. A principal reason for this is that, as noted earlier, the four-step sequential procedure used in this application underestimates the impact of a transfer. In the case of commuter rail, which has relatively long headways, this upward bias in favor of the new transit alternative is maximized.

Table 2 also indicates that the sequential method estimates fewer inbound busway boardings than does the elasticity method. Two major factors account for this difference. First, the provision of high-quality park-and-ride service under the busway alternative attracts a significant number of park-and-ride passengers. As noted previously, the elasticity method treats a new access mode as a new travel market, and the sequential method simply assigns a fixed number of transit riders to the minimum path access mode. Second, the elasticity method will estimate a greater number of new trips, all else being equal, because a fixed total trip table is not assumed.

The results of the two methods are also similar in several important ways. First, the light rail and busway alternatives attract considerably more trips than do the commuter rail and express bus alternatives in both methods. Second, a significant share of North Corridor boardings occur at the stations within Baltimore City where the population density is higher and incomes and automobile ownership are lower (relative to stations in Baltimore County). Finally, a significant minority of outbound trips are destined for the Towson area, an employment and population center in Baltimore County. These similarities increase our confidence in the patronage forecasts. In addition, our ability to explain differences in the forecasts based on assumptions implicit in the methods increases our confidence in the validity of the elasticity-based approach. [Note: the elasticity-based figures are being used as the final patronage results for local and Urban Mass Transportation Administration (UMTA) decisionmaking purposes in this UMTA-sponsored alternatives analysis.]

ADDITIONAL APPLICATIONS OF THE ELASTICITY-BASED METHOD

The Baltimore alternatives analysis patronage forecasting work represents one of the most complex applications of the elasticity-based method. The multitude of modest alternatives currently being considered in urban transportation clearly need easy-to-use forecasting methods for assessing their travel consequences. The forecasting methods should be subject to strict reasonableness tests. The validity of the results suggests that the method can be even more easily applied to transportation alternatives that have fewer network analysis requirements (e.g., requirements that affect travel on fewer existing links with fewer submodes.) Other applications of the method include estimating the ridership response to improvements in existing transit routes; determining the travel impacts of highway improvements and other automobile level-of-service changes, including parking strategies and gasoline price changes; and determining the optimum mix of fare and service changes for maximizing transit revenues.

The elasticity-based method provides a quick-turnaround, relatively inexpensive alternative to conventional large-scale travel models. The method saves personnel and computational resources without sacrificing accuracy. It relies on easily available ground count data and can be applied manually or with the use of simple computers. Also, the structure of the method is easily understood by transportation planners and its transparency allows the analyst to determine the impact of each step of the estimation procedure on the resulting forecasts.

ACKNOWLEDGMENT

The Baltimore patronage forecasting work described here was performed by Charles River Associates under subcontract to Parsons Brinckerhoff Quade and Douglas, Inc. The work was undertaken for the Maryland Department of Transportation, Mass Transit Administration, for the Baltimore North Corridor-Metro-Center alternatives analysis and environmental impact statement sponsored by UMTA. Katherine O'Neil, Robert Hirschey, and Kim Honetschlager of CRA participated in the collection and analysis of the Baltimore data. Terry Rosapep, Charles Goodman, and David Dunlap of the Regional Planning Council were extremely helpful and cooperative in providing the results and insights on the UTPS forecasting work. We are also deeply indebted to Ted von Briesen, MTA's project director, and Vern Bergelin, PBQD's project manager, who provided helpful comments and continuous support throughout the course of this study.

REFERENCES

1. B. Brand. Theory and Method in Land Use and Travel Forecasting, HRB, Highway Research Record 422, 1973, pp. 10-20.
2. W. Alonso. The Quality of Data and the Choice and Design of Predictive Models. In Urban Development Models, HRB, Special Rept. 97, 1968, pp. 178-192.
3. T.A. Domencich and G. Kraft. Free Transit. Charles River Associates, Cambridge, MA, 1970.
4. Y. Chan; Pennsylvania Transportation Institute. Review and Compilation of Demand Forecasting Experiences: An Aggregation of Estimation Procedures. U.S. Department of Transportation, June 1977.
5. A. Talvitie. A Direct Demand Model for Downtown Work Trips. Transportation, Vol. 2, 1973.
6. D. McFadden. The Measurement of Urban Travel Demand. Journal of Public Economics, Vol. 3, 1974.
7. Peat, Marwick, Mitchell and Company. Implementation of the N-Dimensional Logit Model: Final Report. Comprehensive Planning Organization, San Diego County, CA, 1972.
8. R.H. Pratt and Associates, Inc.; DTM, Inc. Development and Calibration of Mode Choice Models for the Twin Cities Area. Metropolitan Council, Minneapolis-St. Paul, MN, Nov. 1976.

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

Analytic Models of Trip Length Distributions

MOSHE BEN-AKIVA AND NICOLAOS LITINAS

This paper develops analytic models of trip length distributions. The models are derived from a destination choice model for a range of assumptions about the distributions of transportation level-of-service attributes and opportunities over the urban space. These models include all previously reported analytic trip length distributions. Their derivation from an explicit model of individual choice behavior illuminates their underlying assumptions about the urban space. It is shown how the parameters of the derived trip length distributions can be interpreted and estimated from available data that include estimated parameters of travel demand models and other readily available statistics on average speeds and fuel consumption. This makes these models useful for simplified analyses of various urban transportation policies, especially areawide pricing and travel time changes.

This paper derives a wide variety of analytic trip length distributions from underlying assumptions about travel behavior, transport system performance, and spatial distribution of travel opportunities. It develops the relations between parameters of trip length distributions and aggregate measures of transport level-of-service, land use, and socioeconomic variables. The resulting trip length distributions have policy-sensitive parameters and therefore can be used for simplified analyses of urban transportation policies and land use changes. The aggregate impacts of changes in the patterns of travel speeds and travel costs can be predicted by such a model with fewer input data and calculations than by a discrete destination choice model that distributes traffic among a number of origins and destinations.

A variety of models have been proposed and used to describe trip length patterns and urban densities (population, employment or joint populations, and employment densities). These include the exponential model [attributed to Clark (1) but first applied by Bleicher in 1892 (2) to analyze Frankfurt data], the square root exponential model (3), the gamma model (3,4), the normal model (5), the shifted normal model (6), the generalized normal model (7), the generalized gamma model (8), the power model (9-11), the beta-type model (12), and the combined exponential and gamma model (13). The normal model has also been used with a directionally dependent variance (7,14,15). Joint population and employment densities were modeled by the bivariate normal model (14) and the quadrivariate normal model (16-18). Theoretical justification is provided for several of the above models through the framework of spatial equilibrium of deterministic utility functions with or without a competitive housing market (17,19-23). Several models were derived by using the entropy maximizing approach (24-26) and from the gravity model (18). A few models have also been derived from random utility theory. Ben-Akiva and Watanatada (27) derived a truncated gamma-2 trip length distribution based on the continuous logit model and Goodwin (28) derived a gamma trip length distribution and Mogridge (29,30) a Weibull distribution based on a slightly different approach. Empirical validations and comparisons of alternative models exist in a number of sources [for example, Casetti (21), Genest (31), Pearce and others (32), McDonald and Bowman (33), Clickman and Oguri (34), and Horowitz (35)].

In this paper all the trip length distributions found in the published literature are derived as special cases of the continuous spatial choice logit model. Their derivation from the continuous logit model clarifies their underlying assumptions and offers ways for their improvement. It also offers a basis for comparison and selection among alternative models for specific applications. Furthermore, a few new, more general models are derived.

ASSUMPTIONS ABOUT URBAN SPACE UNDER CIRCULAR SYMMETRY

The derivation of the trip length distributions in this paper is based on the assumption of circular symmetry around the decisionmaker's origin. This is an approximation of the complex urban patterns that can be employed here for analytical convenience because of the nondirectional nature of the analysis. The results demonstrate that even this highly simplifying assumption leads to valid trip length distributions that are expressed as functions of a small number of parameters.

Assumptions About Generalized Transport Cost Surface

Assume a circularly symmetric (around the decisionmaker) generalized cost surface, $B(l,\phi|r_w,\theta_w)$, in units of generalized cost per unit distance, given by (see the coordinate system in Figure 1)

$$B(l,\phi|r_w,\theta_w) = [c/(1 + c_o)] + b_1(1 + b_o)^{\nu-1} \quad (1)$$

where c, c_o, b_o, ν are parameters that have specific values by mode and decisionmaker. Equation 1 says that the generalized cost surface depends only on distance and not on directionality. This approximation is most accurate in situations with no directional congestion and for trips that start or end at the city center. A detailed discussion on the derivation of this surface from transportation system performance and the interpretation of its parameters is presented later. For $\nu = 1$ this surface can be derived from the velocity field used by Blumenfeld and Weiss (36) and is also an approximation of a generalized cost surface based on the velocity field tested by Angel and Hyman (37). For this case b_1 can be interpreted as the generalized cost per unit distance at free flow (i.e., $l = \infty$) and c/c_o is the difference between the generalized cost at the most congested point ($l = 0$) and a free flow location.

From this circularly symmetric generalized cost surface, the value of the utility β for a trip from w to h can be derived as a function of the distance l, as follows (for $\nu>0$):

$$\beta(l) = -\int_0^l B(\zeta,\phi|r_w,\theta_w)d\zeta = c\ln c_o + (b_1/\nu)b_o^\nu - c\ln(1 + c_o) - (b_1/\nu)(1 + b_o)^\nu \quad (2)$$

Let

$$b \equiv b_1/\nu \quad (3a)$$

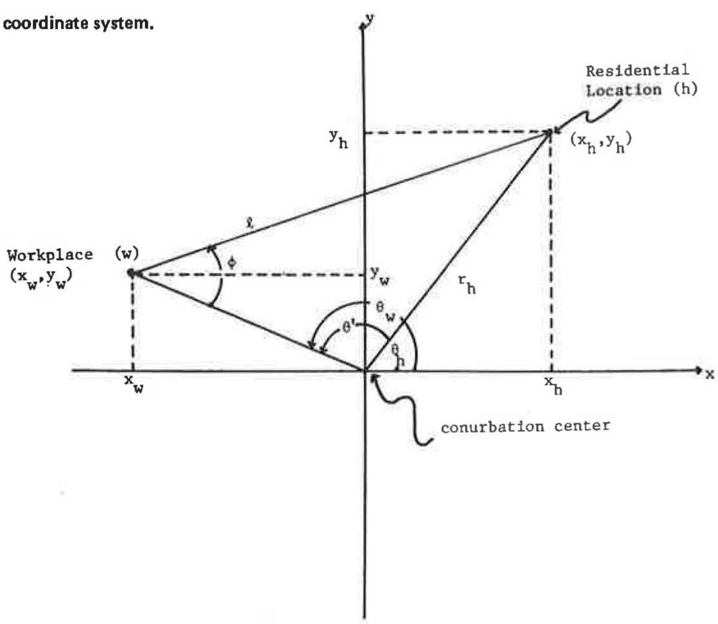
and

$$a_1 \equiv -c\ln c_o - b\, b_o^\nu \quad (3b)$$

to obtain,

$$\beta(l) = -a_1 - b(l + b_o)^\nu - c\ln(l + c_o) \quad (4)$$

Figure 1. Spatial coordinate system.



For the case $b_o = c_o = 0$ the β function can be defined as follows:

$$\beta(l) = -a_1 - bl^{\nu} - c\ln l \qquad (5)$$

This form with $\nu = 1$ appears in the literature under the name Tanner function. Mogridge (29) has also suggested the form bl^{ν} with $0<\nu<1$ as a suitable approximation for the generalized cost.

The average utility of a trip from w to h is given by the above β function plus an additive constant that represents urban area, trip origin, mode, and decisionmaker-specific characteristics. Let a denote the sum of this constant with the above constant of integration a_1 to obtain the average utility as follows:

$$V(l) = -a - b(l + b_o)^{\nu} - c\ln(l + c_o) \qquad (6)$$

Assumptions About Spatial Opportunity Density

Under circular symmetry, one of the most general assumptions for the opportunity density function that represents travel attractions is the form

$$\gamma(l) = \gamma_o(l + \lambda)^{\gamma-2}\exp[-\delta(l + \lambda)^{\xi}] \qquad (7)$$

where γ_o, γ, δ, ξ, and λ are parameters that depend on the urban area and the decisionmaker. This form equals the kernel of the generalized gamma density function. We will show later that the generalized gamma density contains a rich set of density functions, including normal, hydrograph, Rayleigh, Maxwell, Weibull, chi-squared, and gamma [see, for example, Johnson and Kotz (38), a special case of interest primarily because most of the existing models can be derived by using some particular subcase of this form]. It is obtained by setting the translation parameter λ in Equation 7 equal to zero, to get

$$\gamma(l) = \gamma_o l^{\gamma-2}\exp(-\delta l^{\xi}) \qquad (8)$$

CONTINUOUS LOGIT DESTINATION CHOICE MODEL WITH CIRCULARLY SYMMETRIC URBAN SPACE

The spatial choice logit density function for a circular attraction area is expressed in polar coordinates, as follows (27,39):

$$f(l,\phi) = \exp[V(l,\phi)]\,\gamma(l,\phi)/\int_0^{r^*}\int_0^{2\pi}\exp[V(l,\phi)]\,\gamma(l,\phi)\,l\,dl\,d\phi \qquad (9)$$

where r* denotes the radius of the boundary of the attraction area. Substitute in this model the assumptions presented in the previous section (i.e., Equations 6 and 7) to obtain:

$$f(l,\phi) = \gamma_o(l+\lambda)^{\gamma-2}\exp[-\delta(l+\lambda)^{\xi} - a - b(l+b_o)^{\nu} - c\ln(l+c_o)]$$
$$\div \int_0^{r^*}\int_0^{2\pi}\gamma_o(l+\lambda)^{\gamma-2}\exp[-\delta(l+\lambda)^{\xi} - a - b(l+b_o)^{\nu} - c\ln(l+c_o)]\,l\,dl\,d\phi \qquad (10)$$

The trip length distribution is defined as,

$$f(l) = \int_0^{2\pi} f(l,\phi)\,l\,d\phi = 2\pi l f(l,\phi) \qquad (11)$$

Since under circular symmetry the density at a point can be obtained by dividing f(l) by 2πl, the following analysis considers the derivation of trip length distributions only. Substitute Equation 10 in Equation 11 to obtain the following trip length distribution for a circularly symmetric logit model:

$$f(l) = \frac{(l+\lambda)^{\gamma-2}(l+c_o)^{-c}\,l\exp[-b(l+b_o)^{\nu} - \delta(l+\lambda)^{\xi}]}{\int_0^{r^*}(l+\lambda)^{\gamma-2}(l+c_o)^{-c}\,l\exp[-b(l+b_o)^{\nu} - \delta(l+\lambda)^{\xi}]\,dl} \qquad (12)$$

Without the translation parameters ($\lambda = b_o = c_o = 0$) Equation 12 simplifies to

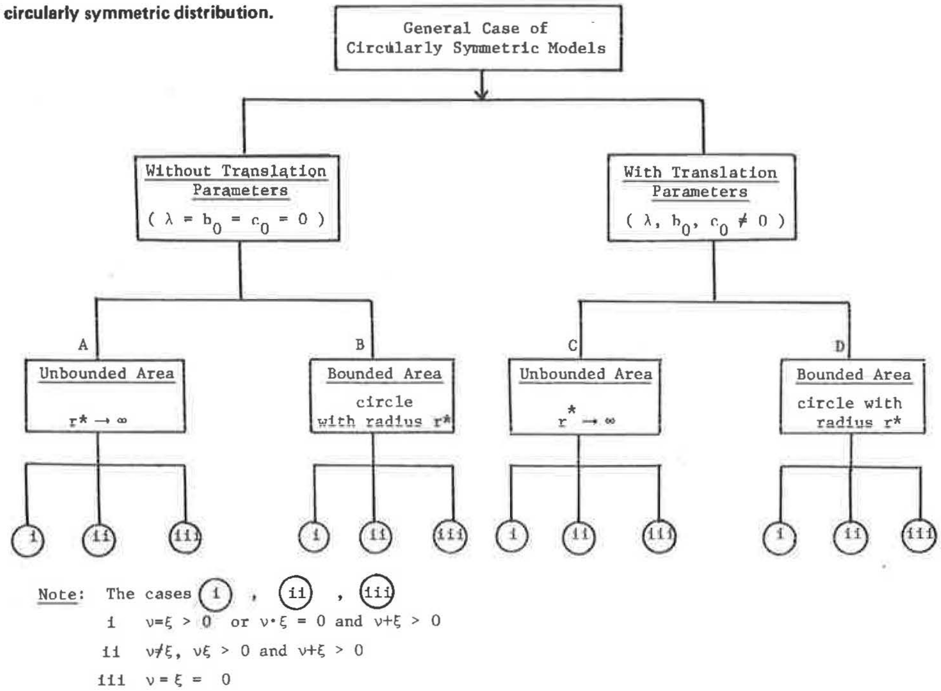
$$f(l) = l^{\gamma^*-1}\exp(-bl^{\nu} - \delta l^{\xi})/\int_0^{r^*} l^{\gamma^*-1}\exp(-bl^{\nu} - \delta l^{\xi})\,dl \qquad (13)$$

where γ^* is identical to $\gamma - c$.

Equation 12 is the general form of the circularly symmetric continuous logit trip length distribution. The integral in the denominator cannot be evaluated analytically except for special cases. These analytic solutions are given in the following section for the classification of special cases shown in Figure 2. Each special case is defined as a combination of the following:

1. With or without translation parameters,
2. Finite or infinite radius r*, and
3. One of the following ranges of values for ν and ξ: (a) $\nu = \xi > 0$ or $\nu \cdot \xi = 0$ and $\nu + \xi > 0$, (b) $\nu \neq \xi$, $\nu \cdot \xi > 0$ and $\nu + \xi > 0$, or (c) $\nu = \xi = 0$.

Figure 2. Special cases of circularly symmetric distribution.



TRIP LENGTH DISTRIBUTIONS

The complete list of the analytic trip length density functions according to the classification of Figure 2 is given in Litinas and Ben-Akiva (39). This section summarizes the key results and discusses specific assumptions and the relations among special cases. Particular emphasis is placed on the special cases that correspond to the models that have appeared in the literature.

Models for Unbounded Urban Area Without Translation Parameters

In this case, the trip length density function is given by Equation 13, with $r^* = \infty$. Two major families of models are derived: the generalized gamma and the shifted generalized gamma. For the third case the distribution vanishes.

Generalized Gamma Models ($\nu = \xi = \nu^*$ or $\nu\xi = 0$ and $\nu + \xi > 0$)

The trip length density is given by

$$f(l) = \text{Generalized gamma}\,(\nu^*, \gamma^*, b^*) \qquad (14)$$

where b^* is identical to $b + \delta$. Note that for the case $\xi = 0$ or $\delta = 0$ the model is

$$f(l) = \text{Generalized gamma}\,(\nu, \gamma^*, b) \qquad (15)$$

and for the case $c = 0$ and $\xi = 0$ (or $\delta = 0$), it is

$$f(l) = \text{Generalized gamma}\,(\nu, \gamma, b) \qquad (16)$$

Thus, the models in Equations 14-16, which are based on different assumptions, have the same form.

The generalized gamma model is equivalent to the model used by Blumenfeld and others (8) and labeled as the generalized Clark and Sherratt model. Depending on the particular values of ν^*, four types of distributions can be distinguished within the generalized gamma family: gamma, generalized Gauss, generalized Weibull, and other.

Gamma Models ($\nu^* = 1$)

For $\nu^* = 1$ or $[\nu = 1$ and $\delta = 0$ (or $\xi = 0$)] or $[\nu = 1$ and $\delta = 0$ (or $\xi = 0$) and $c = 0]$ the generalized gamma density becomes the gamma density:

$$f(l) = \text{Gamma}\,(\gamma^*, b^*) \qquad (17)$$

This model was studied by Ajo (3), Aynvarg (4), and Blumenfeld (7) and has had numerous applications (20-22, 29,33,37,40-42).

Special cases of the gamma model are obtained for different values of γ^*. The model for $\gamma^* = 1$ is

$$f(l) = \exp(b^*) \qquad (18)$$

In this model, if $\delta = 0$ (or $\xi = 0$) the opportunities are assumed to decline with the reciprocal of distance. For $\gamma^* = 2$, the model is

$$f(l) = \text{Gamma}\,(2, b^*) \qquad (19)$$

This model with $\delta = 0$ (or $\xi = 0$) and $c = 0$ represents an assumption of a featureless plane with uniformly distributed opportunities (27) and is equivalent to an exponential density at a point (11). It has had numerous applications (1,2,7,19, 21,23,27,32-34,43-48).

Generalized Gauss Models ($\nu^* = 2$)

For $\nu^* = 2$ or $[\nu = 2$ and $\xi = 0$ (or $\delta = 0$)] or $[\nu = 2$ and $\delta = 0$ (or $\xi = 0$) and $c = 0]$, the model is

$$f(l) = \text{Generalized Gauss}\,[\gamma^*, 0, (1/2b^*)] \qquad (20)$$

It includes as special cases the following distributions: normal, Rayleigh (which is also called circular normal), and Maxwell (which is also sometimes called spherical normal). It is equivalent to the generalized Sherratt model used by Blumenfeld (7).

Special cases of the generalized Gauss model are obtained by specific values of γ^*. The model for $\gamma^* = 1$ is

$$f(l) = 2\ \text{Normal}\ [0, (1/2b^*)] \quad (21)$$

For $\delta = 0$ (or $\xi = 0$) and $c = 0$ it corresponds to declining opportunities with the reciprocal of distance and increasing travel costs with the square of distance. For $\gamma^* = 2$, the model is

$$f(l) = \text{Rayleigh}\ [0, (1/2b^*)] \quad (22)$$

which is the same as the Sherratt model for population density at a point. In rectangular coordinates the density at a point (x,y) is an independent bivariate normal with zero means and variances $1/2b^*$. For $\delta = 0$ (or $\xi = 0$) and $c = 0$ it represents a trip length distribution under a featureless plane assumption. The Rayleigh and the normal models have been used in numerous applications (5,7,20-22,32,33, 49). The model for $\gamma^* = 3$ is

$$f(l) = \text{Maxwell}\ [0, (1/2b^*)] \quad (23)$$

Generalized Weibull Models ($\nu^* = n\gamma^*$)

The generalized Weibull distribution is obtained from the generalized gamma by letting $\nu^* = n\gamma^*$, where n is a positive integer.

$$f(l) = \text{Generalized Weibull}\ (\gamma^*, b^*, n) \quad (24)$$

For $n = 1$ it becomes the Weibull density,

$$f(l) = \text{Weibull}\ (\gamma^*, b^*) \quad (25)$$

This model was also derived by Mogridge (29). The model for $n = 2$ is

$$f(l) = \text{Modified Weibull}\ (\gamma^*, b^*) \quad (26)$$

Other Models

Other values of ν^* result in other types of distributions. A special case of interest is the model for $\nu^* = 1/2$,

$$f(l) = \text{Generalized square root exp}(\gamma^*, b^*) \quad (27)$$

from which more specialized cases can be obtained for different values of γ^*. For example, for $\gamma^* = 1$ the model is

$$f(l) = \text{Square root exp}(b^*) \quad (28)$$

and for $\gamma^* = 2$, it becomes

$$f(l) = \text{Generalized square root exp}(2, b^*) \quad (29)$$

The last model was used by Ajo (1), Cassetti (21), and Papageorgiou (22).

Shifted Generalized Gamma Models ($\nu \neq \xi$ and $\nu\xi > 0$, $\nu + \xi > 0$)

The shifted generalized gamma distribution is defined by Equation 13 with $r^* = \infty$ and by restricting the exponents ν, ξ to unequal positive values. A closed form solution for the whole family of these models does not exist. However, solutions may be obtained for certain specific values of ν and ξ. The shifted generalized Gauss distribution is obtained when one exponent equals 2 and the other equals 1.

For $\nu = 2$ and $\xi = 1$ the model is

$$f(l) = \text{Shifted generalized Gauss}\ (\gamma^*, -\delta/2b, 1/2b) \quad (30)$$

The following special cases are of interest:

$$\text{For } \gamma^* = 1,\ f(l) = \text{Shifted normal}\ (-\delta/2b, 1/2b) \quad (31)$$

$$\text{For } \gamma^* = 2,\ f(l) = \text{Shifted Rayleigh}\ (-\delta/2b, 1/2b) \quad (32)$$

$$\text{For } \gamma^* = 3,\ f(l) = \text{Shifted Maxwell}\ (-\delta/2b, 1/2b) \quad (33)$$

For $\nu = 1$ and $\xi = 2$ the models are obtained by interchanging δ and b in the above distributions.

The shifted Rayleigh trip length distribution is equivalent to the shifted normal for population density at a point. It has numerous applications (6, 20,21,32-34).

Models for Bounded Urban Area Without Translation Parameters

For a bounded urban area without translation parameters three general families of models are derived from Equation 13:

1. Truncated generalized gamma density,
2. Truncated shifted generalized gamma density, and
3. Power density.

Distributions 1 and 2 have identical kernels to the distributions of the generalized gamma and the shifted generalized gamma. In many previous applications these models were applied in a form that contains the kernel of a distribution multiplied by a constant that was not derived explicitly. The power density is obtained from Equation 13 for $\nu = \xi = 0$,

$$f(l) = \text{Power density}\ (\gamma^*, r^*) \quad (34)$$

This distribution was proposed by Harwood (9) and Smeed (10,11), and applied by Pearce and others (32) and Pearce (50).

Models for Unbounded Urban Area with Translation Parameters

Three general families of models are derived from Equation 12 with $r^* = \infty$.

Combined Generalized Gamma Models ($\nu^* = \nu = \xi$)

It was not possible to obtain a closed-form solution for the general model except for the special cases with equal translation parameters, $\lambda = c_o = b_o = \lambda^*$. The model for $\nu^* = 1$ is independent of b_o and in general can be expressed as a sum of gamma densities.

For $\gamma^* = 2$ it reduces to gamma $(2,b^*)$. For $\gamma^* = 3$ it results in a convex combination of gamma $(2,b^*)$ and gamma $(3,b^*)$,

$$f(l) = [b^*\lambda^{\#}/(2 + b^*\lambda^{\#})]\ \text{gamma}\ (2, b^*) + [2/(2 + b^*\lambda^{\#})]\ \text{gamma}\ (3, b^*) \quad (35)$$

where $\lambda^{\#} = c_o = \lambda$.

This model divided by $2\pi l$ is equivalent to the model of population density at a point proposed by Reinhart (13) and further analyzed by Casetti (21) and Papageorgiou (22). Other values of γ^* result in different combinations of gamma densities.

Solutions can also be obtained for special cases of $\lambda \neq c_o$. For example, letting $c_o = 0$ results in a combination of a gamma 2 with an exponential density.

The general model for $\nu^* = 2$ is the combined generalized Gauss density. For $\gamma^* = 2$ it reduces to the shifted Rayleigh density $(\lambda^*, 1/2b^*)$ which has the same form as Equation 32 but with different interpretation of the parameters. The case $\gamma^* = 3$

results in a combination of a shifted Mawell with a shifted Rayleigh density.

Combined Shifted Generalized Gamma Models
($\nu \neq \xi$ and $\nu\xi>0$, $\nu + \xi>0$)

In the combined shifted generalized gamma models case even the special case for $\nu = 2$, $\xi = 1$ or $\nu = 1$, $\xi = 2$ could not be solved in closed form. However, for $\lambda = c_o = \lambda^{\#}$ and $\gamma^* = 2$ the shifted Rayleigh density is obtained and $\gamma^* = 3$ results in a combination of a shifted Maxwell with a shifted Rayleigh. These models have the same functional forms as those of the models for the combined generalized gamma models with $\nu^* = 2$ with different interpretations of the parameters.

Combined Generalized Beta Prime Models
($\nu = \xi = 0$)

The general form of the combined generalized beta prime models can be expressed as a combination of generalized beta prime densities. The model for $c_o>\lambda>0$ is a combination of two generalized beta primes [(γ, $c - \gamma$, λ, c_o) with ($\gamma - 1$, $c - \gamma + 1$, λ, c_o)], which for $\lambda = 0$ reduces to

$$f(l) = \text{Generalized beta prime}\,(\gamma, c - \gamma, 0, c_o) \quad c > \gamma > 0 \tag{36}$$

and further simplifies to

$$f(l) = \text{Beta prime}\,(\gamma, c - \gamma) \tag{37}$$

for $c_o = 1$. Similarly, the model for $\lambda>c_o>0$ is another combination of two generalized beta primes [($2 - c$, $c - \gamma$, c_o, λ) with ($1 - c$, $c - \gamma + 1$, c_o, λ)]. For $c_o = 0$ this model reduces to

$$f(l) = \text{Generalized beta prime}\,(2 - c, c - \gamma, 0, \lambda) \quad \gamma < c < 2 \tag{38}$$

and to

$$f(l) = \text{Beta prime}\,(2 - c, c - \gamma) \tag{39}$$

for $\lambda = 1$. For the third possibility of $\lambda = c_o = \lambda^{\#}>0$, the model is

$$f(l) = \text{Generalized beta prime}\,(2, c - \gamma, 0, \lambda^{\#}) \quad c > \gamma \tag{40}$$

which for $\lambda^{\#} = 1$ becomes

$$F(l) = \text{Beta prime}\,(2, c - \gamma) \tag{41}$$

The beta prime densities in Equation 37 with $\gamma = 2$, Equation 39 with $c = 0$, and Equation 41 correspond to the simple potential models proposed and used by Stewart and Warntz (51) and Warntz (52) for rural population densities. The generalized beta prime models in Equation 36 with $\gamma = 2$, Equation 38 with $c = 0$, and Equation 40 correspond to a modified simple potential model for population densities investigated by Papageorgiou (22). Casetti (21) used the simple potential model with good results in peripheral areas of a number of cities.

Models for Bounded Urban Area With Translation Parameters

For models for bounded urban area with translation parameters Equation 12 is used to derive the truncated counterparts of the distributions in the models for unbounded urban area with translation parameters.

In the combined generalized beta prime models for the following opportunity density function

$$\gamma(l) = \gamma_o^* (r^* - l)^{\gamma - 2} \tag{42}$$

and for $c_o = 0$, the model is

$$f(l) = \text{Generalized beta}\,(2 - c, \gamma - 1, 0, r^*) \tag{43}$$

This model with $c = 1$ (or $c = 0$ for the population density at a point) corresponds to a model proposed by Mills (12) and discussed by McDonald and Bowman (33).

ESTIMATION OF PARAMETERS

The purpose of this section is to relate the parameters of the models with the transportation system performance and the decisionmaker characteristics. For illustrative purposes, take the case of travel by automobile (A).

First consider performance of the transportation system. Continuous surfaces are used to describe the travel time and travel cost per unit distance at each point of the urban space. Consider the following circularly symmetric travel time surface centered at the traveler's origin (see Figure 1 for the coordinate system):

$$t_A(l, \phi \mid r_w, \theta_w) = (c_A'/l + c_{oA}) + b_{lA}'(l + b_{oA})^{\nu_A - 1} \tag{44}$$

where $t_A(l,\phi \mid r_w,\theta_w)$ is the travel time [min/unit distance at the point (l, ϕ)] and c_A', c_{oA}, b_{lA}', b_{oA}, ν_A are parameters that can be estimated from observation and can be influenced by policies.

The first term of Equation 44 decreases with increasing l, and the behavior of the second term depends on the value of ν_A. For $0<\nu_A<1$ it also decreases with l but less rapidly than does the first term. For $\nu_A = 1$ it is constant, and for $\nu_A>1$ it increases with l. It is reasonable to assume that $\nu_A \geq 1$. Then, the effect of the first term is more important in locations close to the travelers' origin and the weight of the second term is greater for distant locations. This functional form has enough flexibility to allow the representation of a wide range of travel time fields observed in urban areas and used by Blumenfeld and Weiss (36) and Angel and Hyman (37).

Consider the following travel cost surface:

$$C_A(l, \phi \mid r_w, \theta_w) = AVMMC_A \cdot t_A(l, \phi \mid r_w, \theta_w) \tag{45}$$

where $C_A(l,\phi \mid r_w,\theta_w)$ is travel cost [¢/unit distance (l,ϕ)] and $AVMMC_A$ is the average monetary cost of travel by automobile (¢/min).

Equation 45 implies that the travel cost per unit distance increases as the travel speed decreases. The average travel cost per minute ($AVMMC_A$) can be approximated from an average travel cost per mile ($AVMC_A$) as follows:

$$AVMMC_A = AVMC_A \cdot \bar{V}_A/60 \tag{46}$$

where $\bar{V}_A$ is average travel speed (miles/h) and $AVMC_A$ is average monetary cost of travel by automobile (¢/mile).

The average cost per mile can be related to gasoline price, fuel efficiency, and other costs such as maintenance costs as follows:

$$AVMC_A = (GPRICE/MPG) + MC \tag{47}$$

where

GPRICE = gasoline price (¢/gal),
MPG = miles per gallon of gasoline, and
MC = maintenance costs (¢/mile).

A different travel cost surface can be derived from the following relation of gasoline consumption to specific automobile characteristics:

$$\varphi_A(l,\phi|r_w,\theta_w) = K_{1A} + (K_{2A}/60)t_A(l,\phi|r_w,\theta_w) \quad (48)$$

where

$\varphi_A(l,\phi|r_w,\theta_w)$ = gasoline consumption [gal/unit distance at (l,ϕ)],
K_{1A} = gasoline consumed to overcome the rolling resistance (gal/unit distance), and
K_{2A} = gasoline consumed to overcome mechanical losses (gal/h).

The above equation has been investigated by several researchers in numerous experiments. This equation was found to adequately explain fuel consumption for different drivers who are driving normally in urban traffic and for speed $\lesssim$70 km/h [see for example, Evans and others (53), Evans and Herman (54,55), Chang and others (56), and Chang and Herman (57)]. The use of this equation is associated with a simplified fuel-consumption model theoretically derived by Amann and others (58). The parameters K_{1A} and K_{2A} can be inferred from the weight of the car and the idle fuel flow rate as follows:

$$K_{1A} = k_{1A}W_A \quad (49a)$$
$$K_{2A} = k_{2A}I_A \quad (49b)$$

where

W_A = weight of the car (lb),
I_A = idle fuel flow rate (gal/h), and
k_{1A}, k_{2A} = constants.

Evans and Herman (55) provide values of k_{1A}, k_{2A}, W_A, and I_A for various cars. Based on Equation 48 the travel cost surface is derived as follows:

$$C_A(l,\phi|r_w,\theta_w) = MC + GPRICE\,\varphi(l,\phi|r_w,\theta_w) \quad (50)$$

This cost surface allows a more elaborate analysis of automobile-related policies (for example, the effect of smaller-size cars on gasoline consumption).

The generalized cost surface $[B_{tA}(l,\phi|r_w,\theta_w)]$, which expresses the disutility per unit distance at the point (l,θ) perceived by traveler t, is derived from the above travel time and travel cost surfaces as follows. Assume that the disutility of travel is a linear combination of travel times and travel costs. Then, $B_{tA}(l,\phi|r_w,\theta_w)$ can be written as,

$$B_{tA}(l,\phi|r_w,\theta_w) = MUTT_t\, t_A(l,\phi|r_w,\theta_w) + (MUTT_t/VT_t)C_A(l,\phi|r_w,\theta_w) \quad (51)$$

where $MUTT_t$ is the marginal utility of travel time for decisionmaker t, which can be inferred from existing estimated discrete logit models, and VT_t is the value of time for decisionmaker t (¢/min).

The value of time (VT_t) is often estimated as a percentage of the wage rate as follows:

$$VT_t = (PWRVT_t)(INC_t)/1200 \quad (52)$$

where

INC_t = annual income for decisionmaker t ($),
$PWRVT_t$ = percentage of wage rate for the value of time, and
1200 = factor that converts annual income to wage rate, assuming 250 working days/year, (¢/min).

Substitution of the above travel time and travel cost surfaces in equation 51 yields the form of Equation 1 as follows:

$$B_{tA}(l,\phi|r_w,\theta_w) = [c_{tA}/(l + c_{oA})] + b_{ltA}(l + b_{oA})^{\nu_A - 1} \quad (53)$$

where

$c = c_{tA}$,
$b_1 = b_{1tA}$,
$\nu = \nu_A$,
$c_o = c_{oA}$, and
$b_o = b_{oA}$.

For the cost surface assumption of Equation 45 the parameters are evaluated as follows:

$$c_{tA} = MUTT_t\, c_A'\left(\{[(GPRICE/MPG) + MC]\,20\bar{V}_A/(PWRVT_t) \times (INC_t)\} + 1\right) \quad (54a)$$

$$b_{ltA} = MUTT_t\, b_{lA}'\left(\{[(GPRICE/MPG) + MC]\,20\bar{V}_A/(PWRVT_t) \times (INC_t)\} + 1\right) \quad (54b)$$

For the cost assumption of Equation 50 and for the case of $\nu_A = 1$ the following expressions for the parameters are obtained:

$$c_{tA} = MUTT_t\, c_A' \{1 + [20k_{2A}I_A GPRICE/(PWRVT_t)(INC_t)]\} \quad (55a)$$

$$b_{tA} = MUTT_t\{b_{lA}' + [20k_{2A}I_A b_{lA}' GPRICE/(PWRVT_t)(INC_t)] + [1200/(PWRVT_t)(INC_t)](MC + k_{1w}W_A GPRICE)\} \quad (55b)$$

Note that in this case $b_{tA} = b_{1tA}$.

Thus, all the parameters of the generalized cost surface of Equation 1 have been related to the transportation system performance and the decisionmaker characteristics. The β function derived from this surface is given in Equation 2. The average utility function (V) given in Equation 6 equals the sum of this β function with the trip origin, mode, and traveler-specific constants. However, the additional parameter (a) of Equation 6 (denoted here as a_{tAw}) does not enter the expressions for the automobile trip length distributions and therefore it will not be evaluated here in terms of other variables.

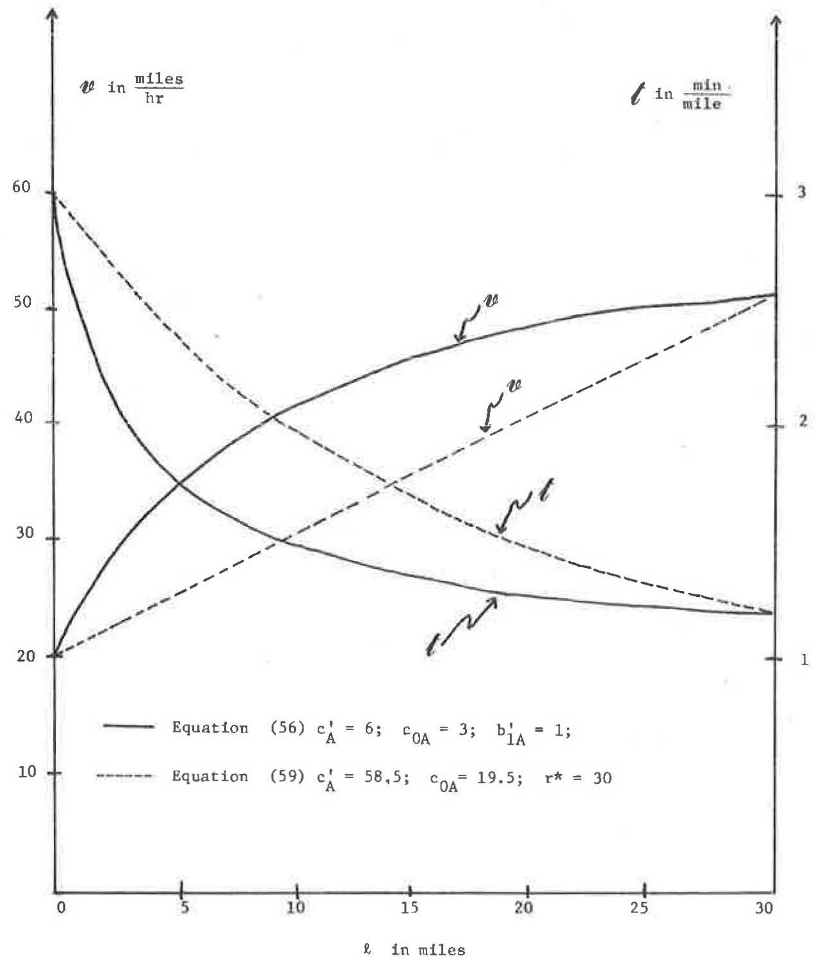
The above relations give a behavioral interpretation and a method of calculation for the parameters of all the derived trip length distributions and permits the use of these models for simplified policy analysis.

Below, two special cases of the above results are presented in more detail. For these cases a slightly different interpretation of the parameters is also possible, such as the case $\nu_A = 1$. This case covers a broad range of the derived trip length models. The travel time surface of Equation 44 for $\nu_A = 1$ is

$$t_A(l,\phi|r_w,\theta_w) = [c_A'/(l + c_{oA})] + b_{lA}' \quad (56)$$

This surface has been used by Blumenfeld and Weiss

Figure 3. Simplified velocity and travel time surface.



(36) and its shape is shown in Figure 3. Note that the following relation exists between its parameters:

$$c'_A/c_{oA} = tt^o_A - tt^*_A \quad (57)$$

where tt^*_A is the free-flow travel time [min/unit distance (i.e., $l = \infty$)], which is equal to b'_{1A} and tt^o_A is the travel time [min/unit distance at (r_w,θ_w)].

Apply the previously described assumptions to derive the average utility function as follows:

$$V_{tA}(l) = -a_{tAw} - c_{tA} \ln(l + c_{oA}) - b_{tA}\, l \quad (58)$$

To get the parameter values of Equation 58, first estimate c'_A, c_{0A}, and b'_{1A} from travel time field observations. Assume the travel cost surface of Equation 45 and obtain estimates of GPRICE, MPG, MC, $\bar{V}_A$, $PWRVT_t$, INC_t, and $MUTT_t$. Substitute these values in Equation 54 to compute c_{tA}, b_{1tA}. For this case $b_{tA} = b_{1tA}$.

For the case of a bounded urban area another interpretation of Equation 58 is possible. Assume that the travel speed is linearly increasing from the traveler's origin to the city boundary. The resulting travel time surface (see Figure 3) is as follows:

$$t_A(l, \phi | r_w, \theta_w) = c'_A/(l + c_{oA}) \quad (59)$$

where

$$c_{oA} = r^* V^o_A/(V^*_A - V^o_A), \quad (60a)$$

$$c'_A = 60\, r^*/(V^*_A - V^o_A), \quad (60b)$$

V^o_A = travel speed (mph at traveler's origin),
V^*_A = travel speed (mph at r*), and
r* = the radius of city boundary from traveler's origin (miles).

Assume that the travel cost surface is given by Equation 50. Then substitute Equation 60 in Equation 55 for $b'_{1A} = 0$ to obtain the values of c_{tA} and b_{tA}.

For the case of $c'_A = 0$ it is also required that $b_{oA} = 0$. The travel time surface is then

$$t_A(l, \phi | r_w, \theta_w) = b'_{lA}\, l^{\nu_A} \quad (61)$$

Assume that the travel cost surface of Equation 45 is applicable. Then, these assumptions result in the following utility function:

$$V_{tA}(l) = -a_{tAw} - b_{tA}\, l^{\nu_{tA}} \quad (62b)$$

where b_{tA} is given by Equation 54b. Note that this utility function covers all the trip length distributions of categories A and B (i.e., without translation parameters) by substituting $\gamma^* = \gamma$.

Now consider the following alternative behavioral assumption that leads to the same functional form of the utility function. The decisionmaker perceives the disutility of travel as a generalized cost to the ν_t power. Then, the following average utility is derived

$$V_{tA}(l) = -a_{tAw} - b_{tA}\, l^{\nu_A} \tag{62a}$$

where

$$\nu_{tA} = \nu_t \nu_A \tag{63a}$$

and

$$b_{tA} = \text{MUTT}_t \left[(b'_{iA}/\nu_A) \left(\left\{ [(\text{GPRICE}/\text{MPG}) + \text{MC}]\, 20\bar{V}_A \div (\text{PWRVT}_t)(\text{INC}_t) \right\} + 1\right)\right]^{\nu_t} \tag{63b}$$

This interpretation allows the use of models that have an exponent $\nu_{tA} \neq 1$ under the assumption of a constant travel time surface (i.,e., $\nu_A = 1$). For this case, the more elaborate travel cost surface of Equation 50 may also be used.

REFERENCES

1. C. Clark. Urban Population Densities. Journal of the Royal Statistical Society, Series A114, 1951, pp. 490-496.
2. H. Bleicher. Statische Berschreibung Der Stadt Frankfurt Am Main Und Ihrer Bevolkerung. Frankfurt Am Main, Germany, 1892.
3. R. Ajo. On the Structure of Population Density in London's Field. Acta Geographica, Vol. 18, 1965, pp. 1-17.
4. Y. Aynvarg. Zones of Influence of Middle Sized Cities, Their Boundaries and Passenger Flows. Soviet Geography 10, 1969, pp. 549-558.
5. G.C. Sherratt. A Model for General Urban Growth. In Mathematical Sciences: Models and Techniques (O.W. Churchman and M. Verhulst, eds.), Proc., 6th International Meeting of the Institute of Management Science, Pergamon Press, New York, 1960, pp. 147-159.
6. B.E. Newling. The Spatial Variation of Urban Population Densities. Geographical Review, Vol. 59, 1969, pp. 242-252.
7. D.E. Blumenfeld and G.H. Weiss. Some Radical and Directional Dependent Models for Densities of Homes and Workplaces. Transportation Research, Vol. 8, 1974, pp. 149-155.
8. D.E. Blumenfeld and others. Spatial Distributions of Homes for Journeys to Work by Different Modes of Transportation. Transportation Research, Vol. 9, 1975, pp. 19-23.
9. E.M. Horwood. A Three Dimensional Calculus Model for Urban Settlement. Highway Research Board, Bull. 347, 1962, pp. 143-146.
10. R.J. Smeed. The Road Space Required for Traffic in Towns. Town Planning Review, Vol. 33, 1963, pp. 270-292.
11. R.J. Smeed. Road Development in Urban Areas, Journal of the Institute of Highway Engineers, Vol. 10, 1963, pp. 5-26.
12. E.S. Mills. Urban Economics. Scott, Foresman, Glenview, IL, 1972.
13. F.R. Reinhart. A Test of Hypothesis Specifically Related to the Cross Sectional Distribution of Population Densities of Cities. Univ. of Minnesota, Minneapolis, unpublished MSc thesis, 1959.
14. J.C. Wilkins. Sherratt's Model and Consumer Travel. Transportation Science, Vol. 3, No. 2, 1969, pp. 93-98.
15. R.J. Vaughan. Decentralization: A Mathematical Explanation. In Traffic Flow and Transportation (G. Newell, ed.). American Elsevier, New York, 1971, pp. 405-412.
16. R.J. Vaughan. Traffic Activity Described in Terms of Some Characteristics of an Urbanized Area. Transportation Research, Vol. 8, 1974, pp. 553-565.
17. R.J. Vaughan. Some Traffic Characteristics of Sydney. Australian Road Research, Vol. 7, No. 4, 1977.
18. D.E. Blumenfeld. Modeling the Joint Distributions of Home and Workplace Location in a City. Transportation Science, Vol. 2, No. 4, 1977.
19. E.S. Mills. Urban Density Functions. Urban Studies, Vol. 7, 1969, pp. 5-20.
20. J.C. Amson. The Dependence of Population Distribution on Location Costs. Environment and Planning, Vol. 4, 1972, pp. 163-181.
21. E. Casetti. Alternate Urban Population Density Models: An Analytical Comparison of Their Validity Range. In Studies in Regional Science (A.J. Scott, ed), Pion Ltd., London, 1969, pp. 105-116.
22. G.T. Papageorgiou. A Theoretical Evaluation of the Existing Population Density Gradient Functions. Economic Geography, Vol. 47, 1971, pp. 21-26.
23. J. Yellin. A Model of Population Distribution, Traffic Congestion and Neighborhood Crowding in a Circular City. Journal of Urban Economics, Vol. 5, 1978, pp. 305-328.
24. A.G. Wilson. Entropy in Urban and Regional Modeling. Pion, Ltd., London, 1970.
25. J.S. Pipkin and D.P. Ballou. A Model of Central City and Suburban Trip Termination Patterns. Journal of Regional Science, Vol. 19, No. 2, 1979, pp. 179-190.
26. J.L. Batty. Urban Density and Entropy Functions. Journal of Cybernetics, Vol. 4, No. 2, 1974, pp. 41-55.
27. M. Ben-Akiva and T. Watanatada. Applications of a Continuous Spatial Choice Logit Model. In Structural Analysis of Discrete Data: With Econometric Applications (C.F. Manski and D. McFadden, eds.), M.I.T. Press, Cambridge, MA, 1981.
28. P.B. Goodwin. Travel Choice and Time Budgets. In Determinants of Travel Choices (D.A. Hensher and Q.M. Dalvis, eds.), Saxon House, Fornborough, England, 1978.
29. M.J.H. Mogridge. Changing Social Patterns in the Journey to Work. Planning and Transport Research and Computation International Company, Ltd., Summer Annual Meeting, Univ. of Warwick, Coventry, Warwickshire, England, 1978.
30. M.J.H. Mogridge. Multimodal Trip Length Distributions. Planning and Transport Research and Computation International Company, Ltd., Summer Annual Meeting, Univ. of Warwick, Coventry, Warwickshire, England, 1978.
31. B. Genest. Population Distribution Functions for Urban Areas. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Res. Rept. R-70-53, 1970.
32. C.E.M. Pearce and others. Urban Density Models. In Traffic Flow and Transportation (G.E. Newell, ed.), American Elsevier, New York, 1972, pp. 439-453.
33. J.F. McDonald and H.W. Bowman. Some Tests of Alternative Urban Population Density Functions. Journal of Urban Economics, Vol. 3, 1976, pp. 242-252.
34. N.J. Clickman and Y. Oguri. Modelling the Urban Land Market: The Case of Japan. Journal of Urban Economics, Vol. 5, 1978, pp. 505-525.
35. A.J. Horowitz. Developable Area and Trip Distribution in Residential Location: A Detroit Case Study. Journal of Regional Science, Vol. 18, No. 3, 1978.
36. D.E. Blumenfeld and G.H. Weiss. Routing in a

Circular City with Two Ring Roads. Transportation Research, Vol. 4, 1970, pp. 235-242.
37. S. Angel and G.M. Hyman. Urban Fields, Pion, Ltd., London, 1976.
38. N.L. Johnson and S. Kotz. Distribution in Statistics. Wiley, New York, 1970.
39. N. Litinas and M. Ben-Akiva. Behavioral Modelling of Continuous Spatial Distributions of Trips, Residential Locations and Workplaces. Center for Transportation Studies, Massachusetts Institute of Technology, Cambridge, working paper CTS-RAMP-79-3, 1979.
40. M.J.H. Mogridge. Some Thoughts on the Economics of Intra-Urban Spatial Location of Homes, Worker Residencies and Workplaces. Urban Economics Conference, Univ. of Keele, Keele, Staffs, England, Paper No. 8, 1973.
41. M.J.H. Mogridge. Land Use and Transport Interaction: An Analysis of the 1966 Journey to Work Data by Mode in London and Its Applications. Planning and Research Computation International Company, Ltd., Univ. of Warwick, Coventry, Warwickshire, England, Paper 8, 1974.
42. P.B. Giannopolos. An Analysis of Trip Length and Land Use Patterns in the Greater Athens Area. Transportation, Vol. 6, 1977, pp. 379-392.
43. C. Clark. Urban Population Densities. Bull., Institute de Statistique, 36, Pt. 4, 1958, pp. 60-68.
44. C. Clark. Transport-Maker and Breaker of Cities. Town Planning Review, Vol. 28, 1959.
45. C. Clark. Population Growth and Land Use. McMillan, London, 1968.
46. R.F. Muth. Cities and Housing. University Press, Chicago, IL, 1969.
47. J.L. Barr. Transportation Costs, Rent and Intraurban Location. Washington Univ., St. Louis, Working Paper, 1970.
48. E.S. Mills. Studies in the Structure of Urban Economy. Johns Hopkins Press, Baltimore, MD, 1972.
49. J.C. Tanner. Factors Affecting the Amount of Travel. Department of Scientific and Industrial Research, Road Research Laboratory, London, Tech. Paper 51, 1961.
50. C.E.M. Pearce. Locating Concentric Ring Roads in a City. Transportation Science, Vol. 8, 1974.
51. J.Q. Stewart and W. Warntz. Physics of Population Distribution. Journal of Regional Science, Vol. 1, 1958, pp. 93-123.
52. W. Warntz. Macroscopic Analysis of Some Patterns of the Geographical Distribution of Population in the United States, 1790-1950. In Qualitative Geography Part 1 (W.L. Garrison and D.F. Marble, eds.), Northwestern Univ., Evanston, IL, Studies in Geography, No. 13, 1967, pp. 191-218.
53. L. Evans and others. Multivariate Analysis of Traffic Factors Related to Fuel Consumption in Urban Driving. Transportation Science, Vol. 10, No. 2, 1976.
54. L. Evans and R. Herman. A Simplified Approach to Calculations of Fuel Consumption in Urban Traffic Systems. Traffic Engineering and Control, Vol. 17, Nos. 8 and 9, 1976.
55. L. Evans and R. Herman. Automobile Fuel Economy on Fixed Urban Driving Schedules. Transporation Science, Vol. 12, No. 2, 1978.
56. M.-F. Chang and others. Gasoline Consumption in Urban Traffic, TRB, Transportation Research Record 599, 1976, pp. 25-30.
57. M.-F. Chang and R. Herman. An Attempt to Characterize Traffic in Metropolitan Areas. Transportation Science, Vol. 12, No. 1, 1978.
58. C.A. Amann and others. Fuel Consumption in the Passenger Car System. General Motors Corporation, Warren, MI, Res. Publication GMR-1632, 1975.

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

Optimal Bus Scheduling on a Single Route

YOSEF SHEFFI AND MORIHISA SUGIYAMA

This paper develops a method for scheduling bus runs on a single route so as to minimize the total waiting time for patrons. The demand for bus travel is assumed to be time-dependent with a given origin-destination pattern. The problem is formulated as a mathematical program subject to bus capacity and possible other (e.g., service standards) constraints. A dynamic programming procedure is suggested for the solution of this program. Finally, some issues associated with optimization of the schedule under stochastic demand are explored as well.

The conventional wisdom in setting up a schedule for a bus route typically involves supplying enough capacity at the maximum load point, subject to some service standards. This approach is useful when the period for which the schedule (timetable) is set is one in which the arrival rate of patrons is constant, and when this period is long in comparison with the bus roundtrip time on the route. When these conditions are not met (for example, the maximum load point may not be stationary or the peak period may be shorter than a bus trip), the schedule may not be optimal in the sense that unnecessary crowding may exist on some buses and slack capacity may exist on others. Better scheduling may bring about a higher level of service and increased productivity.

The problem referred to in this paper is that of preparing a desirable schedule for a given route, not of scheduling the actual buses to the runs. In other words, bus availability and interlining considerations are not taken into account. The inputs to the schedule preparation problem discussed here are the route geometry (including stops and speeds between stops), the total number of bus runs to be operated, and the desired trip rates (i.e., the demand pattern). The output is the route's schedule.

The objective of the schedule preparation is to find the timetable that would give passengers the maximum level of service for a given level of resources. The level of service is expressed in terms of total waiting time that should be minimized, given a fixed number of bus runs.

The practice in the industry, which is also reflected in the literature, is to assume that the demand rate is constant throughout a given design period (e.g., the morning peak). The problem is then only to find the optimal frequency with which to operate the route. Once this frequency is determined, runs are evenly spaced to minimize waiting time. In reality, however, demand rates are not constant but vary over time even within design periods and, in fact, many bus schedules are not evenly spaced for any period. The focus of this paper is on the optimal schedule given a time-dependent demand pattern. The first four sections of this paper consider the demand rates to be deterministic (even though the rates change randomly from day to day). This is due to the planning and optimization perspective of the approach, which focuses on the mean service rates and run times rather than the variability in these measures. The last section attempts to consider some stochastic effects.

A schedule optimization for the simple case of one destination and where the stop dwell time is independent of the number of boardings is analyzed by Newell (1). Newell's approach is used as a starting point to the cases analyzed in this paper. Similar cases, with the inclusion of the bus round-trip constraint, were analyzed by Salzborn (2) and Hurdle (3,4).

The effect of boarding time on the scheduling problem was briefly discussed by Friedman (5) and extensively by the many researchers who looked at the bus bunching phenomenon [e.g., Osuna and Newell (6), Barnett and Kleitman (7), Newell (8), Chapman and Michel (9), and Jordan and Turnquist (10)].

The final section introduces some stochastic considerations and concludes that such modeling should actually be attempted in the context of real time control rather than in the context of scheduling.

This paper does not include numerical examples of the methods shown for reasons of brevity. The interested reader can find complete numerical examples with detailed solutions to all the problems discussed here in Sugiyama (11).

SIMPLE SCHEDULING PROBLEM

This section analyzes a bus route with multiple boardings and one destination point (the final stop). The stops are numbered consecutively i = 0,1,2...,m, and passengers are assumed to board at stops 1 through m-1 and are all destined for m (the starting point, 0, represents a garage or layover point, which does not have to be a part of the route). In this problem it is further assumed that the bus speed is constant where $\Delta t_i, i = 1,2...,m$ represents the bus travel times between stop (i-1) and stop i, and the dwell times are assumed independent of the number of boardings.

The input to this problem includes the demand at each stop expressed in terms of the cumulative arrivals $[F_i(t)]$ during the design period; the number of bus runs available during the design period (T); and the interstop travel times (Δt_i). Following Newell (1), we define the shifted cumulative demand at point 0 as

$$F(t) = \sum_{i=1}^{m-1} F_i\left(t + \sum_{k=1}^{i} \Delta t_k\right) \quad (1)$$

Assume n buses are dispatched from 0 at times $t_j, j = 1,2,...,n$, as shown in Figure 1. Obviously the total waiting time (w) is the shaded area between F(t) and the step function created by $F(t_j)$ for $t_j \leq t \leq t_{j+1}$.

In other words, the total wait time (w) is given by

$$w = \sum_{j=0}^{n-1} \int_{t_j}^{t_{j+1}} [F(t) - F(t_j)]\, dt \quad (2)$$

where $F(t_0) = 0$ and $F(t_n)$ is the total number of boardings along the route ($t_n = T$, the end of the design period).

Following Newell we observe that in order to minimize w, every passenger has to be picked up by the first bus that comes along after his or her arrival at the stop. This, in fact, can be thought of as a necessary condition for a minimum. This observation simplifies any solution procedures significantly because it reduces the feasible space, as shown below.

Newell presents the analytical solution to the minimization of w with no capacity constraints. In our case we are interested in minimizing w subject to the following constraints:

$$F(t_j) - F(t_{j-1}) \leq C \quad \text{for } j = 1, 2, \ldots, n \quad (3a)$$

$$F(t_0) = 0 \quad (3b)$$

and

$$0 < t_1 < t_2 \ldots < t_n = T \quad (3c)$$

where C is the capacity of a bus. The fixed capacity (C) can be readily replaced with C_j where each bus has a different capacity without affecting any of the solution procedures.

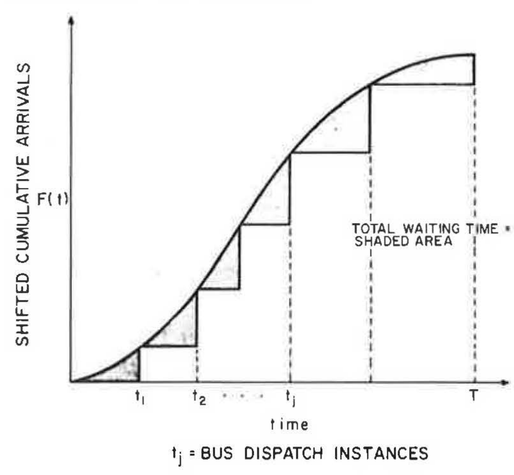
This problem can be solved by a simple dynamic programming (DP) procedure that uses the abovementioned necessary condition, linking the load on each bus with its dispatch time.

The DP procedure can be formulated in several ways for this program. For example, since the dispatch time of the last bus is fixed at $t_n = T$, the procedure can work backwards in time by using the recursive relation:

$$r_j(t_j) = \underset{t_j}{\text{MIN}} \left\{ r_{j+1}(t_{j+1}) + \int_{t_j}^{t_{j+1}} [F(t) - F(t_j)]\, dt \right\} \quad \text{for } j = n-1, n-2, \ldots 1 \quad (4)$$

where $r_j(t_j)$ is the total wait time for all passengers carried by the last (n-j) buses where these buses have been dispatched optimally. The

Figure 1. Shifted cumulative arrivals.



stages of the system thus correspond to the bus dispatch decisions and the states to the dispatch times. The state dynamics are obvious in this formulation as the solution to each stage is the state variable itself. A forward recursion similar to Equation 4 can be formulated also since $F(t_0) = 0$. In any event the recursive relations are subject to the constraint set in Equation 3.

In order to execute the DP procedure, the time line can be divided into discrete parts, say whole minutes. Such a procedure would ignore the integer nature of the number of passengers who use each bus [this discrete nature of the problem is masked by the use of the continuous approximation F(t) to the arrival pattern]. In the case where the number of patrons is small in comparison with the length of the design period and this approximation is inaccurate, the DP procedure can be formulated in terms of loads rather than dispatch times, making the load on the jth bus the state variable. (Alternatively, the time state space, in the original formulation, can be searched in a manner that would ensure the integer nature of the load.) Note, however, that the load and F(t) correspond to the average conditions and therefore the number of passengers who use a particular bus does not have to be an integer on the average.

A simpler solution to the optimal dispatch problem can be obtained if the objective function is to equalize the load factors among buses rather than to minimize the waiting time. In such a case the optimal load on each bus is $s^* = (1/n)\ F(T)$ for all i, assuming that $s^* \leq C$. The schedule in this case can be found graphically or by the simple recursion:

$$t_{i-1} = F^{-1}\,[F(t_i) - s^*] \tag{5}$$

This procedure will not, however, yield the minimum total wait. The minimum wait time schedule, in general, will be such that buses dispatched at a period where the slope of F(t) is high should pick up more passengers than buses dispatched when F(t) is relatively flat. Under the even load policy all buses will be equally loaded and therefore the waiting time will not, in general, be minimized.

SCHEDULING WITH MULTIPLE ORIGINS AND DESTINATIONS

The problem discussed in this section is identical to the optimal scheduling problem, but we assume that patrons can board and alight at any stop along the route. Thus, let P_{ij} represent the fraction of patrons who board at station i who alight at station j. P_{ij} is assumed to be constant over time (during the design period), and

$$\sum_{j=i+1}^{m} P_{ij} = 1 \quad \text{for } i = 1, 2, 3, \ldots m-1 \tag{6}$$

The problem of finding the schedule that minimizes the total travel time is more complicated now because the capacity constraint cannot be formulated in a straightforward manner as in Equation 3a. In order to formulate this constraint let $G_i(t)$ denote the cumulative alightings at stop i by time t, in other words:

$$G_1(t) \equiv 0 \tag{7a}$$

$$G_2(t) = F_1(t - \Delta t_2)P_{12} \tag{7b}$$

$$G_3(t) = F_1(t - \Delta t_2 - \Delta t_3)P_{13} + F_2(t - \Delta t_3)P_{23} \tag{7c}$$

$$\vdots$$

$$G_i(t) = \sum_{l=1}^{i-1} F_l\left(t - \sum_{k=l+1}^{i} \Delta t_k\right)\cdot P_{li} \quad i = 2, 3, \ldots, m \tag{7d}$$

where Δt_i is the bus travel time between stop (i-1) and i and $F_i(t)$ is the cumulative number of arrivals at the ith stop. In many cases $G_i(t)$ can be obtained directly from the measurements (it is easier to measure than P_{ij}) and P_{ij} need not be used [$G_i(t)$ will be used directly as input to the analysis].

Let $N_i(t)$ represent the cumulative number of passengers carried out of the ith station (imagine that we look at the system only at bus departure times). The functions $N_i(t)$ are obtained as follows:

$$N_1(t) = F_1(t)$$
$$N_2(t) = N_1(t - \Delta t_2) + F_2(t) - G_2(t)$$
$$N_3(t) = N_2(t - \Delta t_3) + F_3(t) - G_3(t)$$
$$\vdots$$
$$N_{m-1}(t) = N_{m-2}(t - \Delta t_{m-1}) + F_{m-1}(t) - G_{m-1}(t).$$

Therefore,

$$N_1(t) = F_1(t)$$
$$N_2(t) = F_1(t - \Delta t_2) + F_2(t) - G_2(t)$$
$$\vdots$$
$$N_i(t) = \sum_{k=1}^{i} [F_k(t - \sum_{\ell=k+1}^{i} \Delta t_\ell) - G_k(t - \sum_{\ell=k+1}^{i} \Delta t_\ell)].$$

where $\sum_{\ell=k+1}^{i} \Delta t_\ell \equiv 0$ for $k+1 > 1$ and $F_m(t) = G_1(t) = 0$.

Since the objective function in the problem is expressed in terms of shifted demands, let us shift $N_i(t)$ to the origin as well. Let $N_i'(t)$ denote the shifted cumulative number of passengers on link (i, i+1). It can be calculated as (see Equation 1)

$$N_i'(t) = N_i\left(t + \sum_{k=1}^{i} \Delta t_k\right) \tag{8}$$

and, substituting the last expression for $N_i(t)$, we get

$$N_i'(t) = \sum_{k=1}^{i} \left[F_k\left(t + \sum_{l=1}^{k} \Delta t_l\right) - G_k\left(t + \sum_{l=1}^{k} \Delta t_l\right)\right]$$
$$\text{for } i = 1, 2, 3, \ldots m-1 \tag{9}$$

The minimization problem now is

$$\min w = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} [F(t) - F(t_{j-1})]\, dt \tag{10}$$

subject to

$$N_i'(t_j) - N_i'(t_{j-1}) \leq C \quad \text{for } i = 1, 2, \ldots m-1; j - 1, 2, \ldots, n \tag{11a}$$

$$F(t_0) = 0 \tag{11b}$$

$$0 < t_1 < t_2 < \ldots t_n < T \tag{11c}$$

The constraints Equation 11 parallel Equation 3 with the capacity constraint (Equation 11a) defined by Equation 9.

Up to this point the discussion assumed that the origin-destination (O-D) pattern in terms of P_{ij} is constant over time. The DP formulation above can be easily extended to handle the case where $P_{ij} = P_{ij}(t)$ even though it is difficult to estimate $P_{ij}(t)$. A reasonable approximation may be different P_{ij}'s for different periods, assuming a constant pattern within each period.

The minimum number of bus runs needed to serve the route (n_{min}), can be found by looking at the number of buses needed to carry passengers at the maximum load point; i.e.,

$$n_{min} = \text{INT}\left\{\max_i [N_i'(t)/C]\right\} + 1 \quad (12)$$

where INT[•] is the integer value of the argument and T is the end of the design period.

The multiple O-D scheduling problem can also be solved by using a DP procedure. In a fashion similar to the last section the recursive relations that define the DP procedure are given by Equation 4 and the only difference in the execution is that the constraints (Equation 11a) have to be checked at every stage to ensure feasibility.

CASE WITH STOP DWELL TIMES DEPENDENT ON BOARDING VOLUMES

This section discusses the optimal scheduling program under the assumption that all patrons are destined to one point only (stop m in our scheme) but the dependence of the dwell time on the number of boarding passengers is taken into account. Thus it is assumed that

$$\text{Dwell time} = \Theta \cdot (\text{No. of boardings}) \quad (13)$$

where θ is a constant.

The problem in introducing this relation is that the travel time between departures from consecutive stops is not constant any more. Thus, one cannot construct the shifted demand function F(t), and each $F_i(t)$ has to be dealt with individually.

Assume that buses are dispatched at times t_j, j = 1,2,...n from 0. The jth bus arrives at stop 1 at $t_j + \Delta t_1$ to find $b_{1,j}$ passengers waiting, where

$$b_{1,j} = F_1(t_j + \Delta t_1) - F_1(t_{j-1} + \Delta t_1) \quad (14)$$

and it is delayed for $\theta \cdot b_{1,j}$ min. At stop 2 this bus picks up $b_{2,j}$ passengers, where

$$b_{2,j} = F_2(t_j + \Delta t_1 + \Theta b_{1,j} + \Delta t_2) - F_2(t_{j-2} + \Delta t_1 + \Theta b_{1,j-1} + \Delta t_2) \quad (15)$$

Let $t_{i,j}$ denote the time that the jth bus departs from the ith station. In the context of our problem this time is given by

$$t_{i,j} = t_j + \sum_{k=1}^{i} \Delta t_k + \Theta \sum_{k=1}^{i-1} b_{k,j} \quad (16a)$$

This time can be computed recursively for the jth bus, for stops i=1,2,...m by using

$$t_{i+1,j} = t_{i,j} + \Delta t_{i+1} + \Theta b_{i,j} \quad (16b)$$

The number of passengers who board the jth bus at the ith station ($b_{i,j}$) is given by the difference between the cumulative boardings between the jth and (j-1)th bus at the ith station. In other words,

$$b_{i,j} = F_i(t_{i,j}) - F_i(t_{i,j-1}) \quad (17)$$

and the capacity constraints can be now formulated as

$$\sum_i b_{i,j} \leq C \quad \text{for } j = 1, 2, \ldots, n \quad (18)$$

Given the capacity constraint, the objective functions can be formulated by using the necessary optimality condition mentioned earlier. Since each bus should pick up all the waiting passengers, the total waiting time for the patrons picked up by bus j at stop i ($w_{i,j}$) is given by

$$w_{i,j} = \int_{t_{i,j}}^{t_{i,j+1}} [F_i(t) - F_i(t_{i,j})]\, dt \quad (19)$$

In order to facilitate the presentation of the objective function, let us assume (without loss of generality) that the design period is defined such that there are no arrivals at its beginning. Accordingly, let $t_{i,o}$ define the time where $F_i(t_{i,o})$ = 0. By using these notations, the total waiting time for all buses at the ith stop ($w_i = \sum_{j=0}^{n-1} w_{i,j}$) and the total waiting time ($w = \sum_{i=1}^{m-1} w_i$). In other words,

$$w = \sum_{i=1}^{m-1} \sum_{j=0}^{n-1} \int_{t_{i,j}}^{t_{i,j+1}} [F_i(t) - F_i(t_{i,j})]\, dt \quad (20)$$

Again, the problem of minimizing Expression 20 subject to the capacity constraint (Equation 18) can be solved by using dynamic programming. The stages in the DP procedure would correspond to the dispatched buses and the states to the dispatch times. The basic (forward) DP recursion is given by

$$F_i(t_j) = \text{MIN}_{t_j} [r_{j-1}(t_{j-1})] + \sum_{i=1}^{m-1} \int_{t_{i,j-1}}^{t_{i,j}} [F(t) - F(t_{i,j-1})]\, dt \quad (21)$$

A detailed numerical example of this procedure is given by Sugiyama (11), which outlines all the computational details.

Note that the approximate procedure can be very beneficial for this problem. The approximation consists of scheduling the buses based on the requirement that each bus should carry the same load. If this guideline is adopted for the problem considered here, one can follow the procedure outlined below.

Assume that each bus carries s^* passengers where $s = \sum_{i=1}^{m-1} F_i(T)/n$. The total boarding time delay is thus equal to Θs^* (assuming that each bus actually stops at each stop). If we let $\Delta \ell_i$, i= 1,2,...m denote that distance between the (i-1)th and the ith stop, the average speed of all the buses (b) is given by

$$v = \sum_{i=1}^{m} \Delta l_i \Big/ \left(\Theta s^* + \sum_{i=1}^{m} \Delta t_i\right) \quad (22)$$

The constant speed enables one now to shift $F_i(t)$ to the origin to obtain an approximate shifted cumulative demand function F(t); i.e.,

$$F(t) = \sum_{i=1}^{m} F_i\left(t + \sum_{k=1}^{i} \Delta l_k / v\right) \quad (23)$$

The solution can now be obtained by using the recursion $t_{j+1} = F^{-1}[F(t_j) + s^*]$ for j = 1,2,...n-1 where t_o is the time for which F(t) = 0.

Note that by measuring interstop bus travel times as the times between departures (as done in many studies) the ideas in this method may be in use implicitly.

SCHEDULING WITH MULTIPLE O-D AND BOARDING-DEPENDENT DWELL TIMES

This section analyzes the most general case discussed thus far--that of the optimal scheduling for a bus route with a multiple origin-destination demand pattern, assuming that the dwell times are a function of the number of boarding passengers. Again, we consider a bus route where buses are dispatched from 0 to a succession of stops i = 1,2,...m. As before, let Δt_i denote the bus travel time between the (i-1)th and the ith stop, let $F_i(t)$ denote the cumulative arrivals at stop i, and let $t_{i,j}$ denote the time that the jth bus leaves the ith stop.

The dwell time may now be a function of the number of boarding as well as alighting passengers. Let $a_{i,j}$ denote the number of passengers that alight the jth bus at the ith stop and let $d_{i,j}$ denote the dwell time of the jth bus at the ith stop. A common formula for estimating this dwell time is

$$d_{i,j} = \max(\Theta_1 a_{i,j}, \Theta_2 b_{i,j}) + E \quad (24)$$

In this expression $\theta_1 < \theta_2$ if people pay when boarding and $\theta_1 > \theta_2$ if they pay when alighting. E is a constant that represents the time it takes for the bus to pull into and out of a stop (in this paper we assume that each bus stops at all stops and thus E is included in the definition of Δt_i. This assumption holds for the problem discussed in this paper as we focus on average boarding volumes. If the average is zero for some bus stops, they should, of course, be abolished.) The number of alighting patrons can be computed from the cumulative alighting $[G_i(t)]$ as

$$a_{i,j} = G_i(t_{i,j}) - G_i(t_{i,j-1}) \quad (25a)$$

in parallel with Equation 17, which gives the number of boarding passengers. The cumulative alightings $[G_i(t)]$ can be input directly to the analysis or computed from the origin-destination matrix $(P_{i,j})$; i.e.,

$$a_{i,j} = \sum_{k=1}^{i-1} [F_k(t_{k,j}) - F_i(t_{k,j-1})] P_{k,i} \quad (25b)$$

In order to formulate the capacity constraints for the problem under study, note that

$$t_{i,j} = t_{i-1,j} + \Delta t_i + d_{i,j} \quad (26)$$

The number of passengers on the jth bus as it departs the ith station $(N_{i,j})$ is given by

$$N_{i,j} = \sum_{k=1}^{i} (b_{k,j} - a_{k,j}) \quad (27)$$

and the capacity constraint is given by

$$\mathop{\mathrm{MAX}}_{i=1}^{m} \{N_{i,j}\} \leq C \quad \text{for } j = 1, 2, \ldots n \quad (28)$$

The objective function is again formulated in terms of minimization of the total waiting time (w) where

$$w = \sum_{i=1}^{m-1} \sum_{j=0}^{n-1} \int_{t_{i,j}}^{t_{i,j+1}} [f_i(t) - F_i(t_{i,j})]\, dt \quad (29)$$

where $t_{i,j}$ is defined by the recursion (Equation 26). Note that the decision variables in this problem (the states of the dynamic programming) are still the dispatch times (t_j). These times are expressed implicitly in both the constraints and the objective function since

$$t_{i,j} = t_j + \sum_{k=1}^{i} \Delta t_i + \sum_{k=1}^{i-1} d_{k,j} \quad (30)$$

In the last expression $d_{k,j}$ can be expressed in terms of the inputs $[F_i(t)$ and $P_{ij}]$ as (see Equations 24, 25b, and 17)

$$d_{k,j} = \max\left\{\Theta_1 \sum_{l=1}^{k-1} [F_l(t_{l,j}) - F_l(t_{l,j-1})] P_{lk} \,;\, \Theta_2 [F_k(t_{k,j}) - F_k(t_{k,j-1})]\right\} \quad (31)$$

Thus $t_{i,j}$ can be computed recursively given the schedule of the (j-1)th bus and the time when the jth bus left the (i-1)th stop.

The dynamic programming formulation for this problem is similar to the formulation discussed in the previous section. A numerical example under the assumption discussed is given by Sugiyama (11), who shows all the details of the DP procedure.

This deterministic optimization program can be extended to include various constraints on the minimum headway and on the time that certain buses have to visit certain stops. The problem can also be formulated in terms of buses and bus cycles rather than bus runs. This means that considerations of layover times and equipment availability can be factored into the programming of the solution.

It will be more difficult computationally to take into account the possibility of express runs, zone buses, and shortlining strategies in the optimization, even though the formulation may not pose a particular problem. The difficulty is that the number of diminsions of the state space will grow considerably under these conditions, creating a significant computational burden on the DP procedure.

OPTIMAL SCHEDULING WITH STOCHASTIC DEMAND

This section extends some aspects of the deterministic formulation of the optimal scheduling problem to include stochastic (yet time dependent) elements. The assumption is that the input demand functions are random variables that are distributed according to a Poisson probability law with time-dependent parameter $[\lambda(t)]$.

The total waiting time under these conditions (W) is a random variable and it is natural to choose the expectation of the waiting time, E[W], as the objective function to be minimized. This expectation can be decomposed as follows:

$$E[W] = E[W|A]\, \Pr(A) + E[W|\bar{A}] \cdot \Pr[\bar{A}] \quad (32)$$

where $\bar{A}$ is the event that every passenger boards the first bus that he or she sees and $\bar{A}$ is the complementary event, that some passengers are left behind at some point. The latter event is difficult to handle as it includes, for example, the case where some passenger cannot board any bus and the wait time is infinity. In general, however, the following relations hold:

$$E[W|A] < E[W|\bar{A}] \quad (33a)$$

and

$$\Pr[A] + \Pr[\bar{A}] = 1 \quad (33b)$$

Therefore, in order to minimize E[W] one can maximize Pr[A]; i.e., maximize the probability that every passenger boards the first bus. This objective function will not be capacity constrained as the capacity is included in the objective function.

Let us now analyze the simple scheduling case discussed earlier with the added assumption that the passengers' arrival rate to the ith stop follows a Poisson process with parameter $\lambda_i(t)$. This rate can be shifted to the origin stop by transforming the time by using $\lambda_i(t = \sum_{k=1}^{i} \Delta t_k)$ where Δt_k is the (deterministic) bus travel time between stop (k-1) and k. Furthermore, since arrivals at each stop are independent, the total (shifted) passenger arrivals can be described by a Poisson process with parameter $\lambda(t)$ where

$$\lambda(t) = \sum_{i=1}^{m-1} \lambda_i \left(t + \sum_{k=1}^{i} \Delta t_k\right) \quad (34)$$

If buses are dispatched at times t_j, j = 1,2,...n, the objective function is given by

$$\Pr(A) = \left[\bigcap_{j=1}^{n} A(t_{j-1}, t_j) < C\right] \quad (35a)$$

where $A(t_{j-1},t_j)$ is the number of passengers who arrive between t_{j-1} and t_j, and t_o is the beginning of the design period. These variables are mutually independent and, therefore,

$$\Pr(A) = \prod_{j=1}^{n} \Pr[N(t_{j-1}, t_j) < C] \quad (35b)$$

The distribution of the random variable $N(t_{j-1},t_j)$ is Poisson and thus the objective function can be expressed as

$$\Pr(A) = \prod_{j=1}^{n} \sum_{k=0}^{C} (1/k!)\left[\int_{t_{j-1}}^{t_j} \lambda(t)dt\right]^k \cdot \exp\left[\int_{t_j}^{t_{j-1}} \lambda(t)dt\right] \quad (35c)$$

Alternatively, one can maximize the logarithm of this objective function,

$$\log \Pr(A) = \sum_{j=1}^{n} \log \sum_{k=0}^{C} (1/k!)\left[\int_{t_{j-1}}^{t_j} \lambda(t)dt\right]^k \cdot \exp\left[\int_{t_j}^{t_{j-1}} \lambda(t)dt\right] \quad (36)$$

This objective function can be maximized by using a DP procedure with the recursive relations:

$$r_j(t_j) = \underset{t_j}{\mathrm{MAX}} \left\{ r_{j+1}(t_{j+1}) + \log \sum_{k=0}^{C} (1/k!)\left[\int_{t_{j-1}}^{t_j} \lambda(t)dt\right]^k \right.$$

$$\left. x \exp\left[\int_{t_j}^{t_{j-1}} \lambda(t)dt\right] \right\} \quad j = n-1, n-2, \ldots 1 \quad (37)$$

where $t_k = T$. The possible dispatch instances (t_j) can be constrained by minimum and maximum headway policies to narrow the state space. Computational speed can be enhanced by using an efficient approximation for the integral.

An alternative solution method can be constructed by differentiating log Pr (A) in Equation 36 with respect to t_j, which results in the set equations:

$$\sum_{k=0}^{C} (1/k!) \left\{ \left[\int_{t_j}^{t_{j-1}} \lambda(t)dt\right]^{k-C} - \left[\int_{t_{j-1}}^{t_j} \lambda(t)dt\right]^{k-C} \right\} = 0$$

$$j = 1, 2, \ldots n-1 \quad (38)$$

This set of n-1 simultaneous equations in n-1 unknowns can be solved by Newton's method or any other algorithm for solving a set of simultaneous equations.

In parallel with the extension of the simple scheduling problem, one can extend the simple stochastic case discussed here to include a multiple O-D flow pattern and dwell times that depend on the number of boarding and alighting patrons. The extension to multiple O-D pairs is straightforward, by using the ideas expressed earlier in this paper. The other extension is somewhat more complicated analytically because some of the independence properties used in the simple problems are lost. Sugiyama (11) formulates some of these problems and suggests some solution algorithms.

Use of a more sophisticated stochastic formulation may not be very interesting because a solution to the problem when all inputs are both stochastic and time dependent does not seem feasible. The deterministic approach, however, may provide some good schedules that may actually be applied and modified, when necessary, by real time controls that are designed to contend with the randomness of transit operations.

CONCLUSION

This paper dealt with optimal scheduling of bus runs on a route. The inputs to the analysis include the passenger arrival rate as a function of time at each step, the run times between the stops, bus capacity, the O-D trip pattern (or the alighting rate at each stop), and some (estimated) parameters related to the relations between the dwell time and the number of boarding and alighting patrons. Once the problem is formulated, the optimal schedule can be found by using a dynamic programming procedure.

In its basic formulation, the solution is subject to simple capacity constraints only, but many other constraints such as service standards and nonuniform bus capacities can be incorporated easily within this framework. Further extensions were suggested.

The last section looked at some of the issues associated with modeling the arrival process as a stochastic phenomenon. Here the formulation takes on a somewhat different form as, rather than minimizing the total waiting time, the objective is to maximize the probability that each patron boards the first bus. The stochastic case is more difficult to generalize and it is concluded that deterministic methods may be the most suitable for the optimal scheduling problem.

The suggested procedures can be applied repeatedly to a given route so that a planner may trace a trade-off curve between number of buses and total passenger waiting time. Over a reasonable range of frequencies the demand for bus transport can be assumed to be inelastic to the waiting time or crowding. The abovementioned trade-off curve can thus be used for policy analysis and the setting of service standards.

In closure, we should add that the collection of detailed demand data as required by this method is an expensive and time-consuming task. Recent advances in automatic vehicle monitoring systems, however, overcome this problem. Many of these systems provide continuous information on loads and travel times that can be input to an optimization procedure.

ACKNOWLEDGMENT

This paper was revised following the valuable comments of Mark Turnquist of Cornell University.

REFERENCES

1. G.F. Newell. Dispatching Policies for a Transportation Route. Transportation Science, Vol. 5, No. 1, 1971, pp. 91-105.
2. F.J.M. Salzborn. Optimum Bus Scheduling. Transportation Science, Vol. 6, No. 2, 1972, pp. 137-148.
3. V.F. Hurdle. Minimum Cost Schedules for a Public Transportation Route, I. Theory. Transportation Science, Vol. 7, No. 2, 1973, pp. 109-137.
4. V.F. Hurdle. Minimum Cost Schedules for a Public Transportation Route, II. Examples. Transportation Science, Vol. 7, No. 2, 1973, pp. 138-157.
5. M. Friedman. A Mathematical Programming Model for Optimal Scheduling of Buses' Departures Under Deterministic Conditions. Transportation Research, Vol. 10, No. 2, 1976, pp. 83-90.
6. E.E. Osuna and G.F. Newell. Control Strategies for an Idealized Public Transportation System. Transportation Science, Vol. 6, No. 1, 1972, pp. 52-72.
7. A. Barnett and D.J. Kleitman. Optimal Scheduling Policies for Some Simple Transportation Systems. Transportation Science, Vol. 7, No. 1, 1973, pp. 85-99.
8. G.F. Newell. Control of Pairing of Vehicles on a Public Transportation Route, Two Vehicles, One

Control Point. Transportation Science, Vol. 8, No. 3, 1974, pp. 248-264.
9. R.A. Chapman and J.F. Michel. Modelling the Tendency of Buses to Form Pairs. Transportation Science, Vol. 12, No. 2, 1978, pp. 165-175.
10. W.C. Jordan and M.A. Turnquist. Zone Scheduling of Bus Routes to Improve Service Reliability. Transportation Science, Vol. 13, No. 3, 1979, pp. 242-268.
11. M. Sugiyama. Optimal Bus Scheduling on a Single Route. M.S. Thesis, Dept. of Civil Engineering, Massachusetts Institute of Technology, Cambridge, MA, 1981.

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