TRANSPORTATION RESEARCH RECORD 976

Travel Model Development and Operations Research Methods

TRB
TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1984

Transportation Research Record 976
Price: $8.60
Editor: Naomi Kassabian
Compositor: Lucinda Reeder

modes
1 highway transportation
2 public transit

subject areas
12 planning
13 forecasting

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

Printed in the United States of America

Library of Congress Cataloging in Publication Data
National Research Council. Transportation Research Board.
Travel model development and operations research methods.

(Transportation research record; 976)
1. Traffic engineering–Mathematical models–Congresses. 2. Transportation–Planning–Congresses. I. Series.
TE7.H5 no. 976 38.5 s 85-246 [HE322]
[388.4'0724] ISBN 0-309-03758-1 ISSN 0361-1981

Contents

Authors of the Papers in This Record

Axhausen, Kay W., Department of Civil and Environmental Engineering, 1415 Johnson Drive, University of Wisconsin-Madison, Madison, Wisc. 53706
Hendrickson, Chris, Department of Civil Engineering, Carnegie-Mellon University, Pittsburgh, Pa. 15213
Horowitz, Joel L., Departments of Geography and Economics, University of Iowa, Iowa City, Iowa 52242
McNeil, Sue, 12 DeVite Trail, Hopatcong, N.J. 07843
Miller, Eric J., Department of Civil Engineering, University of Toronto, Toronto, Ontario M5S 1E4, Canada
Morris, Daniel E., Texas Transportation Institute, Texas A&M University System, College Station, Tex. 77843
O'Kelly, Morton E., Department of Geography, Ohio State University, Columbus, Ohio 43210
Parody, Thomas E., Charles River Associates, Inc., 200 Clarendon Street, Boston, Mass. 02116
Rutherford, G. Scott, Washington State Transportation Center, 121 More Hall, FX-10, University of Washington, Seattle, Wash. 98195
Smith, Robert L., Jr., Department of Civil and Environmental Engineering, 1415 Johnson Drive, University of Wisconsin-Madison, Madison, Wisc. 53706
Stokes, Robert W., Texas Transportation Institute, Texas A&M University System, College Station, Tex. 77843
Strombom, Cathy J., Parsons Brinckerhoff, Dexter Horton Building, Suite 960, 710 Second Avenue, Seattle, Wash. 98104

Testing Disaggregate Travel Demand Models by Comparing Predicted and Observed Market Shares

JOEL L. HOROWITZ

ABSTRACT

An intuitively appealing and popular method for testing a disaggregate choice model of travel demand, such as a logit model, consists of comparing the model's predictions of the market shares of travel alternatives in population groups with observations of these shares. Excessively large differences between the predicted and observed shares indicate that the model being tested is incorrect. In current practice, the decision whether differences between predictions and observations are large is made judgmentally, thereby raising the possibility that a correct or approximately correct model will be rejected because of the effects of random sampling errors. A statistical test is described that enables one to distinguish between the effects of random sampling errors and those of true model errors when predicted and observed market shares are compared. Five easily programmable steps for implementing the test are given, and commercially available software that can help with the computations is identified. A numerical example of the application of the test is presented, and the role of the test in practical model development is discussed.

An intuitively appealing way of testing any model of travel behavior is to compare its predictions with actual observations. In the case of disaggregate choice models, such as logit and probit models, that predict individuals' choices among sets of discrete alternatives, this approach to testing often consists of comparing predictions of market shares of alternatives in population groups with observations of the actual market shares in the same groups. Large differences between predicted and observed shares constitute grounds for rejecting the model. For example, to test a model of mode choice the population of interest might be grouped according to characteristics such as income, automobile ownership, location of residence or work, and so forth. The model's predictions of the proportions of individuals in each group that use each mode would be compared with the observed proportions. The model would be rejected as incorrect if the differences between the predicted and observed proportions were excessively large.

In current practice, the decision whether differences between predictions and observations are excessive is made judgmentally. This is unsatisfactory because differences between predicted and observed market shares are subject to random sampling errors. These errors are not relevant to the question of whether the model under consideration is correct. However, depending on the details of the model, selection of population groups, and size of the data set being used, they can produce differences between predictions and observations that are large by reasonable judgmental standards, even if the model being tested is correct. In other words, when predictions and observations are compared judgmentally, random sampling errors may cause a correct model to be rejected.

To minimize the likelihood of this undesirable outcome, it is necessary to have a method for distinguishing between random sampling errors and true model errors in comparisons of predicted and observed market shares. In the terminology of statistics, it is necessary to have a test of the statistical significance of differences between predictions and observations. The main objectives of this paper are to describe such a test and to present a numerical example illustrating its use. Subsidiary objectives are to discuss briefly two important questions relating to the use of the test. These are as follows:

1. In carrying out the test, should the data used for testing (i.e., for computing predicted and observed market shares) be independent of the data used for model estimation or should all of the available data be used for both estimation and testing?
2. The test based on comparisons of predicted and observed market shares is only one of several procedures that are available for testing disaggregate choice models. How should one choose among these procedures in practical, empirical work?

The remainder of this paper is organized as follows. The test statistic is described in the next section and the first question posed previously is answered. Then the numerical example of the use of the test statistic is presented. In the final section the second question posed previously is answered and some concluding comments are presented.

THE TEST STATISTIC

The most frequently used form of disaggregate choice model is the multinomial logit model with a linear-in-parameters utility function. Therefore, to minimize the complexity of the discussion, it will be assumed here that the model being tested has this form. The test statistic for a general choice model is given by Horowitz (1).

In the multinomial logit model with a linear-in-parameters utility function, the probability $P(i|m,\theta)$ that individual m chooses alternative i from a set of I available alternatives is (2,3)

$$P(i|m,\theta) = (\exp X'_{mi}\,\theta)/\sum_k (\exp X'_{mk}\,\theta) \qquad (1)$$

where X'_{mk} (k = 1, ..., I) is a row vector of explanatory variables evaluated for individual m and alternative k, θ is a column vector of constant parameters, and the sum in the denominator is over all available alternatives. In practice, the values of the parameters θ usually are not known a priori and must be estimated by fitting the model to data. In accordance with usual practice, it is assumed in this paper that θ is estimated by the method of

maximum likelihood using a disaggregate data set consisting of observations of the choices and X-values corresponding to M randomly selected individuals. This data set will be called the estimation data set. Let $\hat{\theta}$ denote the estimated values of the parameters θ. Then the estimated probability that individual m chooses alternative i is $P(i|m,\hat{\theta})$.

Derivation

Suppose that the model is to be tested by using a test data set consisting of observations of the choices and X-values corresponding to N randomly selected individuals. The test data set is assumed to be either the same as the estimation data set (in which case N = M) or independent of it. Let the individuals in the test data set be organized into J mutually exclusive groups (J > 1) either randomly or according to the values of characteristics such as income, automobile ownership, location of residence or work, and so on and let N_j denote the number of individuals assigned to group j (j = 1, ..., J). Let $z_{ni} = 1$ if individual n in the test data set chose alternative i and let $z_{ni} = 0$ otherwise. Then the observed market share of alternative i in population group j of the test data set ($\hat{Q}_{ij}$) is

$$\hat{Q}_{ij} = \sum_{\substack{n \text{ in} \\ \text{group } j}} (z_{ni}/N_j) \tag{2}$$

The predicted market share of alternative i in population group j according to the estimated model ($\hat{P}_{ij}$) is

$$\hat{P}_{ij} = \sum_{\substack{n \text{ in} \\ \text{group } j}} P(i|n,\hat{\theta})/N_j \tag{3}$$

The difference between the predicted and observed shares ($\hat{D}_{ij}$) is $\hat{Q}_{ij} - \hat{P}_{ij}$ or

$$\hat{D}_{ij} = \sum_{\substack{n \text{ in} \\ \text{group } j}} [z_{ni} - P(i|n,\hat{\theta})]/N_j \tag{4}$$

$\hat{D}_{ij}$ contains two sources of random error that cause it to differ from zero in general, even if the model $P(i|n,\theta)$ with suitably chosen parameter values is correct. First, the individuals in the test data set are sampled randomly, so their choice indicators z_{ni} are random. Second, the estimated parameter values $\hat{\theta}$ are random because they depend on the choices of the randomly selected individuals in the estimation data set. To develop a test statistic that enables one to distinguish between the effects on $\hat{D}_{ij}$ of random sampling errors and those of true errors in the model, it is necessary to know how large the sampling errors might cause $\hat{D}_{ij}$ to be if the model being tested is in fact correct. [Throughout this paper, the term "large" refers to both large positive and large negative values of $\hat{D}_{ij}$.] This in turn requires knowledge of the probability distribution of the random variable $\hat{D}_{ij}$ for the case of testing a correct model.

To obtain this distribution, let θ_0 denote the true (but unknown) values of the parameters. In other words, $P(i|n,\theta_0)$ is a correct model. Then as discussed by Chernoff (4), $P(i|n,\theta)$ is given approximately by

$$P(i|n,\hat{\theta}) = P(i|n,\theta_0) + \sum_k P(i|n,\theta_0)(X'_{ni} - X'_{nk}) \times P(k|n,\theta_0)(\hat{\theta} - \theta_0) \tag{5}$$

where the sum on the right-hand side is over all of the available alternatives. Substituting Equation 5 into 4 yields

$$\hat{D}_{ij} = \sum_{\substack{n \text{ in} \\ \text{group } j}} \{[z_{ni} - P(i|n,\theta_0)]/N_j\} - K'_{ij}(\hat{\theta} - \theta_0) \tag{6}$$

where K'_{ij} is the row vector defined by

$$K'_{ij} = \sum_{\substack{n \text{ in} \\ \text{group } j}} \sum_k P(i|n,\theta_0)(X'_{ni} - X'_{nk})P(k|n,\theta_0)/N_j \tag{7}$$

The first term on the right-hand side of Equation 6 gives the effect on $\hat{D}_{ij}$ of random variations in the z_{ni}'s. This term is normally distributed with a mean of zero by virtue of the central limit theorem and the mean value of z_{ni} of $P(i|n,\theta_0)$. The second term in the right-hand side of Equation 6 gives the effect on $\hat{D}_{ij}$ of random sampling errors in the estimated parameter values. This term also is normally distributed with mean zero because $\hat{\theta}$ is normally distributed with mean θ_0 when $\hat{\theta}$ is obtained by the method of maximum likelihood (5). [The N_j values and the term K'_{ij} that multiplies $\hat{\theta} - \theta_0$ in Equation 6 can be treated as constants in samples of practical size.] Because sums and differences of normally distributed random variables are normally distributed, it follows that for each i and j, $\hat{D}_{ij}$ has the normal distribution with a mean of zero when the model being tested is correct.

To complete the derivation of the probability distribution of $\hat{D}_{ij}$, it is necessary to calculate the variance of $\hat{D}_{ij}$. Before doing this, however, it is important to observe that there are IJ different values of $\hat{D}_{ij}$, one for each alternative-population group pair. To conclude that the model being tested is incorrect, it is necessary for these IJ values to be excessively large collectively rather than individually. As will be seen, carrying out a test of the sizes of the $\hat{D}_{ij}$'s considered collectively requires knowledge of the covariances of D_{ij}'s corresponding to different alternative-population group pairs as well as knowledge of the variances of individual $\hat{D}_{ij}$'s. Thus it is necessary to obtain the covariance terms $\text{cov}(\hat{D}_{ij},\hat{D}_{rs})$ for each combination of pairs ij and rs.

It can be seen from Equation 6 that each $\hat{D}_{ij}$ contains two parts, one associated with random variations in the z_{ni}'s (i.e., the first term on the right-hand side of Equation 6) and one associated with random variations in $\hat{\theta}$ (i.e., the second term on the right-hand side of Equation 6). Let A_{ijrs} denote the covariance of the first term on the right-hand side of Equation 6 and let B_{ijrs} denote the covariance of the second term. Then as shown by Horowitz (1),

$$A_{ijrs} = \sum_{\substack{n \text{ in} \\ \text{group } j}} [P(i|n,\theta_0)\delta_{ir} - P(i|n,\theta_0)P(r|n,\theta_0)]\delta_{js}/N_j^2 \tag{8}$$

where $\delta_{ir} = 1$ if $i = r$, $\delta_{ir} = 0$ if $i = r$, and δ_{js} is defined similarly. B_{ijrs} is given by (1)

$$B_{ijrs} = K'_{ij} V K_{rs} \tag{9}$$

where V is the covariance matrix of the parameter estimates $\hat{\theta}$. Note that A_{ijrs} is associated with random variations in the z_{ni}'s, whereas B_{ijrs} is associated with random variations in $\hat{\theta}$.

The covariance $\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ can be obtained by combining A_{ijrs} and B_{ijrs} in the appropriate way. Consider first the case of independent estimation and test data sets. In this case the two terms on the right-hand side of Equation 6 are independent (because they are computed using different data sets), so $\hat{D}_{ij}$ is the difference of two independent random variables. Therefore, $\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ is simply the sum of the covariances arising from the independent components of $\hat{D}_{ij}$ and $\hat{D}_{rs}$ (5). In other words,

$$\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs}) = A_{ijrs} + B_{ijrs} \tag{10}$$

when the estimation and test data sets are independent.

Derivation of $\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ for the case in which the estimation and test data sets are the same is more complicated because $\hat{\theta}$ depends on the z_{ni}'s in this case, which causes the two terms on the right-hand side of Equation 6 to be correlated. Horowitz has shown (1) that when the estimation and test data sets are the same, $\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ is the difference between A_{ijrs} and B_{ijrs}. Thus,

$$\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs}) = A_{ijrs} - B_{ijrs} \tag{11}$$

when the estimation and test data sets are the same.

To obtain a statistic that tests whether the $\hat{D}_{ij}$'s considered collectively are larger than can be explained by random sampling errors, it is convenient to organize the $\hat{D}_{ij}$'s into a vector and the covariances $\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ into a matrix. The appropriate vector is given in row vector form by

$$\hat{D}' = (\hat{D}_{11}, \hat{D}_{21}, \ldots, \hat{D}_{I1}, \hat{D}_{12}, \hat{D}_{22}, \ldots, \hat{D}_{I2}, \ldots, \hat{D}_{1J}, \hat{D}_{2J}, \ldots, \hat{D}_{IJ}) \tag{12}$$

Let $\hat{D}$ denote the corresponding column vector. The appropriate matrix is most conveniently defined in terms of submatrices. Define the I x I submatrix S_{js} (j,s = 1, ..., J) by

$$S_{js} = \begin{bmatrix} \mathrm{cov}(\hat{D}_{1j}, \hat{D}_{1s}) & \mathrm{cov}(\hat{D}_{1j}, \hat{D}_{2s}) \ldots \mathrm{cov}(\hat{D}_{1j}, \hat{D}_{Is}) \\ \mathrm{cov}(\hat{D}_{2j}, \hat{D}_{1s}) & \mathrm{cov}(\hat{D}_{2j}, \hat{D}_{2s}) \ldots \mathrm{cov}(\hat{D}_{2j}, \hat{D}_{Is}) \\ & \ldots \\ \mathrm{cov}(\hat{D}_{Ij}, \hat{D}_{1s}) & \mathrm{cov}(\hat{D}_{1j}, \hat{D}_{2s}) \ldots \mathrm{cov}(\hat{D}_{Ij}, \hat{D}_{Is}) \end{bmatrix} \tag{13}$$

Then the desired IJ x IJ covariance matrix S is

$$S = \begin{bmatrix} S_{11} & S_{12} & \ldots & S_{1J} \\ S_{21} & S_{22} & \ldots & S_{2J} \\ & \ldots & & \\ S_{J1} & S_{J2} & \ldots & S_{JJ} \end{bmatrix} \tag{14}$$

This matrix contains all the covariances $\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ organized in a way that is conformable with the vector $\hat{D}$. In matrix notation $E(\hat{D}\hat{D}') = S$. If the estimation and test data sets are independent, the elements of S are given by Equation 10. If the estimation and test data sets are the same, the elements of S are given by Equation 11.

The matrix S is singular (i.e., its determinant is zero) owing to the existence of exact linear relations among the $\hat{D}_{ij}$'s. For example, for any population group j

$$\sum_{i=1}^{I} \hat{D}_{ij} = 0 \tag{15}$$

because the sums over all alternatives of the predicted and observed market shares must both equal 1. Thus, the elements of $\hat{D}$ are normally distributed with mean zero and singular covariance matrix S when the model being tested is correct.

Define the random variable C by

$$C = \hat{D}' S^{-} \hat{D} \tag{16}$$

where S^{-} is any matrix satisfying

$$SS^{-}S = S \tag{17}$$

S^{-} is called a generalized inverse of S and, as will be explained shortly, can be computed from S using standard computer software. C is a collective indicator of the sizes of the $\hat{D}_{ij}$'s. Roughly speaking, if the $\hat{D}_{ij}$'s are large, C is large, and if the $\hat{D}_{ij}$'s are small, C is small. Moreover, C has a known probability distribution when the model being tested is correct. Specifically, C has the chi-square distribution with degrees of freedom equal to the rank of S (6). As a result, C can be used to test whether the differences between the predicted and observed market shares are larger than can be explained by random sampling errors. If the D_{ij}'s differ from zero only because of random sampling errors, with probability $1 - \alpha$ the value of C will be less than the $1 - \alpha$ quantile of the chi-square distribution with degrees of freedom equal to the rank of S. Therefore, if the value of C exceeds the $1 - \alpha$ quantile of this distribution, the model being tested is rejected at the α significance level (i.e., either the model being tested is incorrect or an event of probability α has occurred). By choosing a small enough value of α (values of 0.05 and 0.01 are used frequently in practice), one ensures that there is little likelihood of rejecting a correct model because of the effects of random sampling errors.

Two problems remain to be solved before the test based on the C statistic of Equation 16 (hereafter called the C test) can be implemented in practice. First, the elements of the matrix S depend on the unknown true parameter values θ_0 and on the covariance matrix V of the parameter estimates (see Equations 7-9). A means must be found for estimating these quantities. Second, a method is needed for computing the matrix S^{-}. The first problem can be solved by using the parameter estimates $\hat{\theta}$ in place of the unknown true values θ_0 in Equations 7 and 8. The matrix V can be estimated by the inverse of the Fisher information matrix of the estimation data set evaluated at the parameter values $\hat{\theta}$ (2,5). This estimate of V is one of the standard outputs of many logit estimation computer programs.

The matrix S^{-} can be obtained in several different ways. One is by using a commercially available software package for computing generalized inverses of matrices. For example, the operator GINV in the SAS procedure MATRIX computes the generalized inverse of a matrix that is supplied as input to the procedure (7). Another way to compute S^{-} is by using the eigenvalues and eigenvectors of S. Let W be the matrix whose columns are the eigenvectors of S and let λ_k (k = 1, ..., IJ) denote the corresponding eigenvalues of S. The λ's and the ma-

trix W can be obtained using commercially available computer software such as the subroutine EIGRS of the International Mathematical and Statistical Library (IMSL) (8) or the operator EIGEN in the SAS MATRIX procedure. For each k define λ_k^* by $\lambda_k^* = 1/\lambda_k$ if $\lambda_k = 0$ and $\lambda_k^* = 0$ otherwise. Then the matrix whose (p,q) element is

$$S^-_{pq} = \sum_k \lambda_k^* W_{pk} W_{qk} \quad (18)$$

is a generalized inverse of S (5). A singular matrix has infinitely many generalized inverses, and not all procedures for computing generalized inverses yield the same results. However, any generalized inverse can be used in carrying out the C test.

Implementation

The results of the preceding discussion can be organized into five steps for comparing predicted and observed market shares by means of a C test. These steps, which can easily be programmed for implementation by computer, are as follows:

1. Use the method of maximum likelihood to estimate the values of the parameters $\hat{\theta}$ of the model to be tested and the covariance matrix V of the parameter estimates.
2. Organize the test data set into the desired population groups. For each alternative i and group j, compute the observed market share $\hat{Q}_{ij}$ (Equation 2), the predicted market share $\hat{P}_{ij}$ (Equation 3), and the difference $\hat{D}_{ij}$ between the observed and predicted shares. Organize the elements $\hat{D}_{ij}$ into the vector D (Equation 12).
3. Using the estimate of V and the estimated parameter values $\hat{\theta}$ in place of θ_0, compute the quantities A_{ijrs}, B_{ijrs}, and $\text{cov}(\hat{D}_{ij},\hat{D}_{rs})$ (Equations 7-9 and either 10 or 11, depending on whether the estimation and test data sets are independent or the same). Organize the quantities $\text{cov}(\hat{D}_{ij},\hat{D}_{rs})$ into the matrix S (Equations 13 and 14). Determine the rank of S. An easy way to do this is by computing the eigenvalues of S using one of the methods described earlier. The rank of S equals the number of nonzero eigenvalues.
4. Compute S^- using one of the methods described earlier.
5. Compute the test statistic C (Equation 16). Reject the model under consideration at the α significance level if the value of C exceeds the $1 - \alpha$ quantile of the chi-square distribution with degrees of freedom equal to the rank of S.

A numerical example illustrating the application of these steps is given later is this paper.

Division of Data

A statistical test can identify an incorrect model with high probability only if the effects of random sampling errors on the test statistic are less than those of errors in the model. In other words, to maximize a test's power (or ability to identify erroneous models) it is necessary to minimize the effects of random sampling errors.

In the case of the C test, Equations 10 and 11 indicate that the effects of random sampling errors are likely to be smaller when the estimation and test data sets are the same than when they are independent. There are two reasons for this. First, recall that the covariance components A_{ijrs} and B_{ijrs} respectively represent the effects on the $\hat{D}_{ij}$'s of random sampling errors in the z_{ni}'s and the parameter estimates $\hat{\theta}$. Equations 10 and 11 show that the latter errors are added to the former when the estimation and test data sets are independent, whereas the latter are subtracted from the former when the estimation and test data sets are the same. This causes the sampling errors in the $\hat{D}_{ij}$'s to be smaller when the estimation and test data sets are the same than when they are independent. Second, use of independent estimation and test data sets means that only part of the available data is used in each of the two stages, estimation and testing, whereas all of the data is used in both stages when the estimation and test data sets are the same. This also makes the random sampling errors in the $\hat{D}_{ij}$'s smaller when the estimation and test data sets are the same than when they are independent.

Because the random sampling errors in the $\hat{D}_{ij}$'s are smaller when the estimation and test data sets are the same than when they are independent, the power of the C test usually is larger when the same data are used for estimation and testing than when independent data sets are used. [A more formal presentation of this argument is given elsewhere (1).] In summary, when a model is tested by comparing predicted and observed market shares in population groups, all the available data should be used for both estimation and testing whenever this is feasible (i.e., whenever the full data set is not too large for use in model estimation). The data should not be divided into separate estimation and test data sets.

A NUMERICAL EXAMPLE

The example consists of testing the logit model of mode choice. The modes are automobile and transit. The data used in the example consist of 500 observations of choices and the relevant explanatory variables. The observations were generated by simulation from a logit mode-choice model specified as in Equation 1 with

$$X'_{mi}\,\theta_0 = -0.0865T_{mi} - (0.24881C_{mi}/Y_m) + 3A_mR_i - 4.5R_i \quad (19)$$

where

- i = automobile or transit,
- T_{mi} = individual m's travel time (min) by mode i,
- C_{mi} = individual m's travel cost (cents) by mode i,
- Y_m = individual m's annual income ($10,000s),
- A_m = number of automobiles owned by individual m's household, and
- R_i = 1 if mode i is automobile and 0 if mode i is transit.

[Those who wish to work through this example in detail may obtain a copy of the data set and the FORTRAN program that generated it from the author on request.]

The logit model that is tested in the example has the specification

$$X'_{mi}\,\theta = \theta_1 T_{mi} + (\theta_2 C_{mi}/Y_m) + \theta_3 R_i \quad (20)$$

This model is incorrect because it does not include the variable A_mR_i. To maximize the probability that the C test will identify the model of Equation 20 as incorrect, the same data set will be used for both estimation and testing.

The C test will now be implemented by carrying out the five steps given in the previous section.

Step 1

The maximum-likelihood estimates of the parameters of the model of Equation 20 and the estimate of the covariance matrix V are given in Table 1. Note that the signs of the time and cost coefficients are consistent with expectation and that the t-statistics of these coefficients are satisfactory. Thus, neither the signs nor the t-statistics suggest that the model is incorrect. The nonsignificance of the estimated coefficient of R_i does not indicate that the model is incorrect because there are no strong a priori expectations concerning the importance of this variable.

Step 2

Two population groups are used in this example: Group 1 consists of individuals whose households own one car, and group 2 consists of individuals whose households own two cars. All individuals in the data set used in the example are from one-car or two-car households. The values of the $\hat{Q}_{ij}$'s, $\hat{P}_{ij}$'s, and $\hat{D}_{ij}$'s are shown in Table 2. Organizing the D_{ij} values into the vector $\hat{D}$ yields

$$\hat{D} = \begin{bmatrix} -0.1124 \\ 0.1124 \\ 0.0942 \\ -0.0942 \end{bmatrix} \tag{21}$$

TABLE 1 Estimation Results for the Logit Model of Equation 20

Parameter	Estimated Value	t-Statistic
θ_1	-0.05056	-2.494
θ_2	-0.2148	-9.382
θ_3	0.3194	1.158

$$V = \begin{bmatrix} 0.4110 & -0.3542 & 4.318 \\ -0.3542 & 0.5242 & -2.388 \\ 4.318 & -2.388 & 76.08 \end{bmatrix} \times 10^{-3}$$

TABLE 2 Values of $\hat{Q}_{ij}$, $\hat{P}_{ij}$, and $\hat{D}_{ij}$ for the Numerical Example

i (Mode)[a]	j (Group)	$\hat{Q}_{ij}$	$\hat{P}_{ij}$	$\hat{D}_{ij}$
1	1	0.3465	0.4588	-0.1123
2	1	0.6535	0.5412	0.1123
1	2	0.6618	0.5676	0.0942
2	2	0.3382	0.4324	-0.0942

[a] Mode 1 is automobile, and mode 2 is transit.

Step 3

The row vectors K_{ij}' defined in Equation 7 are

$$K_{11}' = (-0.9799, 0.4670, 0.09388) \tag{22a}$$

$$K_{21}' = (0.9799, -0.4670, -0.09388) \tag{22b}$$

$$K_{12}' = (-0.8070, 0.2279, 0.08145) \tag{22c}$$

$$K_{22}' = (0.8070, -0.2279, -0.08145) \tag{22d}$$

The covariance terms A_{ijrs}, B_{ijrs}, and $\text{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ are shown in Table 3. Organizing the quantities $\text{cov}(\hat{D}_{ij}, \hat{D}_{rs})$ into the matrix S yields

$$S = \begin{bmatrix} 0.2033 & -0.2033 & -0.1704 & 0.1704 \\ -0.2033 & 0.2033 & 0.1704 & -0.1704 \\ -0.1704 & 0.1704 & 0.1429 & -0.1429 \\ 0.1704 & -0.1704 & -0.1429 & 0.1429 \end{bmatrix} \times 10^{-3} \tag{23}$$

The eigenvalues of S were computed using the subroutine EIGRS of the IMSL and are (0, 0, 0, 6.923×10^{-3}). Because there is only one nonzero eigenvalue, the rank of S is 1.

Step 4

The generalized inverse of S as computed from Equation 18 is

$$S^{-} = \begin{bmatrix} 424.2 & -424.2 & -355.6 & 355.6 \\ -424.2 & 424.2 & 355.6 & -355.6 \\ -355.6 & 355.6 & 298.0 & -298.0 \\ 355.6 & -355.6 & -298.0 & 298.0 \end{bmatrix} \tag{24}$$

Step 5

The test statistic C is computed by substituting Equations 21 and 24 into Equation 16. The result is C = 62.15. The 0.95 quantile of the chi-square distribution with 1 degree of freedom is 3.841. Because the computed value of C exceeds this, the model of Equation 20 is rejected as incorrect at the 0.05 significance level.

CONCLUSIONS

Testing a disaggregate choice model by comparing predicted and observed market shares has much intuitive appeal, but it is not the only way in which these models can be tested. A large number of test procedures based on likelihood ratio tests, Lagrangian multiplier tests, and the likelihood ratio index goodness-of-fit statistic are also available (9,10). It is worthwhile to consider how one might choose among these tests in practical model development.

An important difference between the C test discussed in this paper and the other tests is that the other tests require the analyst to specify an alternative model (e.g., a logit model with a different specification of the utility function $X_{mi}'\theta$) against which the model under consideration is to be tested. In effect, these tests attempt to determine whether the alternative model fits the available data better than the model under consideration does, in which case the model under consideration is rejected as being incorrect. In contrast, the C test does not require specification of an alternative

TABLE 3 Covariance Terms for the Example

A_{ijrs}

		s = 1		s = 2	
		r=1	r=2	r=1	r=2
j = 1	i = 1	4.117×10^{-4}	-4.117×10^{-4}	0	0
	i = 2	-4.117×10^{-4}	4.117×10^{-4}	0	0
j = 2	i = 1	0	0	2.995×10^{-4}	-2.995×10^{-4}
	i = 2	0	0	-2.995×10^{-4}	2.995×10^{-4}

B_{ijrs}

		s = 1		s = 2	
		r=1	r=2	r=1	r=2
j = 1	i = 1	2.084×10^{-4}	-2.084×10^{-4}	1.704×10^{-4}	-1.704×10^{-4}
	i = 2	-2.084×10^{-4}	2.084×10^{-4}	-1.704×10^{-4}	1.704×10^{-4}
j = 2	i = 1	1.704×10^{-4}	-1.704×10^{-4}	1.566×10^{-4}	-1.566×10^{-4}
	i = 2	-1.704×10^{-4}	1.704×10^{-4}	-1.566×10^{-4}	1.566×10^{-4}

$\mathrm{cov}(\hat{D}_{ij}, \hat{D}_{rs})$

		s = 1		s = 2	
		r=1	r=2	r=1	r=2
j = 1	i = 1	2.033×10^{-4}	-2.033×10^{-4}	-1.704×10^{-4}	1.704×10^{-4}
	i = 2	-2.033×10^{-4}	2.033×10^{-4}	1.704×10^{-4}	-1.704×10^{-4}
j = 2	i = 1	-1.704×10^{-4}	1.704×10^{-4}	1.429×10^{-4}	-1.429×10^{-4}
	i = 2	1.704×10^{-4}	-1.704×10^{-4}	-1.429×10^{-4}	1.429×10^{-4}

model. In effect, it tests the model under consideration against all alternatives simultaneously.

The ability of a test against a specific alternative to identify an erroneous model and the relative power of the C test and a test against a specific alternative depend on the choice of the alternative. As is discussed elsewhere (1,11), a test of an erroneous model against an alternative that is a good approximation to the correct model is likely to be much more powerful than a C test. However, a test against an alternative that is a poor approximation to the correct model can be less powerful than a C test.

These considerations suggest that the C test and tests against specific alternatives are complements, rather than substitutes, and that both types of tests should be carried out during the process of developing empirical models. A practical approach to this begins by formulating several alternatives to the model currently under consideration. The developer of an empirical model can virtually always do this. Although there can be no assurance that any of these alternatives is correct or approximately correct (if there could be such assurance, the tests being discussed here would be unnecessary), one or more of the alternatives may nonetheless provide a powerful test of the current model if the model is seriously erroneous (9). Accordingly, the current model should be tested against the alternatives using likelihood ratio tests or other appropriate procedures (9,10). However, it is not possible to test a model against all reasonable alternatives, and failure to reject the current model in tests against specific alternatives may be the result of a poor choice of alternatives rather than an indication that the current model is correct. Therefore, if the current model is not rejected in tests against specific alternatives, a C test should be carried out. The C test amounts to a test of the current model against all remaining alternatives and may have a higher probability of detecting errors in the current model than do the tests against specific alternatives if the specific alternatives are themselves highly erroneous.

ACKNOWLEDGMENT

The work reported in this paper was supported in part by a cooperative agreement between UMTA and the University of Iowa.

REFERENCES

1. J.L. Horowitz. Testing Probabilistic Discrete Choice Models of Travel Demand by Comparing Predicted and Observed Aggregate Choice Shares. Transportation Research, in press (1985).
2. T.A. Domencich and D. McFadden. Urban Travel Demand: A Behavioral Analysis. American Elsevier, New York, 1975.
3. D.A. Hensher and L.W. Johnson. Applied Discrete-Choice Modelling. Croom Helm, London, 1981.
4. H. Chernoff. Large-Sample Theory: Parametric Case. Annals of Mathematical Statistics, Vol. 27, 1956, pp. 1-22.

5. C.R. Rao. Linear Statistical Inference and Its Applications, 2nd ed. Wiley, New York, 1973.
6. C.R. Rao and S.K. Mitra. Generalized Inverse of Matrices and Its Applications. Wiley, New York, 1971.
7. SAS User's Guide: Statistics. SAS Institute, Inc., Cary, N.C., 1982.
8. The IMSL Library. IMSL, Inc., Houston, Texas, 1982.
9. J. Horowitz. Specification Tests for Probabilistic Choice Models. Transportation Research, Vol. 16A, 1982, pp. 383-394.
10. J.L. Horowitz. Evaluation of Usefulness of Two Standard Goodness-of-Fit Indicators for Comparing Non-Nested Random Utility Models. In Transportation Research Record 874, TRB, National Research Council, Washington, D.C., 1982, pp. 19-25.
11. A. Wald. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large. Transactions of the American Mathematical Society, Vol. 54, 1943, pp. 426-487.

Evaluation of Heuristic Transit Network Optimization Algorithms

KAY W. AXHAUSEN and ROBERT L. SMITH, JR.

ABSTRACT

Changes in urban land use and travel demand have created the need to restructure many existing mass transit networks. Heuristic network optimization as one of the available methodologies to improve transit networks is described. The characteristics and results of the algorithms developed in Europe are summarized and a short description of the American algorithms is given. The potential for applying network optimization methodologies in the context of small to medium-sized American cities is evaluated. The review and evaluation of 13 heuristic methodologies revealed a wide range of approaches that are generally theoretically sound, have reasonable potential for generating improved networks, and are computationally and otherwise feasible. Application of an unproven new algorithm by Mandl to the bus network for Madison, Wisconsin, and the light rail network for Duesseldorf, West Germany, showed that a fairly complex heuristic algorithm can be implemented quickly and easily. Mandl's algorithm, however, did not generate an improved network, primarily because the initial computer-generated network does not follow demand. Better results were obtained with two other heuristic methodologies that have been applied to the Duesseldorf network. The Madison and Duesseldorf applications form the basis for recommendations for further improvement of heuristic methodologies.

The bus transit networks that are the predominant form of public transit in American cities have changed only slowly since the elimination of the streetcar in the 1930s and 1940s. Often the major bus lines still run on the same streets that the streetcars used. Because of the major shifts in population and employment that have occurred in recent years, the bus networks in many cities could probably be restructured to serve the existing demand better and reduce operating costs at the same time. Transit managers are often reluctant to make major changes in routes because of the almost certain political opposition by those who think they will receive poorer service. Also, transit managers generally do not have analytical tools readily available to aid them in generating and evaluating alternative networks. As the result of the current fiscal crisis in transit, transit managers should be more interested in methodologies for restructuring their bus networks.

Chua and Silcock (1) identify six methodologies for transit network restructuring and optimization: manual approach using service standards and guidelines, systems analysis using standard travel demand and trip assignment models, market analysis using manual trip assignment for corridors or small service areas, systems analysis with interactive graphics, heuristic procedures, and mathematical optimization. The first three methodologies are limited by the number of alternative networks that can be evaluated in a reasonable amount of time. By adding interactive graphics to systems analysis, network development and evaluation are greatly enhanced. Many more networks can be tested in much less time. The methodology, however, tends to be biased toward the existing network, so unconventional solutions may not be examined. Furthermore, there is no guarantee that solutions near the optimum will be found.

In contrast, mathematical optimization using linear programming or general integer programming will produce an optimal network within the specified

constraints and should not be biased toward the existing network. Mathematical optimization, however, is limited by the computational requirements to relatively small networks. Even with the recent advances in the speed and memory size of computers, networks are limited to about 70 or 80 nodes, which results in a coarse network for bus systems in larger urban areas. A network of 70 nodes may be adequate for rail systems in larger urban areas or bus systems in small to medium-sized urban areas.

Heuristic methodologies bridge the gap between systems analysis with interactive graphics and the mathematical optimization methodologies. The heuristic methodologies utilize systematic procedures to generate and improve transit networks. The complexity of the overall problem is reduced by breaking it into manageable components. Within each component a good and sometimes optimal solution is obtained. The complexity and computational requirements are further reduced by limiting the amount of interaction among the components. Because the heuristic networks are machine generated, many more networks can be evaluated and they are less likely to be biased by the existing networks. Although the heuristic methodologies do not guarantee an optimum network, the starting conditions and other parameters can be varied to increase the chances that the true optimum that would be obtained from mathematical optimization is included in the range of networks considered. All the network evaluation procedures are constrained by the accuracy of the demand estimates and the simplification of the complexities of the transit network as it exists in a dynamic real world. Thus, even mathematical optimization procedures will only provide an indication of potential network improvements. Good heuristic methodologies will provide similar directions for network improvements.

In the past two decades a number of heuristic network optimization methodologies have been developed and applied in European cities to estimate possible means of restructuring both bus and light rail networks. In contrast, in the United States only one heuristic approach to transit network restructuring has actually been tested (2) and none of the European methodologies have been tested here. It must be noted, however, that there are no reports in the literature of the results of implementing the network improvements and comparing the predicted with the actual network performance. The overall purpose of this paper is to evaluate the potential for applying heuristic network optimization methodologies to improve transit networks in small to medium-sized American cities. The evaluation is limited to small to medium-sized cities because realistic networks for large cities would require prohibitive amounts of computer time. In achieving the overall purpose first, the literature on heuristic methodologies is reviewed and the available algorithms are analyzed in terms of their inherent potential for generating improved networks. The performance of these algorithms that have been applied to actual transit networks is also analyzed. Next, one of the available heuristic algorithms is selected for testing on an American transit network. The results of the American application are documented and evaluated. The same heuristic algorithm is also applied to a European transit network for which the results of the application of two other heuristic algorithms are available. The comparison of the performance of the heuristic algorithm in the two cities and with the two other algorithms provides the basis for recommendations for additional research on heuristic algorithms.

REVIEW OF THE LITERATURE

American Literature

In the United States work on the transit network restructuring and optimization problem has focused on the application of systems analysis with interactive graphics. Most of the work has used Rapp's Interactive Graphics Transit Design System (IGTDS) or the more recent Interactive Graphics Transit Network Optimization System (TNOP) (3-6). An enhanced version of TNOP has been applied to transit network development in Washington, D.C.; Jacksonville; Baltimore; and Buffalo (7).

The American experience with heuristic network optimization is limited to research by Rea (8), Sharp (2), and Hsu and Surti (9-11). Rea's service specification model assigns generalized modes to appropriate links by using a small base network, a fixed demand, and a link service level function in which headways (and the resulting wait time) are a function of link volumes. Rea's algorithm uses an iterative procedure in which a minimum-time-path assignment is followed by adjustment of link service levels to correspond with link volumes. Even for the small test network, convergence to equilibrium conditions was sometimes a problem.

Sharp's iterative route-structuring algorithm is formulated as a multicommodity transshipment problem in which each commodity is represented by unique travel demands for each origin and destination node pair. The objective is to minimize the sum of passenger travel and delay time costs and vehicle amoritization and operating costs while satisfying the travel demands. The algorithm was applied to the Columbus, Georgia, bus network. The improved bus network provided a 5 percent reduction in trip times for the base ridership and generated a 9 percent increase in ridership while increasing vehicle costs by only 3 percent.

Hsu and Surti's decomposition approach to bus network design uses a minimum-time-path algorithm to identify an initial set of routes between manually identified route origins and destinations. Incremental changes in route alignments are accepted if route ridership is increased. In the application to a 59-node bus network for a portion of the Denver urban area, the changes in route alignment were made manually; however, the algorithm to select nodes in the vicinity of the shortest path in searching for improved routes no doubt could be computerized. The model is limited to providing local optima within the corridors defined by the initial route specification. Evaluation of alternative combinations of routes requires manual specification of those combinations.

European Literature

The European research on heuristic network optimization includes 10 studies ranging from Nebelung's in 1961 to Sahling's in 1981 (12-26). Because almost all of the studies are the result of consulting work, there may be more, yet unpublished algorithms. Except for Hasselstroem's approach, which is part of the Volvo Corporation's transportation planning package, none of the approaches appears to have been applied more than once.

Rather than review each of the 10 studies in detail, the key features of the algorithms are outlined and compared in terms of a three-step overall procedure that is common to nearly all of the algorithms (see Tables 1 and 2). Only two of the algo-

TABLE 1 Characteristics of Early European Heuristic Network Optimization Algorithms

Author:	Nebelung (1961)	Lampkin (1967)	Silman (1974)	Hoidn (1975)	Rosello (1976)
Step 1--Base Network Construction					
Designed for:	LRT (Radial)	Bus	Bus (Radial)	Bus/LRT (Radial)	Bus
Nodes are:	Intersections	Zones	Zones	Zones	Zones
Networks Based on:	Terminals	Skeletons	Skeletons	Terminals	Spider[a]
Step 2--Initial Line Development and Selection					
Line Constraints:	1) Pair selected terminals 2) Select complete lines	All feasible skeletons	All skeletons	All lines to center	Start with link & add nodes
Line Constraints:	-1.2*shortest path -Serve all links	Omit infeasible skeletons	Length	-1.3 * shortest path (dist) -1.4*node-to-node min. path	Max. no. of routes
Line Objectives:	Max. direct trips	-Psgr. miles -line length -no. of nodes -intersecting lines	--	--	Avg. total cost
Step 3--Network Development/Improvement					
Procedure:	None	Add nodes to best skeletons	1) Add nodes to best skeleton to form line 2) Update list of best skeletons 3) Go to 1	1) Select complete lines 2) Delete demand as lines are added	-Add nodes -Delete nodes and lines of 2 nodes
Objectives:	--	See Step 2	Psgr. minute/ route length	Psgr.-miles weighted by directness of routing & use of line	Avg. total cost/psgr.
Constraints:	--	Increase in travel time max. of 50%	Max. of 5 lines added per run Max. of 100 best skeletons	--	--
Frequency Optimization:	No	External	Yes-user selects lines	No	External
User Involvement:	Extensive (terminal selection)	Initial	Initial & frequency optimization	Initial	No

[a]Connection of logical adjacent nodes to form a web-like network.

rithms lack any of the three steps. The steps follow a logical progression from base network construction in step 1 to initial line development and selection in step 2. In general, a large set of feasible lines is identified in step 2, from which an optimum set is selected in step 3 based on the specified objective function and constraints. In some cases additional lines are generated in step 3. Also, the initial networks may be modified systematically in various ways to generate improvements as measured by the objective function.

All but one of the algorithms use minimum time paths to identify node-to-node paths for creating lines or assigning demand to links or both. Sonntag uses a multipath assignment to create a loaded spider network as the basis for initial line development and selection in step 2. In a spider network adjacent nodes (zones) are connected to form a web-like network. Rosello uses an all-or-nothing assignment to create a loaded spider network. The advantage of loaded spider networks is that no constraints are placed initially on the ultimate pattern of lines, but the number of possible lines is large. Dubois solves the problem in part by using an initial assignment and cost constraints to reduce the size of the base network. The more common means of reducing the number of lines considered is to require that terminals, ring lines, or skeletons be specified. Ring lines specify three nodes--the terminals and an intermediate node--as the starting point for developing a circular or ring line. Skeletons add a second intermediate node as the basis for a linear line.

In step 2 the initial line development procedure is with one exception based on either a loaded spider network or specification of nodes (terminals, etc.). With the spider network an objective function and constraints are applied to the loaded links either in pairs (Sonntag) or incrementally as nodes are added (Rosello). When terminals and so on are specified, the most common approach to generating a feasible set of base lines is to expand the minimum-time-path connections between the terminals to include all lines that are longer than the minimum time path by up to a specified percentage, usually 20 or 30 percent. The idea is to reduce computing times to a manageable level while still providing a range of lines that includes the optimal or near-optimal network. Lampkin and Mandl, however, only consider minimum-time-path lines between terminals

TABLE 2 Characteristics of Recent European Heuristic Network Optimization Algorithms

Author:	Dubois (1977)	Sonntag (1977)	Hasselstroem (1979)	Mandl (1979)	Sahling (1981)
Step 1--Base Network Construction					
Designed for:	Buses	Buses/LRT	Buses	Buses/LRT	Buses/LRT
Nodes are:	Zones	Intersections	Zones	Zones	Intersections
Networks Based on:	Optimal Street system[a]	Multiple-path spider	Terminals for reduced base network	Terminals & Ring lines	Minimum time paths
Step 2--Initial Line Development and Selection					
Line Building:	Shortest path with max. no. of nodes	Two link base lines	All lines	Shortest with max. no. of nodes (total length)	1) Select productive lines 2) Delete served demand 3) Go to 1)
Line Constraints:	-X*shortest path -All nodes connected	-Lines/link ≤ max. lines/link	X*shortest path	-All nodes served -Network connected -One line per terminal	-All demand served -All links served
Line Objectives:	--	Maximize flow/base line	--	--	Psgr.-miles per line mile
Step 3--Network Development/Improvement					
Procedure:	1) Select complete lines 2) Combine lines & delete portions	1) Connect base lines 2) Break-up & re-arrange lines 3) Go to step 1	Linear programming solution	1) Recombine line segments 2) Generate detours 3) Eliminate detours	None
Objectives:	-Psgrs./line -Transfers	-Direct trips -Transfers -Link flows (max)	Reduce transfers	-Total travel time -Transfers network length	
Constraints:	Costs/no. of buses	-Network length -Total deviation from min. path -Serve all links	-Total cost -Line frequency	Number of buses	
Frequency Optimization:	External	No	Simultaneous with line selection	No	No
User Involvement:	No	No	Initial	Initial	Extensive (interactive)

[a]Select set of streets that minimize travel time subject to total investment cost constraint. Heuristic removal algorithm is used.
[b]Select links to minimize the sum of travel times and operating costs subject to budget constraints and satisfying all demands.

so that a relatively small set of bus lines is obtained. Mandl's lines between terminals and ring lines are augmented by additional shortest-path lines that are independent of the terminals if necessary to meet the service constraints. Neither Dubois nor Sahling relies on terminals and the like to constrain the number of lines considered. Instead, the base lines are generated from the minimum time paths directly. Dubois expands his set of base lines by considering all lines within a specified percentage of the shortest path.

In step 3 several different strategies are used to develop a final network from the base network or improve on the base network from step 2 or both. Both Lampkin and Silman add nodes to skeletons incrementally based on the objective of maximizing the amount of service provided (passenger miles or minutes) for each additional node. Silman's objective function is more relevant because the impact of increasing route length is taken into account. Rosello extends node-by-node network construction to include deletion of nodes and lines that have been reduced to only two nodes. The large number of lines generated in step 2 is improved on or eliminated node by node based on average total costs per passenger. Sonntag also operates at a disaggregate level using base lines of only two adjacent links rather than nodes. In step 3 the best base lines from step 2 are connected incrementally to form full base lines, which are then broken up and rearranged incrementally based on a complex objective function.

Both Hoidn and Dubois select complete lines incrementally in step 3. Hoidn stops when all the demand has been served. In contrast, Dubois has a second stage in which lines are combined and segments of lines deleted based on passengers per line and transfers. Similar but not identical procedures for rearranging the base lines are also used by Sonntag and Mandl. Mandl follows a three-level hierarchical procedure in an attempt to minimize total travel time. First the line segments for two lines are recombined. Next nodes are added to generate detours and, finally, less productive detours are eliminated. If an improvement is found at level 2 or 3, the algorithm begins again at level 1.

Rather than rely in step 3 on heuristics that

provide no assurance of an optimal solution, Hasselstroem applies linear programming with the objective of maximization of through trips per vehicle trip, which is equivalent to minimizing transfers. The resulting network is optimal in terms of providing for through trips but only for the set of lines generated in steps 1 and 2. In contrast, step 3 is not used at all by the authors of the earliest and the most recent algorithms. Both Nebelung and Sahling stop with the generation of a feasible set of base lines. Nebelung's primary purpose was to check the quality of manually produced network improvements.

EVALUATION OF THE HEURISTIC ALGORITHMS

The review of the literature on American and European heuristic network optimization algorithms clearly shows that there is a wide variety of approaches to developing improved transit networks. In order to provide a basis for selecting algorithms for application in the United States, the algorithms are evaluated in three ways:

1. Subjective evaluation of the basic procedures using the information presented in the literature review,
2. Comparison of the level of network improvement predicted (different algorithms applied to different cities), and
3. Comparison of predicted network improvement for the same city and transit demand.

The last two comparisons are severely limited by the lack of data on applications, especially applications to the same network. The evaluation is also limited by the lack of published before-and-after studies that would provide a benchmark for validation of the algorithms. The impact of the network improvements on demand, however, could be estimated with an independent modal-choice model as was done by Hasselstroem.

Subjective Evaluation

The results of the subjective evaluation of the American and European algorithms are presented in Table 3. In order to approximate an optimal solution in the mathematical programming sense, heuristics must consider the entire range of possible lines and use an algorithm that gives good if not consistently near-optimal solutions. The potential for an optimal solution can be increased by increasing the range of feasible lines considered. The most common approach is to assume that feasible lines diverge from minimum-time-path lines by a limited amount, say 20 or 30 percent. Most of the algorithms constrain the base minimum-time-path lines by specifying terminals, but Dubois avoids terminals by selecting the shortest paths that have the most nodes. If base lines generated by X times the shortest path connecting terminals are likely to capture the optimal network, Hasselstroem's algorithms will give excellent results because his algorithm uses linear programming to select the optimal network from the given base lines.

Solutions that are independent of the existing network and the planner's preconceived ideas of good solutions are desirable. Such solutions can potentially be obtained from the incremental line generation approaches of Rosello and Sonntag. The use of the shortest paths with the most nodes by Dubois also has some potential, but demand should be considered as well.

TABLE 3 Subjective Evaluation of Heuristic Algorithms

Evaluation Criterion	Best Procedure or Example
Potential for optimal solution	
Wide range of feasible lines	Terminals plus X * shortest path
Independent of existing network	Incremental line generation from spider network [Rosello (*16*), Sonntag (*18, 19*)]
Partial mathematical programming solution	Hasselstroem (*20, 21, 26*)
Theoretical soundness	
Logical steps and internal consistency	All are minimally acceptable
Appropriate objective function and constraints	Rosello, Dubois (*17*), Sonntag, Hasselstroem, Sharp (*2*)
Ease of implementation	
Reasonable computational requirements	Problems with Sonntag, Mandl (*23, 24*), and possibly Hasselstroem
Maximum allowable network size	Sahling (*25*) allows large, detailed networks
Simple algorithm	Sahling, Nebelung (*12*), Silman (*14*), Hoidn (*15*)
User involvement	Nebelung, Silman, Sahling

The most relevant theoretical basis for evaluating the algorithms is provided by modal-choice theory and practice. The variables used in the objective functions and constraints of the algorithms should be consistent with the variables that are important for modal choice. More complicated and relevant objective functions and constraints are used by Rosello, Dubois, Sonntag, Hasselstroem, and Sharp. Modal-choice theory may also provide a basis for evaluating the procedures used in the algorithms. At a minimum the algorithms should be based on logical steps and be internally consistent. All the algorithms are at least minimally acceptable and cannot be rejected initially on procedural grounds.

As shown in Table 3, some of the algorithms are easier to implement than others, but the ease of implementation in general must be traded off against the lack of complexity and possibly more limited potential for obtaining a good solution. Little information is provided in the literature on the computational requirements of the algorithms and how those relate to network size.

Most of the algorithms require some user input, at least initially, in specifying terminals or other parameters. A few require more extensive user involvement. More opportunity for user input at various stages in the algorithms will provide for greater understanding of the mechanics of the solutions and reduce computing times if inferior options can be identified and eliminated or modified at an early stage.

Predicted Network Improvements

The performance of individual algorithms can be measured by comparison of the improved transit network with the base network. Performance measures are available for only 5 of the 10 European algorithms and only 1 of the 3 American algorithms. As shown in Table 4, substantial improvements in network performance are generally predicted. Some trade-offs may be required as shown by the results for Hoidn's algorithm. Network length was reduced by 9 percent at the expense of a 2 percent decrease in direct trips. Similarly, for Sharp's algorithm increase in ridership and decrease in travel time were achieved at the expense of higher total vehicle costs. Rosello achieved substantial reductions in average costs by reducing the number of lines and increasing ridership dramatically. Because the total demand is fixed, the predicted doubling of transit demand does not appear reasonable. The lack of a sophisticated,

TABLE 4 Improvements Predicted Using Heuristic Network Optimization

Source	Performance Measure	Base Network	Improved Network	Change (%)
Lampkin (1967)	No. of buses	88	66	- 25
	Mileage	NA	NA	- 5
	Wage bill	NA	NA	- 13
Sharp (1974)	Base rider travel time	NA	NA	- 5
	Ridership	NA	NA	+ 9
	Total vehicle costs	NA	NA	+ 3
Hoidn (1975)	Network length (km)	43.01	39.20	- 9
	Direct trips (%)	57.14	56.24	- 1.6
	Satisfied demand (%)	92.20	92.20	0
	Objective function for direct trips	529	686	+ 30
	All trips	2,226	2,723	+ 22
Rosello (1976)	No. of lines	9	8	- 11
	No. of trips	11,200	23,366	+108
	Avg costs ($)	14.66	11.72	- 20
Sonntag[a] (1977)	Direct trips (%)	49	59	+ 20
	Avg travel time (sec)	590	520	- 12
	Possible demand (no. of trips)	84,100	86,100	+ 2
Hasselstroem (1979)	No. of trips	8,636	9,435	+ 9
	Total costs (skr)	9945	8401	- 16
	Cost/trip (skr)	1.15	0.89	- 23

Note: NA = not available.

[a]The results are for an earlier and simpler form of the algorithm that was applied to part of Berlin's bus network.

well-calibrated modal-choice model may be the problem here.

Except for direct validation of the predicted performance, the best measures of performance would be obtained by applying different algorithms to the same network and travel demand pattern. Unfortunately, only one such comparison is presented in the literature. Sonntag (18,19) compared his solution for the Duesseldorf light rail network with that of Nebelung (12). Sonntag's network was marginally better with about 2 percent more direct trips (77.49 versus 76.15 percent) and 2 percent lower average travel times (1,483.7 versus 1,511.8). The small difference between Sonntag's sophisticated and complicated algorithm and Nebelung's simple, only partially computerized algorithm is surprising. The Duesseldorf network, however, was not very complex and may be more amenable to the extensive user involvement required by Nebelung's algorithm.

Overall Conclusions

The evaluation of the existing heuristic algorithms shows that both of the basic approaches to network improvement--incremental development from a loaded spider network versus minimum-time-path base generation of feasible lines--give reasonable results. The magnitude of the predicted improvement for the two incremental approaches (Rosello and Sonntag) is similar to that for the other approaches shown in Table 4. No firm conclusions can be drawn from the direct comparison of Sonntag's incremental approach with Nebelung's early algorithm based on minimum time path.

No published results of applications of the two most recent algorithms (those of Mandl and Sahling) were available. Both algorithms represent a departure from prior algorithms based on minimum time path in which lines close to the shortest path are included in the set of feasible lines. Mandl only considers the shortest paths with the maximum number of nodes and initially only allows one line per terminal. The feasible set of base lines, thus, is quite small. Sahling also only considers minimum time paths for lines. He selects the most productive minimum time paths as lines incrementally and does not attempt to optimize the base network as Mandl does.

Because the algorithms of both Mandl and Sahling represent untested new approaches to heuristic network optimization and are likely to require substantially less computational time, they both should be tested and compared with the other approaches. Additional direct comparisons between algorithms representing the incremental approach versus shortest-path-based generation are also needed to provide an improved basis for selecting heuristic algorithms for practical applications and for providing direction to further research and development of heuristic algorithms.

SELECTION OF A HEURISTIC ALGORITHM FOR TESTING

In selecting a heuristic algorithm for application in the United States, only the most recent algorithms were considered because these are more likely to incorporate the latest shortest path and other routines that take advantage of the advances in computer speed and memory size. Of the five most recent algorithms, only those of Sonntag and Mandl were potentially available at the beginning of this study. Sahling's work was not yet published officially and that of Dubois and Hasselstroem was not discovered until the study was well under way. Unfortunately, Sonntag's algorithm is programmed in ALGLOW, a special version of ALGOL developed at the Technische Universitaet Berlin. The time required for translation into standard ALGOL or FORTRAN was not available. Also, because of its complexity, Sonntag's algorithm would have taken a long time to implement. Consequently, Mandl's algorithm was selected for testing a state-of-the-art heuristic algorithm in the context of a small to medium-sized American city.

Mandl's algorithm is programmed in FORTRAN and is available as part of the Interactive Network Optimization (IANO) system from the Institute for Advanced Studies in Vienna (24). IANO has been implemented on both Sperry Univac and Digital Equipment (VAX) computers. For this study the programs were implemented on a VAX 11/780 with 4 megabytes of main memory, virtual memory, and a time-sharing operating system. Mandl's algorithm is contained in the program BUSOPTMAIN and its related subroutines. Implementing BUSOPTMAIN on the VAS 11/780 required only three minor modifications in the program.

SELECTION OF TEST CITIES

Mandl's algorithm could potentially be applied to a wide range of transit networks in the United States. Both radial and grid networks with buses or light rail could be analyzed. In general, the test city should meet the following criteria:

1. It should have a small to medium-sized transit network,
2. Transit network and transit demand data should be available,
3. The network should need restructuring, and
4. The analyst should be familiar with the transit system and urban activity system characteristics.

The first criterion reduces the computing requirements whereas the second reduces the cost of data collection. The transit demand data can be obtained from either a recent on-board survey or a calibrated

modal-choice model. Most transit systems in the United States could benefit from an evaluation of the need for restructuring routes. Familiarity with local conditions can be obtained quickly if necessary through interviews with local transit staff, traffic engineers, and planners.

Because Madison, Wisconsin, met all the criteria at the lowest cost for this study, it was selected as the test city in the United States. For a city of its size Madison has excellent bus service, partly because the central area is located on a narrow isthmus between two large lakes. As a result, the radial lines to and through the central business district (CBD) provide a high level of service to the central area. The bus operator, Madison Metro, provides service on 19 lines with a fleet of 189 buses (27). A minor restructuring was completed in 1979 to provide direct service between the two major regional shopping centers.

In order to compare the performance of Mandl's algorithm with that of other heuristic algorithms, Duesseldorf, West Germany, was selected as the second test city. Duesseldorf is the only city for which the performance of two other heuristic algorithms (Nebelung's and Sonntag's) has been reported. Duesseldorf also meets all but the fourth criterion for selection of a test city. Thus, Mandl's sparse, shortest-path-oriented algorithm can be compared directly with Nebelung's emphasis on an expanded set of feasible near-shortest-path lines and Sonntag's incremental construction of optimal lines.

IMPLEMENTATION OF MANDL'S ALGORITHM

Model Structure

Mandl's algorithm is divided into two stages. In the first a feasible network is created. An attempt is made to maximize direct trips by creating lines along the longest shortest paths. Only the geometrics of the base network are considered, not the demand. Thus, the resulting lines may not be along the paths with the highest demands. In the second stage the initial lines are recombined in an attempt to minimize transfers through the minimization of total travel time. The flowchart for the two stages is presented in Figure 1. The second stage can be started with either the network generated in the first stage or a user-specified network.

The second stage of the algorithm uses an objective function defined as total travel time including travel and waiting times. The waiting time is assumed to be equal on all lines and is calculated as the ratio of the network length to the given number of buses. Thus, the number of buses can be used to increase or decrease the amount of waiting time. An all-or-nothing assignment algorithm is used with one-half of the demand assigned on the basis of minimum time and one-half on the basis of minimum transfers. The rationale here is that route-choice behavior is not adequately represented by travel time alone. This is an arbitrary split that could be changed with only a minor modification to the computer program.

The second stage follows a hierarchical, three-step search for improvements. If an improvement is found at any step, the search returns to the lowest step. In the first step, the intersecting lines at the node with the highest net transfers are recombined so that the net transfers are reduced. In the second step new transfer points are created by rerouting lines to include nodes with large flows. Thus, feasible detours are created. In the third and highest step transfer nodes with the lowest total activity (flow plus transfers) that are not on the shortest path are proposed for elimination. Those detours that result in higher total travel time are eliminated.

The main limitation of Mandl's algorithm is the initial lack of consideration of the demand patterns. If the range of lines considered is large enough, the minimization of transfers in the second stage should reorient the network to better serve the demand.

Computer Program

The original version of the computer program was designed to accommodate a maximum of 40 nodes with all the data entered interactively. In order to accommodate the Madison network, the maximum network size was increased to 150 nodes. To facilitate handling large networks, the trip table and list of nodes were input from mass storage files.

The most critical subroutine used by the program is the shortest-path algorithm, SHOPAT. SHOPAT uses Floyd's algorithms for the calculation of shortest paths. As a matrix formulation, Floyd's algorithm provides a memory-intensive but fast solution. SHOPAT separates the network into a transfer network and a complementary network. The transfer network includes all nodes where transfers between the lines are possible for the first and for the last time between two lines. This transfer network is reformulated to include links that represent waiting and transfer times. For this network, which is smaller (has a smaller number of nodes) than the original network, the shortest distances are calculated. This procedure saves computing time and storage space, because the requirement for both in Floyd's algorithm grows rapidly with the number of nodes. The shortest distances between all the other nodes of the network are calculated by finding the transfer node, which is nearest to the destination node.

Because the design of large networks is not computationally feasible with Mandl's algorithm, a program was developed to aggregate a base network to a reasonable size. The user specifies the equivalence table for the aggregation and the program computes the new center-of-gravity zonal centroids and generates the corresponding compressed trip table and distance matrix. The selection of the new aggregate zones is critical because the characteristics of the new zones directly affect the results of the optimization.

The main program that generates the optimal network uses only total travel time as the objective function. Thus, a separate program was developed to provide a wider range of evaluation measures, including the length of the base and the line networks, the mean squared error for the difference in travel time for each origin-destination (OD) pair between the optimum network and the actual network, and the average travel times, waiting times, and number of transfers for all users and for every stop. The demand density, defined as the ratio of the number of trips with n transfers to the number of OD pairs connected with n transfers, is also calculated. The evaluation program applies Floyd's shortest-path algorithm to the transit network and performs an all-or-nothing assignment of transit trips at the same time. These evaluation measures allow a more detailed comparison of the various solutions in terms of passenger-oriented performance measures. It would have been desirable to calculate the frequencies of the various lines, more accurate values of waiting and transfer times, and an ap-

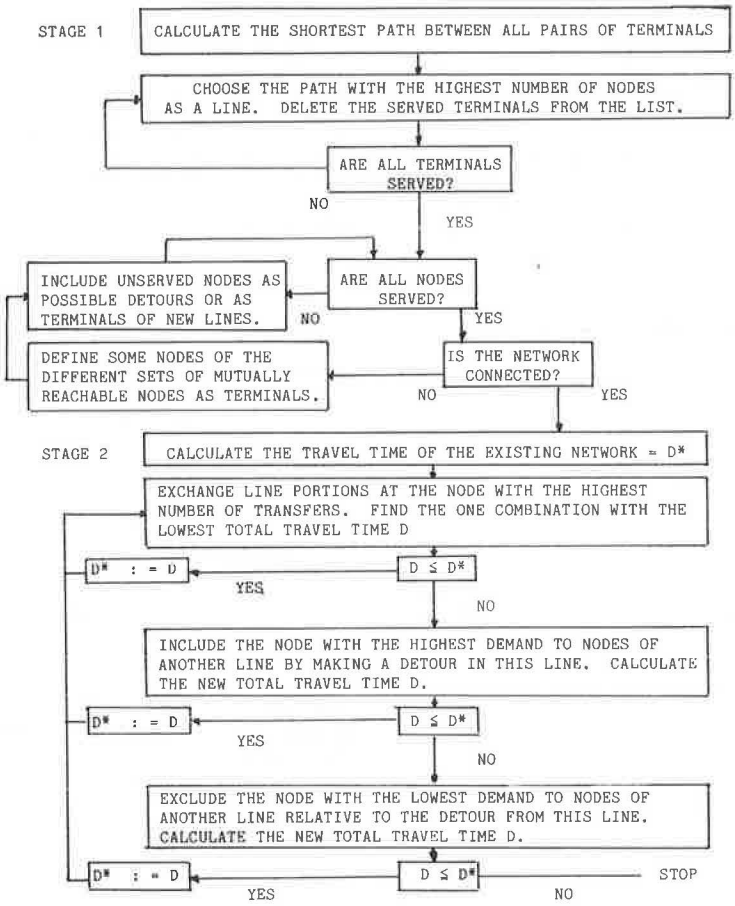

FIGURE 1 Flowchart of Mandl's algorithm.

proximate measure of the operating costs involved, but the time to develop a program capable of generating these performance measures was not available.

Computational Experience

The computer time required by Mandl's algorithm is difficult to predict because the number of iterations through the three-step hierarchical network optimization process is a complex function of the starting point (number and location of terminals and ring lines), the network geometry, and the demand. The range in computer central processing unit (CPU) time as a function of the number of nodes in the network is shown in Figure 2. Both the range in times and the maximum time increase rapidly with increasing network size. The best least-squares fit to the data is given by number of nodes (NN) to the third power. The regression equation for CPU time in minutes is

$$\underset{(-0.25)}{\text{CPU} = -3.96} + \underset{(3.60)}{0.000213 * \text{NN}^3} \qquad (1)$$

with an overall t = 2.07 and an R^2 = 0.35. The two other possible independent variables, number of terminals and number of ring lines, did not contribute significantly to the explanatory power of the regression equation.

The rapid increase in CPU time with increasing network size is characteristic of the assignment and shortest-path calculations that are the most time-consuming elements of Mandl's algorithm. The results for the 73-node network show that unless computer time is nearly free, networks must be limited to 70 to 80 nodes.

APPLICATION TO MADISON

Development of the Data Base

A reasonable transit trip matrix was available from the 1980 on-board survey conducted by the transit operator, Madison Metro. Computerized highway and

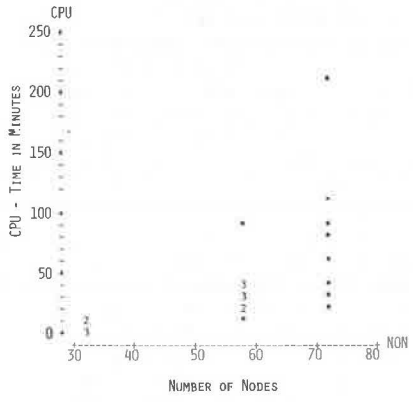

FIGURE 2 Computing time as a function of network size.

transit networks at the level of 377 traffic analysis zones were available from the Wisconsin Department of Transportation. The minimum time paths for the highway network were selected for developing aggregate networks for Mandl's algorithm because the networks would not be biased by the existing transit route pattern. The primary disadvantage of using highway travel time is that node pairs connected by higher-speed highways are favored. The problem is minimized in Madison because the only freeway within the urban area is a short, circumferential facility with no transit service and hence no transit trips.

The 377-zone base network was aggregated first to 89 zones based on Planning Analysis Area (PAA). The PAA zone system is coarse in the downtown area and does not provide for clear delineation of multiple corridors within the central isthmus. A second zonal system with 104 zones was developed in order to define central corridors more clearly. Even for the 104-zone system the definition of the central corridors was limited by the base zonal system, which uses arterial streets as zonal boundaries. For transit planning, zones that straddle main arterial streets would represent potential transit corridors more accurately, so that a finer-grained zonal system would not be needed. At the regional scale the lack of a full range of central corridors was not a major problem because the overall distances involved were small. Also, precise identification of all central corridors was not essential because the overall transit network is represented at a sketch-planning level. Network alternatives are evaluated on a comparative basis with the objective of finding directions for possible improvement.

In order to reduce the size of the network, zones with less than 25 transit trip origins or destinations were consolidated with other zones. After consolidation, the first aggregation contained 58 zones (nodes) and 122 links. The second aggregation contained 73 zones and 153 links.

Results for the First Aggregation

The 58-zone network was used to test the basic operation of Mandl's algorithm and to develop a strategy for selecting a set of input parameters. In running the program the number of buses was held constant at 120. The other two inputs, the number of terminals and the number of ring lines, were varied to test the sensitivity of the algorithm.

The evaluation measures for 11 runs of the model are presented in Table 5. The evaluation program was also run for a network that simulates the existing Madison Metro bus network. The evaluation program was run both with and without a 10-min transfer-time penalty. With the transfer penalty the assignment model in the evaluation program assigns a much higher proportion of the trips to direct routes, that is, routes with no transfers.

The first run uses the obvious ends of lines as terminals. The second run tries to substitute ring lines for some of the terminals as do runs 4, 5, and 9. Runs 3 and 8 specify some high-use terminals more than once to force the algorithm to build lines through high-density corridors. Runs 6 and 7 again use the obvious terminals, but the main purpose of these two runs was to test the sensitivity of the algorithm to the input order of the terminals. The different input orders did change the output network. Thus, the user would be well advised to test the sensitivity of the best output networks. Runs 10 and 11 use extremes. Run 10 gives total freedom to the algorithm to chose the terminals except that one terminal is located in the CBD, whereas run 11 forces the algorithm to build all lines along ring lines that correspond to the lines of the existing Madison Metro network.

In evaluating the results for Mandl's algorithm to find directions for possible network improvements, Table 5 shows clearly that the algorithm, in general, produces substantially shorter networks but at the expense of much greater proportions of transfers and somewhat longer travel times. Because of the shorter network, the indirectness of routing as measured by the mean squared error (MSE) is also generally greater than that for the simulated Madison Metro network.

When analyzed in the context of the structure of Mandl's algorithm, the performance measures shown in Table 5 are consistent with that structure. Solution 10 allows the algorithm to select shortest-path lines unconstrained by terminal locations. The result is the shortest network but at the expense of long average travel times. The addition of terminals and ring lines consistently increases the length of the network and generally provides more direct routing (lower MSE). Additional terminals and ring lines, however, do not guarantee lower average travel times and fewer transfers. The travel times are sensitive to the location of terminals and ring lines. For example, solution 5 has the second lowest travel time and a moderate network length. Excess line length exists because two of the nine lines connect low-demand suburban areas. The problems with solution 5 were solved in solution 6 by selecting fewer terminals and eliminating the ring line. The result is a much shorter network with only a small increase in average travel time.

In terms of total travel time, the best computer-generated network is solution 11, which was designed specifically using ring lines to follow Metro's line pattern (Figures 3 and 4). Although the average travel time is about 15 percent longer than that for the Metro network and the proportion of transfers is much greater, solution 11 actually provides more direct service (lower MSE) with a network of about the same length. As can be seen in Figure 4, solution 11 has considerable extra mileage on the south side of Madison, which is required only to complete the specified ring line. Elimination of that mileage should not affect the average travel times significantly.

A comparison of the pattern of lines of solution 11 with that for the Metro network (Figures 3 and 4) shows that Mandl's solution concentrates lines in one corridor through the isthmus rather than spreading the lines over the isthmus following Metro's pattern. The concentration is caused by limiting the feasible lines to only those on minimum time paths. Other algorithms that consider feasible paths as X times the shortest paths or require service on all links should provide better coverage.

In summary, the sensitivity analysis of Mandl's algorithm to combinations of terminals and ring lines indicates that, at least for the radially oriented Madison network, specifying only a moderate number of terminals and no ring lines will produce a balance between network length and travel time. One strategy for selecting terminals is to start with the obvious ends of lines as terminals and incrementally improve subsequent runs by incorporating the good features and eliminating the problems of earlier runs. Such an interactive process would be speeded up considerably through the use of interactive graphics.

Results for the Second Aggregation

The more detailed 73-zone network provides a better definition of corridors within the central area.

TABLE 5 Performance Measures for the First Aggregation: Madison

Solution	Number of Terminals	Number of Ring Lines	Transfer Penalty	Average Time Vehicle	Wait	Total	Distribution of Transfers (in percent) 0	1	2	MSE[a]	Length Ratio
Metro	19	0	0	9.47	5.63	15.10	63	30	7	39.6	.8912
			10	11.58	5.84	17.12	88	10	1	54.8	
1	12	0	0	13.38	4.05	17.53	47	38	14	56.4	.5441
			10	16.0	6.73	22.73	68	29	3	104.5	
2	8	2	0	13.40	4.54	17.94	53	36	11	47.8	.6510
			10	14.96	6.63	21.59	69	26	5	86.2	
3	19	0	0	13.59	4.60	18.19	54	36	11	26.4	.6567
			10	14.34	8.67	23.01	60	36	4	59.6	
4	9	1	0	13.38	3.92	17.30	47	40	13	35.8	.5310
			10	15.55	6.68	22.23	67	28	5	61.9	
5	15	1	0	11.53	4.49	16.02	55	33	11	27.4	.6398
			10	14.46	7.52	20.18	65	32	2	50.1	
6	13	0	0	14.18	3.38	17.56	69	27	3	118.1	.5647
			10	14.46	6.24	20.70	71	27	1	131.0	
7	13	0	0	13.92	3.95	17.87	57	32	11	72.0	.5722
			10	15.16	6.42	21.58	71	28	1	129.0	
8	30	0	0	13.20	5.31	18.51	61	33	6	44.6	.8274
			10	13.56	8.92	22.48	65	32	3	80.3	
9	13	3	0	13.40	4.40	17.80	70	28	2	39.9	.7505
			10	13.61	7.25	20.86	71	28	1	60.5	
10	1	0	0	18.05	2.75	20.80	63	28	8	393.1	.4296
			10	18.91	5.63	24.54	74	18	8	416.5	
11	0	8	0	12.74	4.86	17.60	73	25	2	18.2	.8668
			10	13.15	6.63	19.78	76	22	2	32.6	

[a]Mean Squared Error

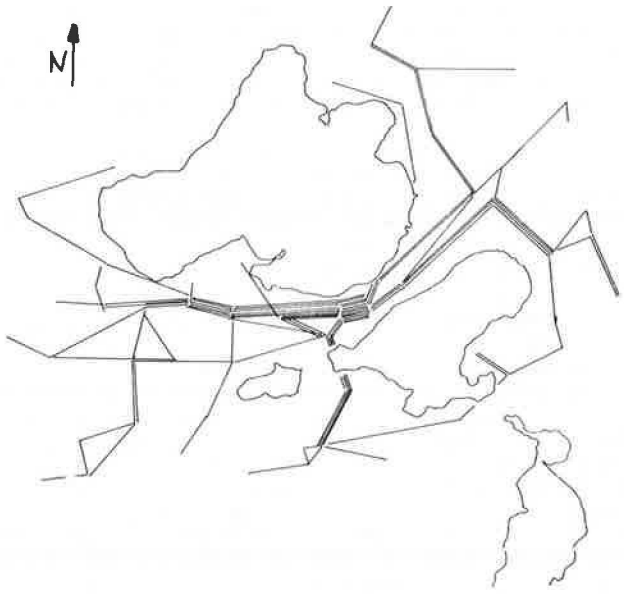

FIGURE 3 Abstraction of the Madison Metro network for the first aggregation.

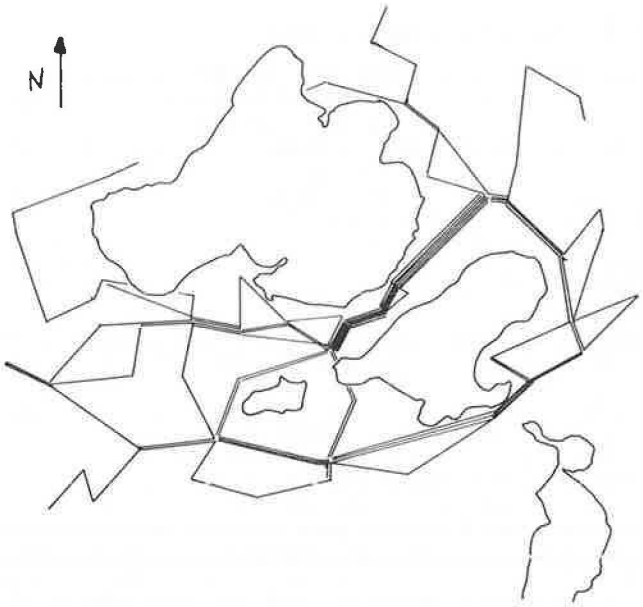

FIGURE 4 Solution 11, generated by Mandl's algorithm for the first aggregation.

Thus, it may be possible to develop a network that provides better coverage through the isthmus.

Because the results of the first aggregation showed that reasonable networks could be obtained from a range of combinations of terminals and ring lines, the same basic strategy of covering the range of possibilities was used. Six runs were made by varying the number of terminals from 1 to 25 and the number of ring lines from 0 to 5. The results of applying the evaluation program to the six runs plus an abstraction of the actual Madison Metro network are shown in Table 6.

TABLE 6 Performance Measures for the Second Aggregation: Madison

Solution	Number of terminals	Number of ring lines	Transfer Penalty	Average Time Vehicle	Wait	Total	Distribution of Transfers (in percent) 0	1	2	MSE	Length Ratio
Metro	19	0	0	13.32	6.78	20.10	75	24	1	36.0	1.0998
			10	13.45	7.81	21.26	82	18	-	58.7	
1	15	0	0	15.55	4.12	19.62	61	37	1	131.1	0.5973
			10	15.64	7.87	23.51	62	37	1	140.5	
2	25	0	0	12.81	6.31	19.12	51	40	10	68.4	0.8139
			10	13.79	10.12	23.91	58	40	2	95.8	
3	15	1	0	13.83	5.25	19.08	54	43	3	79.4	0.7174
			10	14.02	9.63	23.65	54	44	2	87.7	
4	10	3	0	14.66	4.67	19.29	56	40	4	95.1	0.6430
			10	15.02	8.38	23.40	61	37	1	105.1	
5	10	5	0	13.82	5.45	19.27	52	42	7	65.6	0.7242
			10	14.56	8.87	23.43	62	37	1	87.8	
6	1	0	0	17.24	4.28	21.52	37	45	17	310.3	0.4924
			10	18.29	8.29	26.58	55	42	3	377.3	

In comparing the Metro network with the six computer-generated networks to identify directions for possible improvement, the computer-generated networks again are much shorter than the Metro network. The reduction in length is achieved at the expense of more transfers and less direct routing. In terms of travel time, the computer-generated networks are generally better than the Metro network if transfer penalities are not included, but with a 10-min transfer penalty, the Metro network is consistently better. The computer-generated networks are able to provide direct routes that give short travel times only by using many transfers. The much longer Metro network provides many more opportunities for direct travel. When transfer penalties are added, the impact on average travel times is smaller for the Metro network than for the computer-generated networks because the proportion of transfers is much smaller. A similar result was observed for the first aggregation; however, the higher level of aggregation did not permit the computer-generated solutions to match the pattern of demand as closely.

In comparing the performance measures shown in Table 6 for the six solutions, the total travel time both with and without transfer penalties is relatively insensitive to the number of terminals and ring lines. The one exception is solution 6. Here the computer-generated lines are unconstrained by terminals or ring lines, resulting in the shortest possible network but with many transfers and very indirect routing (high MSE).

Because the total travel time for solutions 1 to 5 is about the same, selection of the best computer-generated network depends on the relative importance of the three other performance measures. From a transit passenger's perspective, minimizing transfer will be most important. Solutions 1, 4, and 5 have the lowest and about equal transfer requirements. The basic trade-off then is between indirectness of routing and network length (MSE versus length ratio). Solution 4 provides an intermediate point between the two extremes. It provides more direct routing than solution 1 and a shorter network than solution 5.

The Madison Metro network and solution 4 are compared in Figures 5 and 6. In terms of coverage, solution 4 identifies corridors through the isthmus better than in the first aggregation. Solution 4, however, clearly does not follow the demand-oriented pattern of the Metro network. On the east side of the CBD, the lines for solution 4 are concentrated on the north side of the isthmus whereas the Metro network follows demand on the south side. On the west side of the CBD the concentration of the lines for solution 4 again does not follow Metro's demand-oriented pattern. This illustrates a major weakness of Mandl's algorithm. The initial network selection procedure is not designed to follow demand. Because lines cannot subsequently be added or deleted, the rearrangement of lines to minimize transfers in the second stage cannot significantly change the line pattern to follow demand.

The primary advantage of Mandl's solutions is the much shorter network. Such solutions would be of interest if substantial cutbacks in service are required or light rail networks are being developed. The advantage, however, is outweighed by the lack of consideration for demand patterns. The results of the application to Madison indicate the importance of multiple performance measures in evaluating alternative networks.

APPLICATION TO DUESSELDORF

The data set for this application was generated in 1960 for Nebelung's study about the reorganization of light rail systems (12). Sonntag used this data set for the development of his algorithm, although he intended his algorithm for the design of bus networks (18,19). The application of Mandl's algorithm to the data set permits comparison of Mandl's geometrically oriented algorithm with Nebelung's algorithm, which maximizes direct trips on lines that are at most 1.2 times shortest paths, and Sonntag's, which builds lines incrementally along the demand without too strong an orientation toward the shortest paths.

In searching for the best solution with Mandl's algorithm, six combinations of terminals and ring lines were tested. For the 32-node, 48-link network the computing times ranged between 2 and 8 min on a minicomputer, which is much less than the 30 min of IBM 370-158 time required for Sonntag's algorithm. Using total travel time as the performance criterion, the run with only one terminal and no ring lines gave the best results.

The performance of Mandl's best solution is compared with those of Nebelung and Sonntag in Table 7. The comparison must be interpreted in view of the

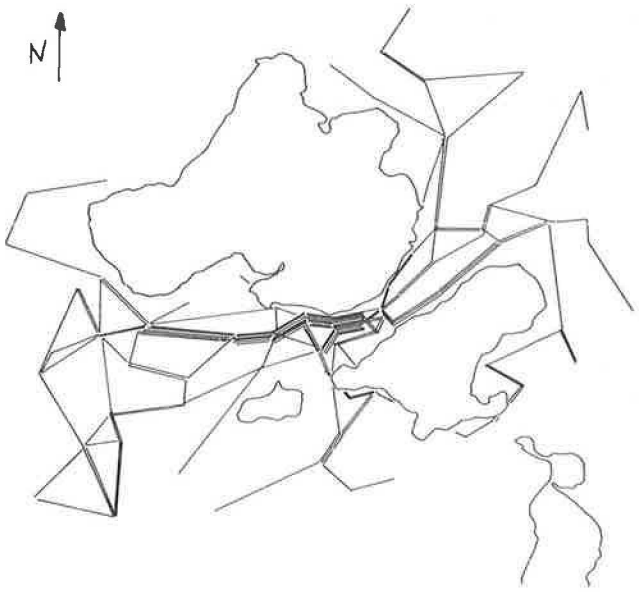

FIGURE 5 Abstraction of the Madison Metro network for the second aggregation.

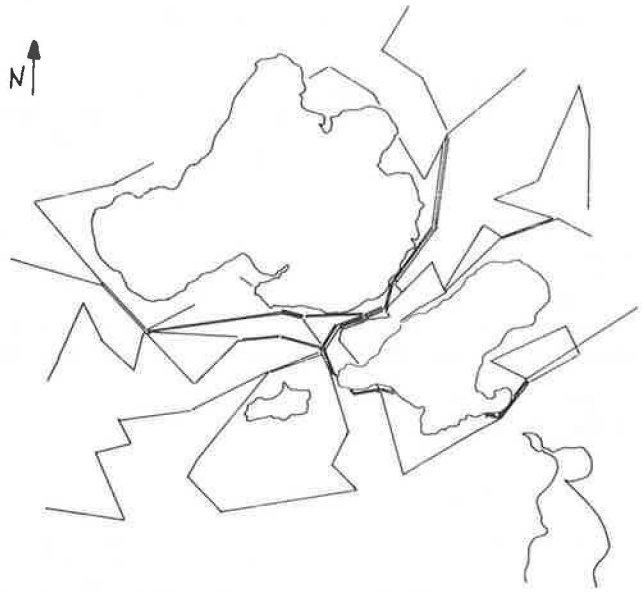

FIGURE 6 Solution 4, generated by Mandl's algorithm for the second aggregation.

difference between Nebelung's and Sonntag's network representation, in which the nodes are the intersections of the light rail tracks, and Mandl's, in which the nodes are zones. Nebelung and Sonntag require that all links be served at least once, whereas Mandl only requires that all nodes be served. Sonntag's algorithm gives the best results with the smallest average travel time for the realistic 10-min transfer penalty, the highest percentage of direct trips, and the most direct connections as indicated by MSE. The differences from Nebelung's solutions are not very large, which might be caused by the relatively small network. Sonntag's solution follows the demand patterns closely. The solution has the highest demand density for the direct connections and the lowest density for the indirect connections.

TABLE 7 Performance Measures for Duesseldorf Network Solutions

	Transfer Penalties (min) by Author					
	Nebelung		Sonntag		Mandl	
Performance Measure	0	10	0	10	0	10
Avg travel time (min)						
In vehicle	19.66	20.21	20.61	21.21	21.21	21.82
Waiting	12.40	16.31	11.96	15.71	9.50	15.24
Total	32.06	36.52	32.57	36.53	30.71	37.06
Transfers (%)						
None	54	55	61	61	38	41
One	41	43	38	38	56	57
Two	5	1	1	0	6	2
Demand density[a]						
No transfers	288	285	299	300	273	271
One transfer	192	188	175	169	226	220
Two transfers	146	121	83	56	127	85
MSE[b]	19.09	19.89	4.06	8.74	22.99	31.78
Network length	1.327	–	1.35	–	0.901	–

[a]Ratio of trips with X transfers to the number of OD pairs connected with X transfers.
[b]Mean square error of difference between actual path and minimum path time.

Mandl's solution requires an unacceptably high level of transfers, although the solutions of Nebelung and Sonntag also have a high level of transfers. Mandl's extremely high transfer levels are the results of a substantial reduction in network length without orienting the reduced network to maximize direct trips. The sparseness of Mandl's network compared with that of Sonntag is shown in Figures 7 and 8. Mandl's network lacks the coverage and many connecting lines provided by Sonntag's network. As for Madison, the advantage of a reduced network is outweighed by the lack of consideration for the demand pattern, which results in too many transfers even for a light rail network.

CONCLUSIONS

The review and evaluation of 13 heuristic transit network optimization methodologies revealed a wide range of approaches that should provide reasonable solutions for transit networks in small and medium-sized American cities. In general, the methodologies

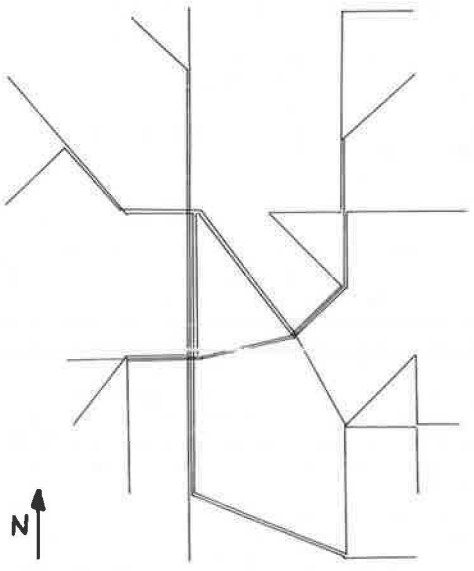

FIGURE 7 Duesseldorf network: solution 6, generated by Mandl's algorithm.

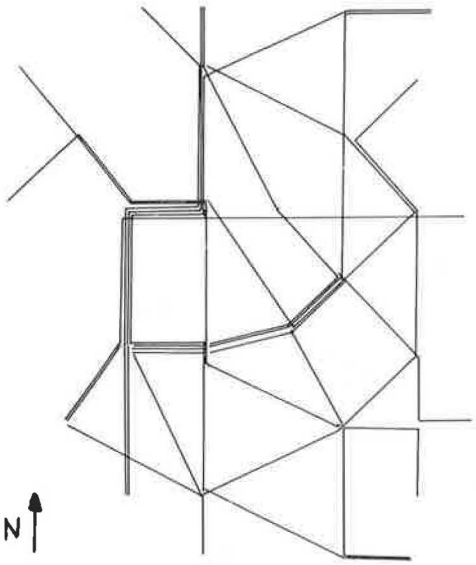

FIGURE 8 Duesseldorf network: Sonntag's solution.

have good potential for generating improved solutions because either a wide range of base lines is considered or an incremental approach to line construction is used. Most of the available heuristics provide a theoretically sound basis for network improvement because factors relevant to modal choice such as travel time, transfers, and cost are generally incorporated into the selection of base lines and the objective function and constraints for the algorithms. Although not all of the heuristics have been applied to actual transit networks, the computational requirements appear to be reasonable and should not be a major constraint on implementation, at least for medium-sized networks. The predicted improvements from applications of a few of the algorithms to one American and several European cities are positive enough to warrant further applications.

One major problem in applying the heuristic methodologies in the United States is the lack of readily available software in either the public or private domain. The emphasis in the United States has been on systems analysis with interactive graphics for transit network improvement. Thus, there has been almost no practical experience with the wide variety of heuristic methodologies that have been developed in Europe.

The potential for applying heuristic network optimization algorithms in the United States was tested with the unproved new algorithm by Mandl. The application of Mandl's algorithm to Madison, Wisconsin, showed first that with appropriate software and documentation, the computer program for a fairly complex heuristic algorithm can be implemented quickly and easily. Second, the available data base from a 1980 on-board OD survey and a standard urban area highway network provided an adequate basis for applying Mandl's algorithm. Such data are available in many American cities. Third, the computational requirements of an algorithm based on minimum time path are affected dramatically by the number of nodes in the network. Initial testing of an algorithm is probably done most efficiently at a reasonably high level of aggregation. A more detailed network, however, many be required to represent critical corridors adequately, as was the case in Madison. The two levels of analysis have the advantage of indicating the sensitivity of the model to the level of aggregation.

The most important result of the application of Mandl's algorithm to Madison was the demonstration of the need for a theoretically sound and fully tested algorithm. Mandl's algorithm held out the promise of a sparse network that could still satisfy demand through transfer optimization. The application to Madison resulted in a substantial reduction in network length but at the expense of an unacceptable increase in transfers for a bus network. Also, the revised network lacked the directness of routing provided by the base network. The importance of having a full range of performance measures available for evaluating the solutions generated by algorithms such as Mandl's was clearly indicated for Madison. If travel time and network length alone had been considered, Mandl's solutions would have appeared to perform well.

The same basic limitations of Mandl's algorithm were observed for the Duesseldorf light rail network application. The reduced network did not serve the demand directly and required high levels of transfers. Although emphasis on a small network is appropriate for the development of a light rail system in order to minimize capital costs, following demand so that ridership is maximized is even more important. Also, for restructuring an existing light rail network, Mandl's algorithm is not really appropriate because the algorithm is not constrained to serve all links.

Finally, the applications of Mandl's algorithm to Madison illustrate the problem of selecting the input parameters. For Madison it was not sufficient to choose only the obvious terminals. Experiments with ring lines and multiple input of certain terminals were necessary to find the best networks.

Based on the results of the Madison and Duesseldorf applications, there is a clear need to make Mandl's algorithm more responsive to demand. One possibility is to use Sahling's approach for selecting lines based on the most productive shortest paths in place of Mandl's selection of the longest shortest path of the base network. Because the number of lines should not increase dramatically, the computing requirements of Mandl's algorithm should remain reasonable. Mandl's second stage could also be applied to the results of other simple algorithms such as Nebelung's and others or to an existing base network. Mandl's second stage should be particularly useful for identifying how existing lines can be reoriented to reduce transfers.

DIRECTIONS FOR ADDITIONAL RESEARCH

The review and evaluation of 13 heuristic methodologies, the application of Mandl's algorithm to Madison, and the comparison of Mandl's solution for Duesseldorf with those of Nebelung and Sonntag suggest two directions for future research. First, the existing transit network should be used as input to the second network improvement stage of a number of algorithms, including those of Rosello, Dubois, Sonntag, Mandl, and possibly Hasselstroem. This strategy is particularly important for Mandl's algorithm because Mandl's base network development phase is defective. For some of the algorithms the base network lines could be expanded to include families of lines that are within X times the base-line lengths.

Second, additional development and applications of full-scale heuristic methodologies are needed. The relative advantages of the three basic approaches to base network development--family of X times the shortest paths, incremental, and incremental selection of shortest paths to maximize productivity (Sahling)--should be evaluated so that

real-world applications can be made with greater confidence. The potential for reducing the complexity of the base network using the network selection algorithms of Rea and Hasselstroem should also be explored further.

Validation of the predictions of the various heuristic methodologies is also a critical need. As an alternative, synthetic validation is possible with a calibrated modal-choice model. An external modal-choice model could also be applied iteratively with a heuristic algorithm to approximate an equilibrium solution to the interaction of supply and demand.

Finally, research is needed on the sensitivity of network performance to frequency optimization. The potential for including both line network generation and frequency optimization in one step using Hasselstroem's linear programming approach should be explored further.

ACKNOWLEDGMENT

The data used in this study were provided by the staff of Madison Metro, the Dane County Regional Planning Agency, and the Wisconsin Department of Transportation. The help and support provided by numerous individuals in these organizations are gratefully acknowledged. Support for computer time was provided by the Engineering Experiment Station at the University of Wisconsin-Madison.

REFERENCES

1. T.H. Chua and D.T. Silcock. The Practice of British Bus Operators in Planning Urban Bus Services. Traffic Engineering and Control, 1982, pp. 66-70.
2. G.P. Sharp. Public Transit System Network Models: Consideration of Guideway Construction, Passenger Travel and Delay Time and Vehicle Scheduling Cost. Ph.D. dissertation. Georgia Institute of Technology, Atlanta, 1974.
3. M. Rapp. Transit System Planning: A Man-Computer Interactive Graphic Approach. In Highway Research Record 415, HRB, National Research Council, Washington, D.C., 1972, pp. 49-61.
4. M. Rapp. Interactive Graphics System for Transit Route Optimization. In Highway Research Record 559, HRB, National Research Council, Washing- ton, D.C., 1976, pp. 73-88.
5. M. Rapp and C.D. Gehner. Transfer Optimization in an Interactive Graphics System for Transit Planning. In Highway Research Record 619, HRB, National Research Council, Washington, D.C., 1976, pp. 27-33.
6. M. Rapp and R. Keller. Netzoptimierungs System fuer die Netzstruktur der oeffentlichen Nahverkehrsmittel in der Region Basel. Baseler Verkehrsbetriebe, Basel, Switzerland, 1975.
7. G.B. Douglas. Applications of Interactive Graphics in Urban Transit Analysis. In Transportation Research Record 866, TRB, National Research Council, Washington, D.C., 1982, pp. 18-24.
8. J.C. Rea. Designing Urban Transit Systems: An Approach to the Route Technology Selection Problem. In Highway Research Record 417, HRB, National Research Council, Washington, D.C., 1972, pp. 48-59.
9. J. Hsu and V.H. Surti. Decomposition Approach to Bus Network Design. Transportation Engineering Journal of the ASCE, Vol. 103, 1977, pp. 447-459.
10. J. Hsu and V.H. Surti. Demand Model for Bus Network Design. Transportation Engineering Journal of the ASCE, Vol. 102, 1976, pp. 451-461.
11. J. Hsu and V.H. Surti. Framework of Route Selection in Bus Network Design. In Transportation Research Record 546, TRB, National Research Council, Washington, D.C., 1975, pp. 44-57.
12. H. Nebelung. Rationelle Umgestaltung von Strassenbahnnetzen in Grossstaedten. Ministerium fuer Wirtschaft, Mittelstand und Verkehr des Landes Nordrhein, Westfalen, West Germany, 1961.
13. W. Lampkin and P.D. Saalmans. The Design of Routes, Service Frequencies and Schedules for a Municipal Bus Undertaking: A Case Study. Operations Research Quarterly, Vol. 18, 1967, pp. 375-397.
14. L.A. Silman, Z. Barzily, and U. Passy. Planning the Route System for Urban Buses. Computers and Operations Research, Vol. 1, 1974, pp. 201-211.
15. H.P. Hoidn. Busnetz von Aarau, Neukonzeption der Linien. Institut fuer Operations Research der ETH Zuerich, Zuerich, Switzerland, 1975.
16. X. Rosello. An Heuristic Algorithm to Generate an Urban Bus Network. In Advances in Operations Research: Proceedings, Elsevier, Amsterdam, 1976.
17. D. Dubois, G. Bel, and M. LLibre. A Set of Methods in Transportation Network Synthesis and Analysis. Journal of the Operations Research Society, Vol. 30, 1977, pp. 797-808.
18. H. Sonntag. Linienplanung im Oeffentlichen Personennahverkehr. Dissertation. Fachbereich Informatik, Technische Universitaet Berlin, 1977.
19. H. Sonntag. Ein heuristisches Verfahren zum Entwurf nachfrageorientierter Linienfuerhrung im oeffentlichen Personennahverkehr. Zeitschrift fuer Operations Research, Vol. 23, 1979, pp. B15-B23.
20. D. Hasselstroem. Connecting Bus Routes at Points of Intersection. In Advances in Operations Research: Proceedings, Elsevier, Amsterdam, 1976.
21. D. Hasselstroem. A Method for Optimization of Urban Bus Route Systems. Report 88-DH-03. Volvo Bus Corporation, Gothenburg, Sweden, 1979.
22. Martin and Voerhess Associates and Volvo Transportation Systems. Coventry Bus Study: Summary Report. West Midlands Passenger Transport Executive, England, 1981.
23. C. Mandl. Applied Network Optimization. Academic Press, New York, 1979.
24. C. Mandl. User Reference and Implementer Manual IANO-System. Institutsarbeit No. 122. Institute for Advanced Studies, Vienna, Austria, 1979.
25. Seminar Verkehrswesen, Sommer Semester 1981: Linienplanung. Institutsnotiz 29. Instituts fuer Verkehrswesen der Universitaet Karlsruhe, West Germany, 1983 (description and application of Sahling's model).
26. D. Hasselstroem. Public Transportation Planning: A Mathematical Programming Approach. Department of Business Administration, University of Gothenburg, Sweden, 1981.
27. Dane County Regional Planning Commission: Transit Development Program 1982-1986. Madison, Wisc., 1982.

Application of an Algorithm for Estimating Freeway Trip Tables

ROBERT W. STOKES and DANIEL E. MORRIS

ABSTRACT

The application of an existing computer program that uses only the marginal totals of a freeway trip table to estimate ramp-to-ramp traffic flows is described. The flows estimated by this program [synthetic origin-destination matrices (SYNODM)] were compared with origin-destination data obtained from a postcard survey and the distribution and magnitude of the estimation errors across the trip table were documented. A modified version of the basic SYNODM algorithm that incorporates limited origin-destination data into the estimation method is also discussed. Relative to the postcard data, the estimates of trip interchanges obtained from the modified version of SYNODM exhibited a substantially smaller average trip error than those obtained from the basic SYNODM program. However, given the limited scope of the study, no generalizations regarding the adequacy of the basic SYNODM program could be made.

A large variety of freeway planning and management activities require information about the pattern of vehicular traffic flows between entry and exit ramps. Traditionally the collection of such information has involved elaborate, time-consuming, and expensive special-purpose surveys such as roadside interviews. It has been suggested that instead of special-purpose surveys being relied on to gather this information, more readily available information such as mainline and ramp traffic volumes could be used to estimate traffic distribution patterns (1).

Reliable procedures for synthesizing trip distribution patterns offer a number of advantages over strictly empirical approaches. First, synthetic trip distribution data would certainly be less expensive and more easily obtained than data obtained from field studies. Second, the use of available (or in any case easily obtainable) information on flows on the system could greatly reduce the lead time between the data collection and model calibration phases of the planning process. Third, the advantages of low cost and quick turnaround time could lead to broader applications of analytic procedures that, because of their extensive data requirements, have in the past been restricted to use in large-scale studies. Finally, reliable procedures for estimating freeway traffic distributions from readily available, longitudinal data, such as observed flows, could alleviate many of the problems attributable to the cross-sectional nature of empirically derived origin-destination (OD) data.

The objective of this research was to evaluate an existing algorithm for estimating freeway traffic distributions and assess the extent to which the algorithm could be used in lieu of more traditional empirical approaches.

BACKGROUND

In recent years the Texas State Department of Highways and Public Transportation and the Texas Transportation Institute have made extensive use of the FREQ simulation programs (2) in evaluating proposed design configurations and control strategies for freeways in Houston and San Antonio (3-6). One of the basic data inputs required by the FREQ programs is an OD table for each of the time periods being simulated. Fortunately, much of the necessary OD data had been collected as part of earlier planning studies. However, as use of the FREQ programs was extended to more and more freeways it became apparent that more efficient and economical procedures for generating reliable OD data were needed.

A preliminary literature review revealed several OD estimation methods. Nihan (7), for example, has presented a summary of several proposed estimation procedures and their limitations. The computer program available to users of the FREQ simulation programs, synthetic origin-destination matrices (SYNODM) (8), was chosen for preliminary evaluation. SYNODM was designed to be internally compatible with the FREQ simulation programs and therefore represented a logical starting point in the search for an acceptable OD estimation method. The basics of the SYNODM algorithm are presented in the following.

DESCRIPTION OF THE ALGORITHM

Basic Approach

Consider the case of a complete OD matrix (i.e., a matrix in which all cells may have nonzero entries). For such a matrix, with r on ramps and c off ramps, estimates of trip interchanges could be obtained from (1)

$$T^*_{ij} = (O_{i\cdot} D_{\cdot j})/S \qquad (1)$$

where

T^*_{ij} = estimate of trips from origin i to destination j,
$O_{i\cdot}$ = total entering volume for origin i,
$D_{\cdot j}$ = total exiting volume for destination j, and

$$S = \sum_{i=1}^{r} O_{i\cdot} = \sum_{j=1}^{c} D_{\cdot j}.$$

Equation 1 is, of course, the familiar maximum-likelihood estimate (MLE) of individual cell values for statistically independent objects in a two-way contingency table (see Table 1). Although this formulation does not account for any observed regularities in travel behavior nor take into consideration the spatial separation between origins and destinations, it does provide a convenient measure for assessing the precision of the estimation method and the adequacy of the underlying assumption of statistical independence of origins and destinations (1,8). Specifically, under the assumption of independence, the statistic

TABLE 1 Typical Complete OD Matrix (Two-Way r x c Contingency Table)

Origin \ Destination	1	2	...	j	...	c	Total
1	T_{11}	T_{12}	...	T_{1j}	...	T_{1c}	$O_{1.}$
2	T_{21}	T_{22}	...	T_{2j}	...	T_{2c}	$O_{2.}$
.							.
.							.
.							.
i	T_{i1}	T_{i2}	...	T_{ij}	...	T_{ic}	$O_{i.}$
.							.
.							.
.							.
r	T_{r1}	T_{r2}	...	T_{rj}	...	T_{rc}	$O_{r.}$
Total	$D_{.1}$	$D_{.2}$	...	$D_{.j}$	...	$D_{.c}$	S

$$\sum_{i=1}^{r}\sum_{j=1}^{c}[(T_{ij} - T^*_{ij})^2/T^*_{ij}] \quad (2)$$

where T_{ij} is the observed flow from origin i to destination j, follows a chi-square distribution with (r - 1) (c - 1) degrees of freedom (1).

The fundamental weakness in Equation 1 is that OD flows are estimated using only the total trips for each origin and destination. Consequently, for two OD matrices that may have completely different cell values but the same trip totals for each origin and destination, Equation 1 would produce two identical matrices. Intuitively, then, one would expect sizeable errors in the T^*_{ij}'s obtained from Equation 1.

Estimation Procedure

The estimation method given by Equation 1, though appealing in its simplicity, is not, unfortunately, directly applicable to simple linear systems with flow in one direction. For a typical freeway OD matrix for flow in one direction, such as that shown in Table 2, the basic contingency-table approach to estimating the T_{ij}'s requires modification. The need for this modification stems from the fact that some trip interchanges are not permissible (i.e., the resulting OD matrix is not complete). The matrix resulting from such an arrangement, then, is roughly upper triangular. The structure of such a matrix

TABLE 2 Typical OD Matrix for Freeway with Flow in One Direction

Off Ramp \ On Ramp	1	2	...	n-2	n-1	n	Total
1		T_{12}	...	$T_{1,n-2}$	$T_{1,n-1}$	T_{1n}	$O_{1.}$
2			...	$T_{2,n-2}$	$T_{2,n-1}$	T_{2n}	$O_{2.}$
.							
.				...	...	...	...
.							
n-2					$T_{n-2,n-1}$	$T_{n-2,n}$	$O_{(n-2).}$
n-1						$T_{n-1,n}$	$O_{(n-1).}$
n							
Total		$D_{.2}$	...	$D_{.(n-2)}$	$D_{.(n-1)}$	$D_{.n}$	

also implies that for those freeway sections where each on ramp is accompanied by an off ramp a few hundred yards downstream, few vehicles will enter the freeway just to exit immediately downstream (1).

Despite these minor complications, a solution to the matrix remains fairly straightforward. For example, for the matrix shown in Table 2, it is clear that $T^*_{12} = D_{.2}$ and $T^*_{n-1,n} = O_{(n-1)}$. Once T^*_{12} and $T^*_{n-1,n}$ have been found the corresponding row and column sums can be adjusted and a solution sought for the reduced matrix. Following this approach, $T^*_{n-2,n-1}$ and $T^*_{n-2,n}$ can be obtained from

$$T^*_{n-2,n-1} = [O_{(n-2)} \cdot D_{.(n-1)}]/[D_{.(n-1)} + D'_{.n}] \quad (3)$$

$$T^*_{n-2,n} = [O_{(n-2)} \cdot D'_{.n}]/[D_{.(n-1)} + D'_{.n}] \quad (4)$$

where $D'_{.n} = D_{.n} - T^*_{n-1,n}$. Except for minor deviations, Equations 3 and 4 follow the notation and formulation given by Hauer and Shin (1). By subtracting the T^*_{ij}'s for row n - 2 from the column sums $D_{.(n-1)}$ and $D'_{.n}$, one can proceed to obtain the T^*_{ij}'s for row n - 3 in the same fashion (1). Note in Equations 3 and 4 that only the second term in the numerator varies across the row. Consequently, the ratio of the current row sum to the current destination grand total can be stored in calculator memory and used as a distribution factor to calculate the T^*_{ij}'s from the corresponding current column sums. It is precisely these transition probabilities (8) that SYNODM uses to apportion entry volumes among the downstream ramps.

Application of the algorithm, then, involves the repetition of two basic steps. In the first step, the T^*_{ij}'s in the current bottom row of the matrix are obtained by multiplying the corresponding current row and column totals and dividing by the current destination grand total. In the second step, the bottom row with the newly obtained estimates is deleted from the matrix and the affected column totals are adjusted accordingly (1).

To summarize, SYNODM distributes the total trips of each origin by working from upstream destinations to downstream destinations using the currently unassigned trips at upstream origins to satisfy the current destination totals (8). That is, total entry volumes are apportioned to downstream ramps in accord with flows leaving the downstream ramps (1,8). Basically, SYNODM estimates the T_{ij}'s by solving a portion of a complete matrix. For those situations where $\Sigma O_{i.}$ does not equal $\Sigma D_{.j}$, SYNODM allows the user to specify whether the distribution matrix is to be balanced with respect to input or output volumes (8).

Although the SYNODM computer program does not (in its present form) permit the user to specify any known T_{ij}'s, such as might have been obtained from a limited field survey, it is possible to manually incorporate such information into the estimation method. Known T_{ij}'s can be incorporated into the estimation method by simply adjusting the appropriate current row and column totals before applying the algorithm. A manual application of the basic SYNODM algorithm that incorporates limited OD data into the estimation method is presented later in the paper.

APPLICATION

Field Data Collection

OD data for the Katy Freeway (I-10W) in Houston, Texas (see Figure 1), were used to evaluate the SYNODM algorithm. The OD data for the Katy Freeway (9) were collected by the driver postcard survey

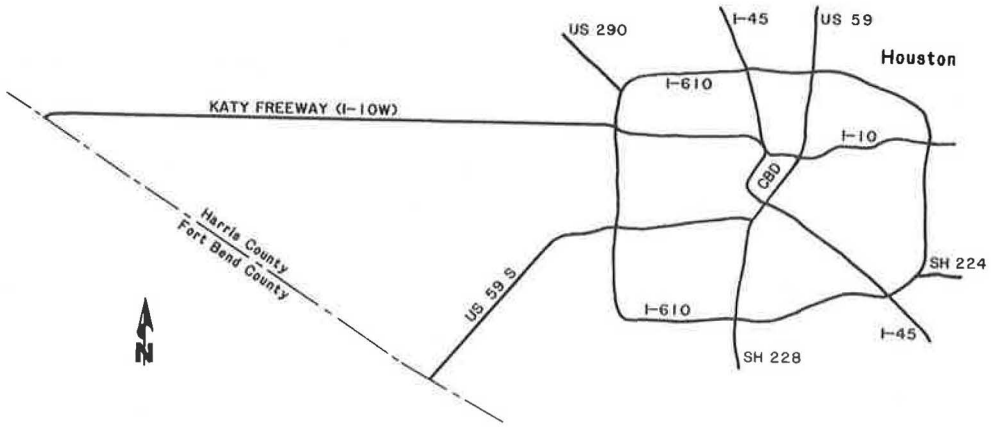

FIGURE 1 Freeways in the Houston region.

method. The survey was designed to sample 50 percent of entrance-ramp traffic for the period 6:00 a.m. to 6:30 p.m. The ramp survey stations from which the samples were drawn are shown in Figure 2.

To estimate total traffic volumes and OD distributions from the sample data, the sample percentages of ramp-to-ramp flows were applied to automatic counts of entrance-ramp volumes made on days just before the field survey. These presurvey volumes were used to compensate for possible diversions attributable to the presence of the survey crews. The sample percentages (distribution factors) that were applied to the presurvey ramp volumes are, of course, analogous to the transition probabilities used by SYNODM.

In order to present a simplified illustration of the SYNODM algorithm, the OD tables developed from the field survey were compressed into a 6 x 6 table representing a 3-mile section of the freeway for the morning peak period (6:00 to 9:00 a.m.)

Summary of Results

A comparison of observed OD distributions (i.e., postcard data) and the OD distributions estimated by SYNODM from row and column sums only is presented in Table 3. As can be seen from Table 3, SYNODM has overestimated the trip interchanges for those off ramps immediately downstream of each on ramp. This tendency to overestimate is particularly evident for

TABLE 3 Observed and Estimated OD Distributions, Eastbound Katy Freeway (6:00 to 9:00 a.m.)

On-Ramp \ Off-Ramp	1 Wilcrest	2 West Belt	3 Gessner	4 Bunker Hill	5 Blalock	6 "Farther East"	TOTAL
1 "Farther West"	822 (822) 0.000	1713 (1428) 56.880	1358 (1048) 91.698	536 (443) 19.524	501 (445) 7.047	7256 (8000) 69.192	12186
2 Wilcrest		22 (307) 264.577	78 (226) 96.920	84 (95) 1.274	139 (96) 19.260	2123 (1722) 93.380	2446
3 West Belt			3 (166) 160.054	51 (70) 5.157	80 (70) 1.429	1437 (1265) 23.387	1571
4 Gessner				18 (81) 49.000	26 (81) 37.346	1578 (1460) 9.537	1622
5 Bunker Hill					9 (62) 45.306	1166 (1113) 2.524	1175
6 Blalock						1997 (1997) 0.000	1997
TOTAL	822	1735	1439	689	755	15557	

XXX Observed
(XXX) Estimated (SYNODM)
XX.XXX Cell Chi-Square

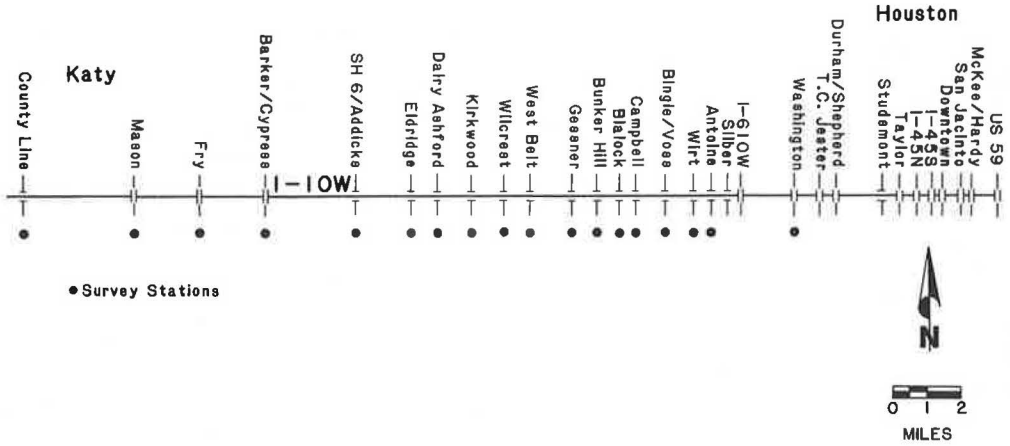

FIGURE 2 OD survey location on Katy Freeway.

those off ramps with low observed interchange volumes. As shown in Figure 3, SYNODM has generally

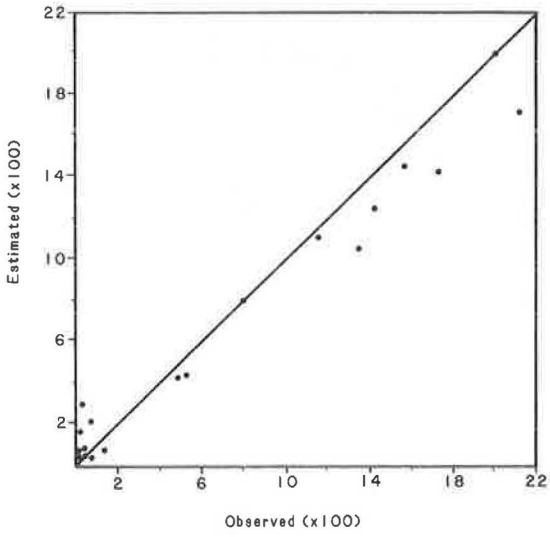

FIGURE 3 Correspondence of observed and estimated T_{ij}'s.

overestimated the number of short trips and underestimated the number of long trips.

The value of the chi-square statistic for the trip table (i.e., 1,053) is clearly much higher than would have been expected if all trip interchanges were equally likely. There appears to be some regularity in travel behavior or some unique attributes of the trip makers that are not accounted for in the algorithm. Intuitively, this conclusion seems entirely reasonable. There are several major employment centers downstream of the portion of the freeway depicted in Table 3. Because the trips represented by Table 3 are primarily work trips, one would expect the majority of the trips originating at the upstream end of the table to exit at or near the downstream end of the table. Approximately 64 percent of the chi-square statistic comes from the five cells associated with ramps upstream of the Bunker Hill exit, which is consistent with this reasoning.

In addition to the cell chi-square statistics shown in Table 3, an average absolute trip error, defined as follows (7), was also calculated:

$$\text{Avg trip error} = \sum_{i=1}^{r} \sum_{j=1}^{c} |T^*_{ij} - T_{ij}|/N \qquad (5)$$

where N is the number of error values.

Whether the resulting average trip error of 147 is excessive or not cannot be answered in general. However, when compared with the low ramp volumes observed at the upstream end of the trip table (Table 3), the average trip error illustrates the relative magnitude of SYNODM's tendency to overestimate the number of short trips.

SYNODM's tendency to overestimate short trips suggests that it might be worthwhile, in terms of improving the precision of the estimation method, to conduct a field survey of the first few upstream ramps. For example, suppose a lights-on survey of the distribution of farther-west traffic among the first three off ramps was conducted. Assume that the survey resulted in T^*_{12} and T^*_{13} values of 1,713 and 1,358, respectively (i.e., identical to those obtained from the postcard survey). Utilizing the total exit volumes shown in Table 3 (i.e., assuming that these values represent ramp counts), estimates of T_{11}, T_{22}, and T_{66} could be obtained by inspection. Estimates of T_{44} through T_{56} would be unaffected by this additional information and would remain as given in Table 3. The trip table reflecting the incorporation of the T_{ij}*'s obtained from the hypothetical field survey is shown in Table 4. The T^*_{23} through T^*_{26} and T^*_{14} through T^*_{16} were obtained by subtracting the known T_{ij}'s from the appropriate row sums and applying Equations 3 and 4. Incorporation of the hypothetical limited field data into the estimation method reduced the average absolute trip error from 147 (for the basic SYNODM program) to 42.

TABLE 4 Observed and Estimated OD Distributions, Eastbound Katy Freeway (6:00 to 9:00 a.m.): Incorporation of Limited Survey Data

Off-Ramp / On-Ramp	1 Wilcrest	2 West Belt	3 Gessner	4 Bunker Hill	5 Blalock	6 "Farther East"	TOTAL
1 "Farther West"	822 ---	1713 ----	1358 ----	--- (413)	--- (416)	---- (7464)	12186 8293[a]
2 Wilcrest		22 --	-- (71)	--- (117)	--- (118)	---- (2118)	2446 2353[a]
3 West Belt			-- (10)	-- (78)	-- (78)	---- (1405)	1571
4 Gessner				-- (81)	-- (81)	---- (1460)	1622
5 Bunker Hill					-- (62)	---- (1113)	1175
6 Blalock						1997 ----	1997
TOTAL	822	1735	1439	689	755	15557	
[a]"Adjusted" row sums.			81	608	612	10987	12288[b]
[b]O'(.n) for row 3.				530	534	9582	10646[c]
[c]O'(.n) for row 2.							

XXX (XXX) Observed Estimated

CONCLUSIONS

Given the limited scope of the study presented in this paper, the question of whether the accuracy of SYNODM is sufficient in some specific case cannot be answered in general. However, a few points deserve note. These observations are offered because they may provide some direction for future research in this problem.

First, for the simple example of an urban freeway during a peak time period with fairly stable OD patterns, the SYNODM algorithm produced reasonably accurate estimates of freeway ramp-to-ramp traffic flows. For freeway planning studies employing macroscopic simulation models, SYNODM's estimates of OD flows may suffice in the absence of any actual OD data. Eldor (8), for example, has shown that the operational measures of effectiveness output by the FREQ simulation programs are not overly sensitive to errors of the magnitudes typically encountered in SYNODM estimates.

Second, incorporation of limited OD data into the estimation method produced a substantial reduction in the average trip error. This suggests that even with the fundamental weaknesses that characterize simplistic algorithms like SYNODM, they could be useful in expanding limited survey data.

In any case, additional research on OD estimation procedures is needed. In this regard, Nihan's (7) incorporation of a trip impedance factor into the estimation method has produced some encouraging results.

ACKNOWLEDGMENT

The research reported in this paper was conducted as part of an ongoing research project entitled West Loop Authorized Vehicle Lane Conceptual Planning and Design sponsored by the Texas State Department of Highways and Public Transportation.

REFERENCES

1. E. Hauer and B.T. Shin. Estimation of Origin-Destination Matrices from Observed Flows: Simple Systems. Research Report 65. Department of Civil Engineering, University of Toronto, Toronto, Canada, March 1980.
2. P.P. Jovanis, W.-K. Yip, and A.D. May. FREQ6PE: A Freeway Priority Entry Control Simulation Model. Research Report UCB-ITS-RR-78-9. Institute of Transportation Studies, University of California, Berkeley, 1978.
3. G.H. English, H. Joiner, and W. Neyland. Traffic Study of Southwest Freeway (US 59) Between Shepherd Drive and Bissonnet Street in Houston, Texas Using the FREQ3CP Simulation Model. Texas State Department of Highways and Public Transportation, Houston, Oct. 1978.
4. R.W. Stokes, J.M. Mounce, D.E. Morris, and R.L. Peterson. Simulation Analyses of Proposed Improvements for the Southwest Freeway (US 59), Houston; Part 1: Long-Term, Capital-Intensive Improvements. Final Report. Texas Transportation Institute, Texas A&M University System, College Station, Aug. 1982.
5. K. Banning, R. Stokes, J. Mounce, and D. Morris. Simulation Analyses of Proposed Improvements for the Eastex Freeway (US 59N), Houston. Texas Transportation Institute, Texas A&M University System, College Station, Feb. 1983.
6. L.G. Nungesser, D.L. Christiansen, and A.J. Ballard. Priority Treatment for High-Occupancy Vehicles on I-10 and I-35 in San Antonio, Texas: A Needs Assessment. Texas Transportation Institute, Texas A&M University System, College Station, March 1983.
7. N.L. Nihan. Procedure for Estimating Freeway Trip Tables. In Transportation Research Record 895, TRB, National Research Council, Washington, D.C., 1982, pp. 1-5.
8. M. Eldor. Integrated Real-Time Freeway On-Ramp Control Strategies. Ph.D. dissertation. Institute of Transportation Studies, University of California, Berkeley, 1976.
9. R.W. Stokes, D. Morris, and J.M. Mounce. Katy Freeway (I-10W) Origin-Destination Study. Texas Transportation Institute, Texas A&M University System, College Station, Jan. 1982.

The views and conclusions expressed in this paper are those of the authors. They are not necessarily those of the Texas State Department of Highways and Public Transportation.

Estimation of Origin-Destination Matrices with Constrained Regression

CHRIS HENDRICKSON and SUE McNEIL

ABSTRACT

The use of constrained generalized least-squares (CGLS) regression to estimate origin-destination travel matrices from aggregate data is described. The CGLS method does not require general surveys but allows any available data to be included. Variances of matrix entry estimates can be estimated and used as measures of uncertainty or to suggest additional sampling strategy. Two case studies are described from applications to data from Portland, Oregon. The first involves expanding a matrix of transit work trips to all transit trips. Second, a gravity-type model of trip distribution for all work trips is estimated. Comparisons are made with other estimation methods with respect to accuracy, computational effort, and the use of uncertainty measures.

Origin-destination (OD) matrices representing the number of trips between zones or locations in a particular time period are widely used in transpor-

tation systems analysis. In particular, they are an important intermediate stage in the Urban Transportation Planning System (UTPS) (1). They are also used for modal operations planning, as in transit route planning [see, for example, discussions by Sheffi and Sugiyama (2) and Turnquist et al. (3)].

Surveys are widely used to estimate such matrices. For example, entries may be obtained from estimates of household travel between specified zones based on home interview surveys. Unfortunately, such surveys are time consuming and expensive. Moreover, updating and expanding survey-based matrices are common problems.

To overcome the costs and delays of general surveys, several alternative methods are used to obtain estimates of matrix entries. Nonsurvey matrix estimation methods range from applying a constant factor to update an existing matrix to more complicated ab initio estimates using, for example, gravity models. More recently, matrix entry estimation methods that include small-sample or traffic-count data or both have been developed. These include entropy maximization (4) and maximum-likelihood methods (5,6). For a survey of these methods see the paper by Chan et al. (7).

Methods used to estimate OD matrices should meet the following requirements. They should be able to

1. Include any available survey-derived or aggregate data,
2. Include socioeconomic characteristics and attitude such as travel time between zones,
3. Provide some measure of the reliability of the estimates, and
4. Account for errors in the available data.

Based on these criteria, a regression formulation is developed of the matrix entry estimation problem that is equivalent to a quadratic programming formulation (3,8). Although this formulation has been applied to the estimation of input-output tables (9) and Markov transition probability matrices (10), in the applications to transportation planning the statistical properties of the estimates have not been explored and these applications have been less general. The method is able to include any linear constraint on the matrix entries in the formulation as well as a linear-in-the-parameters distribution function with known or unknown parameters. The resulting estimates are best linear unbiased estimates and an estimate of their variance-covariance matrix may be obtained.

The plan of the remainder of this paper is as follows. The estimation problem is formulated as a regression problem and the assumptions that are required are described. The estimation method is applied to the problem of expanding a matrix of transit work trips to all transit trips for Portland, Oregon. A work trip OD matrix for Portland is estimated using a distribution function with unknown parameters. Last, the advantages and disadvantages of the regression formulation are discussed and some conclusions are presented.

ESTIMATION PROBLEM FORMULATION

The objective of matrix entry estimation is to form a consistent estimate ($\tilde{Q}$) of an actual but unknown matrix Q of size nm. It is required that the estimate $\tilde{Q}$ be consistent with all relevant information, in the sense that the estimates do not conflict with any available data known a priori.

To formulate the estimation problem, let the matrices Q and $\tilde{Q}$ be rearranged to form vectors $\underline{q}$ and $\underline{\tilde{q}}$ (of size nm) with the columns of the matrices arranged end to end. (Throughout this paper capital letters will be used to represent matrices and underlined lower-case letters to represent vectors.) It is assumed that a linear-in-the-parameters functional relationship exists between some set of attributes X and the values of matrix entries $X\underline{a}$, where $\underline{a}$ is an h vector of either known or unknown parameters. For example, $X\underline{a}$ might be a linear form of a gravity model. With these assumptions a regression equation is formulated as follows:

$$\underline{q} = X\underline{a} + \underline{\varepsilon} \tag{1}$$

where $\underline{\varepsilon}$ is an nm vector of errors that are described in the following.

It is often the case that observations or constraints exist on the sums of some subsets of the elements of q. For example, if Q is an OD matrix, the row totals are the number of trips generated at each origin and the column totals are the number of trips attracted to each destination. Alternatively, the number of trips crossing a corridor may be known. Accordingly, it is assumed that some linear constraints exist on values of $\underline{q}$:

$$\underline{r} = A\underline{q} \tag{2}$$

where $\underline{r}$ is a p vector of constraint totals and A is a p x (mn) incidence matrix of zeros and weights (which will usually be 1's). Known values of $\underline{q}$ can be constrained to equal their values by a simple constraint $r_i = q_i$. The constraint totals $\underline{r}$ may contain measurement errors that may be accounted for in the estimation procedure (11). The existence of inequality constraints is also of interest in many cases, but such constraints introduce some complications. For simplicity it is also assumed that X is a matrix of fixed numbers of full column rank and A is of full row rank. Thus, constraints that are linear combinations of other constraints should be eliminated. For example, if constraints exist on each row and column of Q, one of these n + m constraints is a linear combination of the others and should be discarded because it adds no information for estimation purposes.

The special case in which all $\underline{a}$ are known is of particular interest, and it will be considered first. This case would occur if hypothesized or base-period values of $\underline{q}$ were available rather than the functional form $X\underline{a}$. Such base values, denoted $\underline{y}$, may be obtained from a gravity model or be outdated estimates. In any case, the estimates are generally not consistent with the constraints given in Equation 2. Consistent estimates are obtained by assuming that the base matrix vector $\underline{y}$ is equal to the true vector $\underline{q}$ plus an error term:

$$\underline{y} = I\underline{q} + \underline{\varepsilon} \tag{3}$$

As before, a set of constraints (Equation 2) on $\underline{q}$ is imposed.

The standard assumptions of the generalized least-squares regression model concerning the errors in regression Equation 1 are introduced:

$$E[\underline{\varepsilon}] = \underline{0} \tag{4}$$

$$E[\underline{\varepsilon}\ \underline{\varepsilon}'] = \sigma^2 V \tag{5}$$

so that the expected value of entries in $\underline{\varepsilon}$ is zero and the covariance matrix of $\underline{\varepsilon}$ is a constant (σ^2) times a known nonsingular matrix V, which is discussed in the following. In this case, the model given by regression Equation 3, the linear constraints (Equation 2), and the assumptions about the

error terms (Equations 4 and 5) are identical to that of constrained generalized least-squares regression (CGLS) (12). The CGLS estimate of $\underline{q}$ is

$$\tilde{\underline{q}} = \underline{y} + VA'(AVA')^{-1}(\underline{r} - A\underline{y}) \quad (6)$$

The estimator $\tilde{\underline{q}}$ produces best linear unbiased estimates of $\underline{q}$ and has a covariance matrix:

$$COV[\tilde{\underline{q}}] = \sigma^2[V - VA'(AVA')^{-1}AV] \quad (7)$$

where an unbiased estimator of σ^2 is

$$s^2 = [\underline{y} - \tilde{\underline{q}}]'V^{-1}[\underline{y} - \tilde{\underline{q}}]/p$$

where p is the number of constraints. Assuming normality of observations, confidence intervals and hypothesis testing may be performed using Equations 6 and 7. The square roots of the diagonal elements of $COV[\tilde{\underline{q}}]$ are the standard errors of estimated parameters that are normally reported in regression.

As with other regression models, this special case can be formulated as a quadratic programming problem:

$$P1: \text{Min}_{\underline{q}}[(\underline{q} - \underline{y})'V^{-1}(\underline{q} - \underline{y}) \mid \underline{r} = A\underline{q}] \quad (8)$$

in which the weighted sum of squared errors is minimized subject to a set of constraints.

For the general case, parameters $\underline{a}$ of regression Equation 1 are to be estimated in addition to the matrix entries $\tilde{\underline{q}}$. Estimates of the model parameters $\tilde{\underline{q}}$ and of the matrix entries $\tilde{\underline{q}}$ can be obtained (13) as follows. Substituting Equation 1 into Equation 2 yields

$$\begin{aligned}\underline{r} &= A(X\underline{a} + \underline{\varepsilon}) \\ &= AX\underline{a} + A\underline{\varepsilon} \\ &= X^*\underline{a} + \underline{\varepsilon}^*\end{aligned} \quad (9)$$

This is a linear-regression model with observations $\underline{r}$, with $X^* = AX$, and a random vector $\varepsilon^* = A\varepsilon$. If it is assumed that the assumptions about the error term in Equation 1 are given by Equations 4 and 5, the expectation of the error term is zero, because $E[\varepsilon^*] = E[A\underline{\varepsilon}] = \underline{0}$, from Equation 4. Also, the covariance matrix of error terms is $COV[\underline{\varepsilon}^*] = E[A\underline{\varepsilon}(A\varepsilon)'] = AE[\underline{\varepsilon}\underline{\varepsilon}']A' = \sigma^2AVA'$. Assuming that AVA' is nonsingular, this regression model fulfills the assumptions of Equations 3 and 4 with $V^* = AVA'$, and is a generalized least-squares model.

Following Theil (12), the GLS estimator of $\underline{a}$ is

$$\begin{aligned}\tilde{\underline{a}} &= [(AX)'(AVA')^{-1}(AX)]^{-1}(AX)'(AVA')^{-1}\underline{r} \\ &= [X'A'(AVA')^{-1}AX]^{-1}X'A'(AVA)^{-1}\underline{r}\end{aligned} \quad (10)$$

which has the normal regression properties of being a best linear unbiased estimate of $\underline{a}$. Moreover, the covariance matrix of $\tilde{\underline{a}}$ is

$$COV[\tilde{\underline{a}}] = \sigma^2\,[X'A'(AVA')^{-1}AX]^{-1} \quad (11)$$

where an unbiased estimator of σ^2 is

$$s^2 = (\underline{r} - AX\tilde{\underline{a}})'(\underline{r} - AX\tilde{\underline{a}})/(p - h)$$

The predictor $X\tilde{\underline{a}}$ provides estimates of $\underline{q}$ that can then be used in the CGLS formulation to obtain best linear unbiased estimates of the matrix entries that are consistent with the constraints.

In passing, it might also be noted that the GLS model could be equivalently formulated as a quadratic programming problem. In this case, it is desired to minimize the weighted sum of squared errors, subject to the known constraints, by choosing values of $\underline{q} + \underline{a}$:

$$P1: \text{Min}_{\underline{q},\underline{a}}[(\underline{q} - X\underline{a})'V^{-1}(\underline{q} - X\underline{a}) \mid \underline{r} = A\underline{q}] \quad (12)$$

This programming problem has the same solution for $\underline{a}$ as the problem in Equation 8 [assuming, of course that $(AVA')^{-1}$ exists].

Thus, the estimation problem can be formulated as either a generalized least-squares regression problem or a quadratic programming problem. A third interpretation as a Bayesian estimation problem with a quadratic loss function is also possible (11).

Solution of the estimation equations (Equation 6 or 10) can be performed by any general matrix inversion and multiplication package. Note that estimation of the covariance matrix (Equation 7 or 11) can use the matrix inversion required for calculation of the actual estimates. Alternatively, specialized techniques such as conjugate gradient algorithms can be employed to speed calculations for large problems. Sparse matrix calculation methods can also be used (14).

Before turning to examples, the specification of the error term $\underline{\varepsilon}$ deserves mention. It is assumed that the covariance matrix of error terms is nonsingular and specified in advance. Here some possible specifications for the error term covariance matrix (V) are considered. This is an important subject because specification of the covariance matrix will affect estimates and should reflect the analyst's beliefs concerning errors. In many applications of quadratic estimation techniques, the assumption of independent and identically distributed error terms seems to be made implicitly and without due consideration, solely because of analytical convenience.

The classical assumption in least-squares regression is that error terms are independently and identically distributed. In this case, the matrix V is the identity matrix ($V = I$) and $V[\underline{\varepsilon}] = \sigma^2I$. For this common case, the generalized covariance matrix $AVA' = AIA' = AA'$ must be nonsingular because A is of full row rank. Consequently, estimates of $\tilde{\underline{a}}$ can almost always be calculated from Equation 10; the only exception occurs in the unlikely case that the solution matrix $[X'A'(AA')^{-1}AX]$ is singular.

For physical matrix entries, an appealing alternative to this classical assumption is that error variances are proportional to the corresponding entry in the estimate $\underline{y}$. Thus, entries with small expected value would have small variance, as might be expected in many situations. This assumption results in a weighting function in P1 similar to that of a chi-square goodness-of-fit test:

$$\begin{aligned}P2: \text{Min}_{\underline{q}}&[(\underline{q} - \underline{y})'V^{-1}(\underline{q} - \underline{y}) \mid \underline{r} = A\underline{q}] \\ &= \left[\sum_{k=1}^{nm} (q_k - y_k)^2/y_k \mid \underline{r} = A\underline{q}\right]\end{aligned} \quad (13)$$

where q_k and y_k are the kth entries in $\underline{y}$ and $\underline{q}$, respectively.

One practical advantage of a chi-square weighting scheme of this sort is that the likelihood of obtaining negative entry estimates is much reduced from the case of the classical assumption of constant variance. This result is due to an assumption of small variance for entries of $\underline{q}$ having small initial estimates $\underline{y}$. This result may often remove the necessity to introduce inequality constraints to ensure that entry estimates are nonnegative.

These results are now applied to the problem of expanding an OD matrix.

EXAMPLE 1: EXPANSION OF A TRANSIT WORK TRIP OD TRAVEL MATRIX TO ALL TRANSIT TRIPS

The purpose of the application problems in this and the next section is to illustrate quadratic matrix entry estimation methods. The problems are intended to demonstrate the feasibility and flexibility of the methods developed in the preceding section. The applications also provide some experience with computational problems and illustrate the formulations that might be appropriate for different applications and how these formulations might vary with the quality and quantity of data available.

In the examples the quadratic matrix entry estimation method is applied and evaluated. For validation and evaluation, matrices obtained from surveys are used for comparison with the quadratic estimates. For estimation, a base matrix or data to estimate a matrix entry function as well as some constraints are used. The more constraints, the larger the degrees of freedom and the more reliable are the estimates. The first example expands a transit OD matrix from Portland, Oregon. The second estimates ab initio an OD matrix for work trips in Portland.

The first application is an expansion of the transit work trip matrix to an OD matrix for all transit trips on the Portland Transit Authority's system (Tri-Met). In practice, the transit work trip matrix might have been obtained from data used for another purpose, inferred from census data, or estimated from journey-to-work survey data. The constraints required to estimate the complete matrix might be the result of data routinely collected for each route and some simple surveys. The expansion problem might be a problem facing a transit authority that does not wish to administer a general survey but would like to estimate an OD matrix from such available data.

The OD matrix used as a base matrix and the matrix to be estimated in this application differ slightly from the usual OD matrices. Instead of using OD zones, an OD matrix with transit routes as origins and four types of destinations is estimated. The destinations are defined in two categories--transfer or nontransfer trips. Then within these categories the place where the person alights from the bus is defined as in the central business district (CBD) or outside the CBD, giving a total of four destinations. The transfer trips are particularly interesting to transit planners because although route volumes are generally known, the origin and number of transfer trips can rarely be estimated unless systemwide on-board surveys have been undertaken.

Therefore, the OD matrices in this example are estimated by origin route and four defined destinations. There are 71 routes, so the matrix to be estimated is 71 by 4 (with 284 entries). The ijth element in the matrix represents the number of people beginning their trip on route i with destination j. The Tri-Met route numbers associated with each row (origin) are described elsewhere (11). The destinations are defined to be

j = 1 if the destination is outside the CBD for a nontransfer trip,
j = 2 if the destination is in the CBD for a nontransfer trip,
j = 3 if the destination is outside the CBD for a transfer trip, and
j = 4 if the destination is in the CBD for a transfer trip.

Five different formulations were used to expand the Portland transit OD matrix for work trips to all trips. The five formulations permit a comparison of the estimates using the biproportional method (described in the following) and an ad hoc procedure with estimates obtained using the quadratic method as well as consideration of the effect of additional constraints. For comparison and basic data, a matrix derived from 12 percent on-board survey data is used.

The first formulation is an ad hoc procedure for expanding the matrix and was devised so that the entry estimate equals the matrix entries for work trips multiplied by a factor equal to the total number of all transit trips divided by the number of work transit trips. Thus, this is simply a constant expansion. The expansion factor in this case was 1.683.

The second formulation is a quadratic method using a chi-square objective function, implying that changes in the base matrix are likely to be more variable (larger) if the relative magnitude of the entry is larger. This is believed to be a reasonable assumption. As base values in the formulation, the ad hoc estimates obtained with the first formulation are used. The constraints are row and column totals representing the total number that use a route and the total number that have each of the four destinations, respectively. These data would normally be available from aggregate route counts and some simple surveys.

The third formulation is a biproportional estimation problem. The biproportional estimation is also known as the Fratar method, Bregman's balancing method, the Furness iterative procedure, or the RAS method. The base matrix used in the biproportional formulation is the work trip matrix. Again, row and column totals are assumed known. The biproportional method factors up each entry by a row and column factor so that the constraints are met. This method is a common alternative to least-squares methods.

The fourth and fifth formulations use additional constraints in the form of the total number who transfer on each route. The fourth and fifth formulations are quadratic and biproportional formulations as in the second and third formulations but with the additional constraints. The quadratic formulation simply adds the additional constraints to the first formulation with the row and column totals. To formulate the biproportional problem with the additional constraints, the problem can be split into two independent biproportional problems. The first problem operates on columns 1 and 3 of the matrix and the second problem operates on columns 2 and 4.

The formulations are summarized as follows, where w_{ij} is the ijth entry in the matrix of work trips, t_i is the number of transfers from route i, $\underline{u}$ and $\underline{v}$ are row and column totals, and all other notation has been defined previously.

Ad hoc method, constant factor expansion:

AH: $q_{ij} = aw_{ij}$

where $a = \Sigma_{ij}q_{ij}/\Sigma_{ij}w_{ij}$ with $\Sigma_{ij}q_{ij}$ known from passenger counts.

Quadratic with row and column constraints:

Q1: Min $(\underline{q} - a\underline{w})'V^{-1}(\underline{q} - a\underline{w})$

subject to $\Sigma_i q_{ij} = v_j$ and $\Sigma_j q_{ij} = u_i$.

Biproportional with row and column constraints:

B1: $Q = B\ W\ C$

subject to $\Sigma_i q_{ij} = v_j$ and $\Sigma_j q_{ij} = u_i$. (Note that B and C are diagonal matrices of unknown factors.)

Quadratic with row and column totals and total number of transfers on each route (identical to Q1 but with added constraints):

$$Q2: q_{ix2} + q_{i4} = t_i \; \forall i$$

Biproportional with row and column totals and the total number of transfers on each route:

$$B2I: q_{i1} = b_i w_{i1} c_1 \quad q_{i3} = b_i w_{i3} c_3$$

subject to $\sum_i q_{ij} = v_1$

$$\sum_i q_{i3} = v_3$$

$$q_{i1} + q_{i3} = u_i - t_i$$

$$B2II: q_{i2} = b_i^* w_{i2} c_2 \quad q_{i4} = b_i^* w_{i4} c_4$$

subject to $\sum_i q_{i2} = v_2$

$$\sum_i q_{i4} = v_4$$

$$q_{i2} + q_{i4} = t_i$$

The chi-square formulation Q1 was estimated using a general-purpose quadratic programming package, and both formulations Q1 and Q2 were estimated using FORTRAN programs that used matrix manipulation methods. The specific quadratic programming package used for estimation uses Lemke's method and the matrix manipulation method used a matrix factorization routine from the International Mathematical and Statistical Library (IMSL) package. Similar results were obtained for formulation Q1 using both methods, and minor differences (less than 1 percent) can be attributed to rounding errors. The quadratic programming package has the advantage that the elements may be constrained to be nonnegative and inequality constraints may be included. for example, it is known that each entry in the estimated matrix of all trips has to be greater than or equal to the corresponding entry in the matrix for work trips. This constraint was not applied, because only 5 of the 284 entries (less than 2 percent) violated the constraint. The matrix inversion routines have the advantage that the variance of the estimates can easily be calculated.

Table 1 shows the central processing unit (CPU) computation time on a DEC-20 computer for the matrix manipulation routines and the quadratic programming package. The amounts of time include setting up the problem from the same basic data set. It is indicated in the table that the computation time using the quadratic programming package is significantly greater than that using the matrix manipulation routines. The matrix manipulation routines also calculate variance estimates.

The biproportional estimates were obtained iteratively. The program terminated after 20 iterations or when all the row and column totals were within 1 percent of the constraints. Table 1 shows the computation time for obtaining the biproportional estimates on a DEC-20 to be about one-twentieth of the time required to obtain the quadratic estimates with the matrix inversion routines.

The evaluation of the estimation results presents a problem, because there is no unique method for comparing two or more methods. Several testing routines are possible (11,15,16). Table 2 gives three aggregate measures of estimation errors. Each is calculated by comparison with the on-board survey results. As can be seen, the quadratic and biproportional methods are comparable in accuracy for each case and generally superior to the ad hoc estimates. Adding additional information in the form of transfer totals improves the accuracy of estimates, as expected.

The similarity between the quadratic and biproportional results is not surprising. It has been shown (17) that the chi-square formulation is a first-order approximation to a biproportional problem.

When the estimation problem is formulated as a CGLS regression problem, the estimates can also be evaluated in terms of their uncertainty. This evaluation is based on the assumption that the errors are highly correlated with the standard deviations (variances) of the estimates. Table 3 shows some correlation coefficients between errors and measures of uncertainty for the quadratic estimates. For each set of estimates, there is a positive correlation between the standard deviation and the average absolute error and the coefficient of variation and the average relative error. If one is interested in reducing a particular type of error, entries can be selected for special surveying or data gathering on the basis of the estimated uncertainty measure that

TABLE 1 Computation Times for the Portland Expansion Example Estimates

Method	Formulation	Computation	CPU Time (sec)
Quadratic	Q1 matrix manipulation	Data preparation	14
		Estimation	41
	Q2 matrix manipulation	Data preparation	16
		Estimation	90
	Q1 quadratic programming package	Data preparation	3
		Estimation	93
Biproportional	B1	Data preparation and estimation	2
	B2	Data preparation and estimation	4

TABLE 2 Average Absolute and Relative Errors for Expanding the Portland Work Trip Transit Matrix

Method	Avg Absolute Error	Avg Relative Error[a]	Ratio of Avg Error to Avg Entry
Ad hoc	97.0	0.240	0.166
With row and column constraints			
Q1 quadratic	49.7	0.184	0.085
B1 biproportional	48.4	0.179	0.083
With row and column constraints and constraints on the number of transfers			
Q2 quadratic	35.0	0.155	0.060
B2 biproportional	34.4	0.151	0.059

[a]Calculated as the average ratio of errors to survey matrix entries excluding zero matrix entries.

TABLE 3 Correlation Coefficients Between Errors and Uncertainty Measures for Estimated Entries in Transit Matrix

Method	Uncertainty Measure	Avg Absolute Error	Avg Relative Error
Q1 quadratic with row and column constraints	Standard error	0.509	-0.542
	Coefficient of variation	-0.389	0.542
Q2 quadratic as in Q1 with constraints on transfers	Standard error	0.475	-0.187
	Coefficient of variation	-0.204	0.453

is highly correlated with that error. Using uncertainty measures in this way can be quite helpful. For example, examining the standard errors from the quadratic estimations, a planner might choose those with the largest uncertainty for special surveys.

Although the CGLS regression formulation produces similar results to the biproportional method and is computationally more expensive, there is no way to evaluate the uncertainty of biproportional estimates directly. The estimated variances of the entry estimates represent a measure of the uncertainty associated with each entry estimate and may be used to evaluate the estimates.

EXAMPLE 2: AB INITIO ESTIMATION OF A WORK TRIP OD MATRIX

This application illustrates the estimation of matrix entries ab initio using a distribution function with unknown parameters. The formulation follows that presented in the section on estimation problem formation, in which the matrix entries are assumed to be a linear function with unknown parameters (called a distribution function) of the attributes of the matrix entries. The estimation problem is then formulated as a least-squares estimation problem subject to the appropriate constraints.

One of the advantages of the quadratic estimation method in this application is that it uses any available aggregate data to estimate the matrix entries. Such data generally come from sources other than transportation surveys. For this application the required data can be obtained from sources that are likely to continue to provide the same kinds of information in the future and therefore the modeling approach is unlikely to become obsolete because the required data are not available. The methods used to illustrate ab initio estimation in this section also show how additional aggregate travel-specific data can be added to improve the estimates.

Four parameters must be specified to estimate the ab initio model: a definition of the origin and destination zones, the distribution function, the weighting matrix, and the constraints.

The origin and destination zones are defined in terms of census tracts. For computational ease, the 189 census tracts of the Portland metropolitan area are aggregated to 44 zones. The 57 census tracts that are in the Portland standard metropolitan statistical area (SMSA) but outside the metropolitan area are aggregated to four zones. The aggregation simply joins adjacent areas except where physical barriers such as rivers provide a natural division. Therefore, the OD matrix has 48 zones representing 48 x 48 = 2,304 trip elements to be estimated.

To estimate a model ab initio, a distribution function that is linear in the parameters is required to ensure convexity. A simple model based on one independent variable is used in this application. The dependent variable is a function of the number of workers resident in the origin zone, the employment in the destination zone, and the distance or travel time between the two zones. The number of residents serves as a measure of the trip-generating ability of a zone, whereas the employment serves a similar purpose with respect to the attractiveness of the zone in the sense that it measures the extent to which a zone provides employment. The distance or travel time are measures of the travel impedance, implying that large travel impedances discourage trips. The proposed model is

$$q_{ij} = a\, D_i E_j / (d_{ij}^2 + \varepsilon_{ij}) \qquad (14)$$

where

q_{ij} = number of work trips from i to j,
a = parameter to be estimated,
D_i = number of workers resident in zone i,
E_j = employment in zone j,
d_{ij} = interzonal travel impedance from zone i to zone j, and
ε_{ij} = error term.

The model is a simple gravity model.

The formulation of the model also requires definition of the structure of the weighting matrix (V). The matrix V is the matrix of weights in the quadratic objective function or the variance-covariance matrix of the error terms in the regression equation. It is assumed that V is a diagonal matrix with elements on the diagonal proportional to $D_i E_j / d_{ij}^2$ and therefore the objective function is of the chi-square type.

There are many possible constraints $R\underline{q} = \underline{r}$ that can be used to estimate the unknown parameter (a) and the matrix entries. The number and type of constraints directly affect the accuracy of the estimation. Two different sets of constraints are used to estimate the problem. Each set represents feasible constraints that can be obtained without complete areawide surveys.

The first set of constraints is the most basic, consisting of the row and column totals. Each row represents an origin zone and the row total the number of trips from an origin; therefore the row total is the number of workers residing in that origin zone (D_i). Similarly, the column totals represent the number of trips made to a zone; therefore the column total is the employment (E_j) in that zone. With this set of constraints the formulation is for a doubly constrained gravity model.

The quadratic programming problem is then formulated as follows:

$$\text{Min}_{\underline{q},a} \sum_{ij} \sum [(q_{ij} - aD_iE_j)/d_{ij}^2]^2/(D_iE_j/d_{ij}^2) \qquad (15)$$

subject to $\Sigma_i q_{ij} = E_j$ and $\Sigma_j q_{ij} = D_i$, where the total employment (E_j) and the number of workers resident (D_i) in each zone are the column and row totals, respectively. It is assumed that those with no fixed place of employment work in the zone in which they reside.

The second set of constraints is added to the first set. The additional data are obtained by using the Willamette River as a cordon; this divides the Portland metropolitan area so that there are 15 zones on the east side of the river and 33 zones on the west side.

The additional constraints might be obtained by asking those crossing the cordon in either direction if they are going to work and either where they work or where they live, thereby adding 47 linearly independent constraints. For this estimation problem the 47 constraints obtained by surveying traffic going in both directions and asking travelers where they live are used. Algebraically, the constraints are as follows:

$$\Sigma_{j \varepsilon WS} q_{ij} = C^O_{iES} \qquad i \varepsilon ES$$

$$\Sigma_{j \varepsilon ES} q_{ij} = C^O_{iWS} \qquad i \varepsilon WS \qquad (16)$$

where C^O_{iES} are the constraint totals for each of the origins on the east side, and C^O_{iWS} are the constraint totals for each of the origins on the west side.

The survey-derived matrix is from the 1976 travel-to-work supplement to the annual survey of hous-

ing for Portland. This matrix is used for comparison and to obtain much of the data for the estimation. The interzonal travel impedances were obtained from the Portland Metropolitan Service Center. Such data are commonly available within UTPS models. The peak-period travel time was used in this study because most trips to work are during the peak period and travel time is a more suitable measure of the travel impedance than distance.

Matrix manipulation programs were used to estimate the unknown parameter and matrix entries in each formulation. The results for estimation of the unknown parameter (a) were quite consistent. Using 95 or 145 constraints, an estimate for a of $\tilde{a} = 3.80 \times 10^{-4}$ was obtained, which was also the estimate obtained from regression on all 2,304 surveyed matrix entries. However, the t-statistic for the parameter ranged from 8.15 with 95 constraints to 8.76 with 142 constraints to 18.80 with all 2,304 observations. Clearly, more data reduced the level of uncertainty.

With the estimation of the unknown parameter ($\tilde{a}$), matrix entries could be calculated using Equation 14 as $q_{ij} = \tilde{a} D_i E_j / d_{ij}^2$. However, this would not ensure that the estimates q_{ij} were consistent with the known totals D_i. Accordingly, a second-stage quadratic estimation was applied using the y_{ij} as initial estimates and the constraints defined previously.

Concerning the accuracy of the entry estimates themselves, the average absolute errors were 330 and 310 for the 95- and 142-constraint cases, respectively. Average relative errors were 1.2 in both cases.

In this example, the unknown parameter and matrix entries have been estimated using a linear form of a doubly constrained gravity model and the same model with additional constraints. Each estimation involved two stages. The first stage estimated the parameter in the gravity model, subject to the constraints, using generalized least-squares regression. The second stage uses the gravity model to predict the entries in the OD matrix and reconciles them with the constraints using CGLS regression. Thus, the estimation methods described here do not require iteration to obtain parameter estimates for a linear-in-the-parameters model. They also provide a measure of the reliability of the estimates.

RELATIVE ADVANTAGES AND DISADVANTAGES OF THE REGRESSION FORMULATION

In the preceding sections, CGLS formulations of the matrix entry estimation problem have been developed and applied to the problem of estimating two different OD matrices. The applications demonstrated that the methods are computationally feasible and produced estimates comparable in terms of accuracy with estimates obtained using other techniques. Some of the specific advantages and disadvantages of the regression formulation are as follows.

In particular, the regression formulation provides some unique advantages:

1. It is flexible in terms of the information that can be included as constraints. For example, individual elements can be constrained to particular values or a linear function of several entries; that is, the sum of two entries can be assumed known.
2. It is flexible in its formulation. Distribution functions with known or unknown parameters may be included in the formulation.
3. It is able to provide a measure of the reliability of the entry estimates in terms of the variances. In all the applications presented, the variances were relatively large because of the minimal amount of data used to estimate the matrix entries. The variances in turn can be used to evaluate the entries and derive strategies for obtaining additional information to improve estimates. In the expansion example the variance was found to be highly correlated with the absolute error in the estimates and the coefficient of variation with the percentage error, indicating that such reliability measures are highly desirable. The variance measures can also be used to derive sampling strategies for obtaining additional information.
4. It is of comparable accuracy with the biproportional methods when a chi-square formulation is used. The result is not surprising because the biproportional objective function may be shown to be a first-order approximation to the chi-square objective function. In cases in which the error distribution is not of a chi-square nature, a properly specified CGLS function can be expected to be more accurate than the biproportional method.
5. It is able to account for errors in the constraint totals. The problem can be reformulated to account for errors in the constraint totals (11).
6. It is able to produce estimates that are consistent with the available data. The quadratic methods also consistently improve the estimates and reduce the variance as more data are included in the formulation.
7. It is able to estimate unknown parameters in a linear distribution function. The ab initio estimation example clearly demonstrates this and the quadratic method is the only way of doing this without survey-derived data for several individual entries.

The problems associated with the regression formulations are primarily related to the computational burden associated with calculating the estimates. The expansion example used approximately 20 times as much computer time to obtain estimates using CLGS regression compared with the biproportional method but also provided estimates of the reliability of the entry estimates. However, the reduced cost of computers and particularly the availability of personal computers reduces this problem considerably.

Another problem arises if the formulation does not explicitly include nonnegativity constraints. The applications indicated that regression formulation is unlikely to produce negative estimates when a reasonable formulation is used. In the expansion example, no entry was estimated to be negative. For the ab initio estimation problem with row and column totals, negative estimates were not a problem. For the other formulation, the negative values were constrained to zero and the problem was reestimated. Although this approach is not statistically rigorous, it is practically workable.

Although it is generally an advantage to have the flexibility obtained by allowing the analyst freedom to specify any linear function and the structure of the variance-covariance matrix of the errors, it does create some problems if the formulation is incorrectly specified. Fortunately, the impact of the specification error can be evaluated if the analyst is aware of its existence (11). For example, if the variance-covariance matrix is incorrectly specified, the estimates are unbiased but inefficient. Unfortunately, however, only linear function can be included in the formulation to ensure convexity and a global optimal solution (8).

CONCLUSIONS

In this paper methods for estimating matrix entries using generalized regression have been reviewed and developed. With these methods it is possible to include all available and relevant information, including uncertain information and judgment. As well as including all information, the form of the objective function in the formulation is flexible and the resulting estimates are best linear unbiased estimates. Associated with each entry estimate is an estimated measure of their reliability such as the variance. The variance can be used to evaluate the estimates and derive strategies for additional sampling. In applications where the quadratic method is the only possible formulation or the estimates of the entries' uncertainty are relevant, the quadratic method is a feasible estimation method and should be used. The applications provided examples of the use of the quadratic formulation and demonstrated that the techniques produce reasonable results as well as being computationally feasible for fairly large matrices.

REFERENCES

1. P.R. Stopher and A.H. Meyburg. Urban Transportation Modelling and Planning. Lexington Books, Lexington, Mass., 1975.
2. Y. Sheffi and M. Sugiyama. Optimal Bus Scheduling on a Single Route. In Transportation Research Record 895, TRB, National Research Council, Washington, D.C., 1982, pp. 46-52.
3. M.A. Turnquist, A.H. Meyburg, and S.G. Ritchie. TOP: A Transit Operations Planning Model. Technical Report DOT/RSPA/DMA-50/83-25. U.S. Department of Transportation, March 1983.
4. H.J. Van Zuylen and L.G. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. Transportation Research, Vol. 14B, No. 3, 1980, pp. 281-294.
5. H.R. Kirby and M.N. Lesse. Trip Distribution Calculations and Sampling Error: Some Theoretical Aspects. Environment and Planning A, Vol. 10, 1978, pp. 837-851.
6. U. Landau, E. Hauer, and I. Geva. Estimation of Cross-Cordon Origin-Destination Flows from Cordon Studies. In Transportation Research Record 891, TRB, National Research Council, Washington, D.C., 1983, pp. 5-10.
7. V.K. Chan, R.C. Dowling, and A.E. Willis. Deriving Origin-Destination Information from Routinely Collected Traffic Counts. Working Paper UCB-ITS-WP-80-1. Institute of Transportation Studies, University of California, Berkeley, 1980.
8. M. Carey, C. Hendrickson, and K. Siddharthan. A Method for Estimation of Origin/Destination Trip Matrices. Transportation Science, Vol. 15, No. 1, Feb. 1981, pp. 32-49.
9. F. Van Der Ploeg. Reliability and Adjustment of Sequences of Large Economic Accounting Matrices. Journal of the Royal Statistical Society, Series A, Vol. 145, Part 2, 1982, pp. 169-194.
10. T.C. Lee, G.G. Judge, and A. Zellner. Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data. North-Holland Publishing Co., Amsterdam, 1970.
11. S. McNeil. Quadratic Matrix Entry Estimation. Ph.D. thesis. Carnegie-Mellon University, Philadelphia, Pa., 1983.
12. H. Theil. Principles of Econometrics. Wiley, New York, 1971.
13. M. Carey. Constrained Estimation of Direct Demand Functions and Trip Matrices. Working Paper. Carnegie-Mellon University, Philadelphia, Pa., 1982.
14. R.P. Byron. The Estimation of Large Scale Social Accounts Matrices. Journal of the Royal Statistical Society, Series A, Vol. 141, 1978, pp. 359-367.
15. M. Butterfield and T. Mules. A Testing Routine for Evaluating Cell by Cell Accuracy in Short-Cut Regional Input-Output Tables. Journal of Regional Science, Vol. 20, No. 3, 1980, pp. 293-310.
16. G.J.D. Hewings and B.N. Janson. Exchanging Regional Input-Output Coefficients: A Reply and Further Comments. Environment and Planning A, Vol. 12, 1980, pp. 843-854.
17. M. Bacharach. Biproportional Matrices and Input-Output Change. Cambridge University Press, New York, 1970.

Characteristics of Multistop Multipurpose Travel: An Empirical Study of Trip Length

MORTON E. O'KELLY and ERIC J. MILLER

ABSTRACT

An empirical study is presented of several issues associated with the trip lengths of multistop multipurpose nonwork travel; a 2-week travel diary survey of households in Hamilton, Ontario, is used. These issues include the relationship between average stop-to-stop (link) travel times and stop purpose and number, the extent to which multistop trips follow minimum paths, and the extent to which stop sequences are ordered by distance from home. The analysis indicates that significant differences exist in average link travel times among stop purposes but not among stop number. In general, multistop trips do not follow minimum time paths. It is also found that stops nearest the trip maker's home are most likely to be the first or the last stop on a trip. Finally, some implications of these results for future modeling efforts and empirical investigations are briefly discussed.

A set of empirical studies of selected characteristics of multistop multipurpose travel is presented. The importance of multipurpose travel as a form of spatial interaction has received widespread recognition. The work of Hanson (1,pp.81-100) attests to the variety of the research efforts in this field. Furthermore, some theoretical implications of complex behavior have been raised and given a mathematical formulation (2-5). The research reported here represents a second round of empirical study along the lines of the pioneering empirical and theoretical work of Hanson (1), Jones (6), and Hensher (7).

The major difficulty facing modelers of multistop multipurpose trip making is the great complexity and variety of behavior involved. Multistop multipurpose travel is highly discretionary and flexible in terms of the number and timing of trips made, the choice of destinations visited, and the combinations and sequencings of destinations or purposes within multistop tours. Complex but rarely observed temporal (or other) constraints can affect the timing of trips, the sequencing of stops within tours, and the set of feasible destinations available to a given traveler at a given time.

The range of research issues associated with representing this complex behavior is vast. One particularly challenging and fundamental issue concerns the structure of the process involved in choosing a specific multistop multipurpose tour. That is, given the decision to make a tour, how is the composition of this tour determined? Is a set of stops chosen before leaving home or is this set chosen sequentially on a stop-by-stop basis as the tour progresses? If the set of stops on the tour is chosen before leaving home, is the order in which these stops are visited determined in any systematic fashion (e.g., nearest stop next, minimum path sequence)? How does trip length interact with the number and type of stops within a tour? (Is one likely to travel further to visit the fifth stop on a tour or the first stop? Does the distance traveled vary with the purpose of the stop?)

Several investigations are presented that are designed to search for empirical regularities in multistop multipurpose travel relating to the questions raised in the previous paragraph as a first step toward the development of an improved understanding of nonwork travel behavior and ultimately an improved capability to model such behavior. These investigations are explicitly exploratory in nature, reflecting the basic assumption that in many ways so little is known about multistop multipurpose travel behavior that it is difficult at this stage to formulate specific, testable hypotheses and that a more open-ended investigation might prove to be a more fruitful preliminary approach.

The major linkage among these analyses is that they all involve the investigation of some aspect of trip length in multistop multipurpose travel (where in all cases the measure of trip length will be off-peak stop-to-stop automobile travel times). Specifically, these investigations can be grouped into two main topics: the analysis of the interaction between stop purposes and trip length and the analysis of stop sequence within multistop tours.

INTERACTION BETWEEN STOP PURPOSES AND TRIP LENGTH

Studies of multistop multipurpose travel in terms of the types of activities that are linked together reveal several regularities (8-11). Furthermore, activity pattern analysis has now begun to model the entire time sequence of household travel (12). These studies do not explicitly emphasize the actual amounts of travel inherent in various travel patterns, although Damm (13) recognized that increased trip chaining might represent a rational response to higher fuel prices. Of particular interest in this study is the extent to which different types of activities are chained together on the same tour and the impact of various stop-purpose combinations on average travel times.

The motivation underlying this concern is threefold. First, the extent to which average travel time varies by stop-purpose combination may provide insights into the tour choice process. Second, many spatial interaction models require an estimate of average trip length as input data. It is common practice to estimate such an average trip length from survey information on single-stop home-based trips. Several studies indicate, however, that multistop multipurpose travel is common (2,7,14). Recognition of this leads to the need to reevaluate the concept of average trip length, possibly in terms of trip lengths measured over specific links of the trip chain, possibly given specific origin and destination purposes. Third, and perhaps of greatest practical importance, development of nonwork travel models is often motivated by a need to

measure the impact of various transportation policies (e.g., energy prices) on total urban travel. If such models seriously mispredict average travel times (or distances), their usefulness in such policy analyses is seriously compromised.

PATH CHOICE (SEQUENCE OF STOPS) IN MULTISTOP TOURS

In simple single-stop single-purpose trip analyses the problem of stop sequencing does not arise. When there are more than two stops on a tour, however, several paths exist to the desired destinations. How do trip makers organize these paths? Do they choose minimum time paths to their desired destinations? Do they visit the nearest (farthest) stop first (last)? Answers to these questions have implications for the conceptualization of travel. If it can be shown that tours invariably follow a minimum path, the models ought to encompass global optimization as an objective. If, on the other hand, tours do not usually follow a minimum path, a sequential approach to travel modeling may be appropriate.

The issue of ranking of stops by distance from home is also of importance because it further elucidates the problem of path choice. If all stops away from home are equally likely to be the first stop on a tour, trip making is very unstructured. If, on the other hand, most first stops on multistop tours are ranked nearest to home, trip making is likely to be structured; moreover, the activity that takes place first is less likely to be the major purpose of the tour. This last point is important; it reinforces the idea that tours may have some regularity in their ordering of stops spatially, but this does not help to discern the major purpose of a tour. Furthermore, given that it is difficult to uncover the motivation behind a tour simply by observing its progress, there are likely to be problems modeling the overall structure of the tour.

BACKGROUND

The study area for this analysis is the regional municipality of Hamilton-Wentworth, Ontario. The residents of this region were surveyed by using an areally stratified sampling scheme. About 700 households kept a diary for 2 weeks. The respondents reported all trip-making activity by the adult members of the household together with information on household characteristics and children's activities.

For the purposes of this study the Hamilton-Wentworth region was divided into 181 neighborhoods. In the central region these neighborhoods correspond to census tracts in size, whereas in the suburbs the neighborhoods are generally only slightly larger and hence represent a significant disaggregation of census tracts. All information provided by the households was coded by using the neighborhood numbers; therefore a fine breakdown of activity at a spatial scale was possible. Off-peak neighborhood-to-neighborhood (including intraneighborhood) automobile travel times were collected for this zonal system during the survey period and represent an average of up to four timings over the actual road network.

Data on every trip (hereafter generally referred to as a tour if it involves more than one nonhome stop) made by the household during the 2-week survey period were collected, where a trip was defined as a journey in which any person 16 years or older and not in elementary or secondary school leaves the home, visits one or more places, and returns home. A stop on such a tour occurred each time someone stopped traveling in order to shop, work, engage in recreation, socialize, and so on. A change of mode (e.g., walk to bus stop, wait for bus) did not count as a stop. Trips of less than two blocks were not recorded as generating separate stops.

A total of 46 stop purposes (in addition to work) were used to characterize each stop on each tour made by the households during the 2-week survey period. In addition, multiple purposes could be recorded for any given stop on any tour. Thus, considerable detail concerning tour characteristics is available within the study data. For the purposes of this study, however, only one primary purpose was associated with each stop and the 46 nonwork purposes were aggregated into four categories: grocery shopping, nongrocery shopping, social-recreational-other (i.e., all other nonwork, nonshopping purposes, hereafter referred to as SRO), and return home. Further, the analysis dealt only with nonwork tours, that is, with tours that did not include a work stop.

Given these data, it is impossible to establish any direct evidence of travel strategy or travel routines. The diaries simply record actions and not the motivations and objectives surrounding these actions. Information on motivation should perhaps be collected in the future, but it will be difficult to extrapolate from the unique situation of every trip to a general model of travel behavior. To give just a simple example, suppose 90 percent of a given household's travel follows a regular routine. The other 10 percent occurs in response to particular household circumstances (e.g., the need to deliver someone to the airport). This unique trip might disturb an entire sequence of trips and, depending on the observation point, could be highly misleading about the bulk of the household's travel. The simplifying assumption to be made in what follows is that some inferences about the nature of urban travel can be drawn from simple average statistics computed from the diary responses.

In the next section the interplay between the purpose of a stop, the number of the stop within the tour, and the average travel times spent making various transitions will be examined. The path choice issues discussed earlier are dealt with next: the extent to which stop sequences correspond to minimum path tours and the extent to which stop sequences are ordered by distance from home. Finally, the major findings of the paper are summarized and their implications for future modeling efforts are discussed.

TRAVEL TIME BY TRIP PURPOSE OR STOP COMBINATION

Table 1 shows the average one-way stop-to-stop (or link) travel time by stop number and purpose of the stop for all nonwork tours in the sample involving up to five stops. For example, 617 observations were made of nongrocery shopping on the third stop of tours and typically these involved 5.20 min of travel from the previous activity. Information concerning first stops is further broken down by whether the stop is the only nonhome stop on the trip or it is the first stop on a multiple-stop tour. Points to note from this table include the following:

1. Single-stop trips exhibit shorter average stop-to-stop travel lengths than do multistop tours, regardless of stop purpose. Single-stop grocery trips, however, exhibit by far the most dramatic tendency in this regard, with an average single-stop one-way link travel time 1.14 min less than the average first-stop link travel time for multistop grocery tours. This result implies that single-stop

TABLE 1 Average One-Way Travel Time by Stop Number and Purpose

	Grocery Shopping	Non-Grocery Shopping	Social-Recreation-Other	Return Home	All Purposes
Stop No. 1					
Single-stop	3.10 (1386)	4.94 (2016)	5.90 (5292)	-	5.23 (8694)
Multistop	4.24 (495)	5.68 (1536)	6.45 (2363)	-	5.94 (4394)
All trips	3.40 (1881)	5.26 (3552)	6.07 (7655)	-	5.47 (13088)
Stop No. 2	4.33 (865)	4.99 (1668)	5.93 (1861)	5.23 (8694)	5.24 (13088)
Stop No. 3	4.24 (314)	5.20 (617)	6.03 (889)	5.41 (2565)	5.42 (4385)
Stop No. 4	4.84 (135)	5.66 (221)	6.17 (343)	5.62 (1117)	5.67 (1816)
Stop No. 5	4.54 (31)	5.45 (78)	6.52 (130)	5.71 (458)	5.78 (697)
All Stops Multistop	4.33 (1840)	5.33 (4120)	6.19 (5586)	5.51 (4140)	5.57 (15686)
All Trips	3.80 (3226)	5.20 (6136)	6.05 (10878)	5.32 (12834)	5.39 (33074)

Note: Travel time in minutes; sample size given in parentheses.

grocery destinations are more localized than are grocery stops that occur at the beginning of a multistop trip chain.

2. With the exception of the single-stop trips, no clear relationship emerges between average link travel time and stop number within any given purpose. Thus, it is not clear that stop number plays a significant role in determining trip lengths (or the spatial range of alternatives considered) for multistop tours.

3. A consistent pattern in average link travel times exists across purposes; grocery stops always incur the shortest travel time and SRO stops incur the longest. Averaging across all stops for multistop tours, this results in an average of 1.00 min (23 percent) longer for nongrocery stops than grocery stops and an average of 1.86 min (43 percent) longer for SRO stops than grocery stops.

In order to investigate the interaction between stop purpose and average trip length further, average link travel times were computed for each purpose-to-purpose transition (e.g., grocery shopping to nongrocery shopping) for each stop transition (from the first to the second stop, from the second to the third, etc.). As in Table 1, no discernible trends emerged with respect to stop number, so only the average link travel times for each purpose-to-purpose transition across all stops are presented here (see Table 2). The general trend of shorter grocery link times than nongrocery times, which in turn are less than SRO times, is reinforced in Table 2, from which the following observations may be made:

1. Regardless of the purpose of the destination stop, the link travel time is the least if the origin stop's purpose is grocery shopping and the greatest if the origin purpose is SRO;

2. With the exception of the grocery-to-nongrocery transition, the link travel time is the least if the destination stop's purpose is grocery shopping and the greatest if it is SRO, regardless of the purpose of the origin stop; and

3. Times along the main diagonal of Table 2 are monotonically increasing, whereas off-diagonal times are ordered so that a lower-triangle time is always greater than its upper-triangle counterpart (e.g., the average travel time for a nongrocery-to-grocery transition is greater than that for a grocery-to-nongrocery transition).

TABLE 2 Average Travel Time on One-Way Links by Purpose Transition for All Stops

From \ To	Grocery Shopping	Non-Grocery Shopping	Social-Recreation-Other
Grocery Shopping	4.13 (263)	3.99 (358)	4.12 (274)
Non-Grocery Shopping	4.26 (605)	4.84 (1228)	5.81 (854)
Social-Recreation-Other	4.71 (495)	5.86 (1045)	6.32 (2175)

Note: Travel time in minutes; sample size given in parentheses.

The only anomaly that exists in Table 2 is the grocery-to-nongrocery transition, which possesses a smaller average travel time than the nongrocery-to-grocery transition. This transition, however, represents only 13.6 percent of all nongrocery stops in the sample. Thus, one can speculate that nongrocery stops will be linked to preceding grocery stops only infrequently, and then only if they can be chosen from a relatively localized (with respect to the grocery stop) set of destinations.

Tables 1 and 2 imply that link travel time impedance is highest for grocery stops and lowest for SRO stops or that only a relatively localized choice set of possible destinations is considered by grocery shoppers, whereas a considerably more dispersed set of destinations is involved in SRO trips (with, in either case, nongrocery occupying an intermediate position). This is not an unexpected result, but it does reinforce the need for explicitly considering trip purpose in nonwork travel demand modeling, because the trip distribution pattern (and associated average trip lengths) is likely to vary substantially depending on the purpose or purposes involved. Table 2 indicates that this result is accentuated when one controls for origin stop purpose, although it cannot be determined from this simple analysis whether this result is due to an actual interaction between stop purposes (i.e., a nongrocery-to-SRO transition is fundamentally different from an SRO-to-SRO transition) or whether it is simply due to variations in stop destination choice sets given previous stop choices (which in turn has implications for the stop choice process).

Table 3 shows the implications of these results

TABLE 3 Travel Time on Selected Two-Stop Trips

Trip Pattern	Link Travel Time Home - Stop 1	Stop 1 - Stop 2	Stop 2 - Home	Total Travel Time
GGH	4.24	4.16	3.63	12.03
GNH	4.24	3.72	5.01	12.97
GOH	4.24	4.15	6.82	15.21
NGH	5.68	4.22	3.63	13.53
NNH	5.68	4.67	5.01	15.36
NOH	5.68	5.57	6.82	18.07
OGH	6.45	4.52	3.63	14.60
ONH	6.45	5.71	5.01	17.17
OOH	6.45	6.23	6.82	19.50

Note: Travel time in minutes; G = grocery shopping stop; N = nongrocery shopping stop; O = social-recreation-other stop; H = return home.

by presenting representative average travel times for nine different two-stop multipurpose tours, constructed by adding together the appropriate stop-specific average transition times (i.e., these times are not taken from Table 2, which represents averages taken across all stops). As shown in Table 3, these trip times vary from 12.03 min for a grocery-grocery-home tour to 19.50 min for an SRO-SRO-home tour. This represents perhaps the most important implication of this section's analysis for nonwork trip modeling: there is no such thing as one average travel time for nonwork trips. Rather, this average varies with the number and the purposes of the stops involved, and hence models that ignore the multistop multipurpose nature of nonwork travel are unlikely to be able to replicate such travel adequately.

STOP SEQUENCING IN MULTISTOP TRIPS

Given a symmetrical zone-to-zone travel time matrix whose entries all obey the triangle inequality (conditions that apply to the off-peak automobile travel time matrix used in this analysis), it can be shown that the number of distinct paths through n stops away from home is n!/2. For any tour consisting of a given set of stops, these n!/2 distinct paths can be enumerated in order to compute the minimum time path through the set of stops (denoted by a) and the maximum time path (denoted by c) and these quantities can then be compared with the travel time for the path actually chosen by the trip maker (denoted by b). Table 4 presents the results of such an analysis for all nonwork tours in the sample with five or fewer stops. The table shows the number of tours observed having n stops (n = 1, ..., 5), the number of distinct paths associated with an n-stop tour (n!/2), and the breakdown of observations into the following categories:

a = c: minimum and maximum time paths are equal, hence the chosen path must be a minimum;
a = b < c: a minimum time path is chosen;
a < b = c: a maximum time path is chosen; and
a < b < c: a time path between the minimum and maximum times is chosen.

TABLE 4 Observations of Minimum Path Behavior

n	Paths	Observations	a = c	a=b<c	a<b=c	a<b<c
1	1	8694	8694			
2	1	2566	2566			
3	3	1116	155	523 (54.4)	328 (34.1)	110 (11.4)
4	12	458	20	175 (40.0)	64 (14.6)	199 (45.4)
5	60	140	2	33 (23.9)	8 (5.8)	97 (70.3)

Note: n = number of nonhome stops on the trip; a = minimum time path; b = path chosen by trip maker; c = maximum path. Numbers in parentheses express the percentage of observations less the a = c observations that fall into the given category for a given value of n. Thus, for example, 523/(1116 - 155) = 54.5 percent of the three nonhome stop tours for which a ≠ c fell into the a = b < c category.

As shown in Table 4, the percentage of tours that display distance minimization (excluding the a = c case) is 54.4, 40, and 23.9 percent for three-, four-, and five-stop tours, respectively. The decrease in minimization behavior as the number of stops increases reflects the difficulty facing trip makers in finding the minimum path on longer tours.

The percentage of tours that display distance maximization is 34.1, 14.6, and 5.8 percent for three-, four-, and five-stop tours, respectively. These percentages are lower than the corresponding values for distance minimization. Finally, the percentage of tours with intermediate behavior (a < b < c) is 11.4, 45.4, and 70.3 percent, respectively, for three, four, and five stops. Thus, as the number of stops on a tour increases, the probability of traveling on either a minimum or a maximum path decreases and the probability of using some intermediate travel path increases substantially.

Assuming that the relative attractiveness of locations for various purposes does not vary with the order in which they are visited and that temporal constraints do not exist that would force specific sequencing of stops on some tours, any utility-maximizing model of nonwork travel that assumes a simultaneous or joint choice process over the destinations and purposes associated with a given multistop multipurpose tour (with either a fixed or variable number of stops) implicitly assumes that a minimum path will be taken through the selected set of stops (because the utility associated with this set of stops can always be improved by moving from a nonminimum to a minimum path). Because 45.5, 60.1, and 76.1 percent of the three-, four-, and five-stop tours, respectively, in the sample did not choose a minimum path (i.e., they chose a maximum or intermediate path), there is some evidence either that tour chain decisions are made in some sequential stop-by-stop fashion (rather than simultaneously over all stops) or that the assumptions made earlier do not hold. This latter supposition, however, in turn argues for a more dynamic view of nonwork trip decision making, which, again, is probably most compatible with some form of sequential decision structure.

Table 5 presents the average minimum, maximum, and chosen tour times for three-, four-, and five-stop tours for each of the four cases presented in Table 4 (a = c, etc.). It can be seen from Table 5 that

1. The a = c case (all feasible paths have the same travel time) generally occurs for relatively short tours,
2. Travelers tend to choose minimum paths when large differences exist between the minimum and maximum path times and conversely maximum time paths tend to be chosen when the differences between the minimum and maximum path times are relatively small, and

TABLE 5 Average Tour Times Versus Number of Stops and Type of Path

n	Type of Path	Chosen Path Case a = c	a=b<c	a<b=c	a<b<c
3	Minimum Time Path	671	1331	1374	1564
	Chosen Path	671	1331	1599	1564
	Maximum Time Path	671	1709	1599	1866
			(378)	(225)	(362)
4	Minimum Time Path	788	1557	1480	1664
	Chosen Path	788	1557	1892	1866
	Maximum Time Path	788	2218	1892	2278
			(661)	(412)	(614)
5	Minimum Time Path	1849	1574	1933	1860
	Chosen Path	1849	1574	2420	2147
	Maximum Time Path	1849	2618	2420	2957
			(1044)	(487)	(1097)

Note: Travel time in seconds. The numbers in parentheses indicate the differences between the maximum and minimum path times for each case.

3. Intermediate time paths tend to be chosen when the differences between minimum and maximum time paths lie between the two cases described previously (although these differences and the chosen travel times tend to lie closer to the case of minimum path choice than that of the maximum path choice; i.e., people come closer to minimizing than maximizing).

These results imply that people are better able to choose between shorter and longer paths as the differences between these paths become more pronounced. This may simply be because the traveler does not possess (or is willing to gather) sufficiently accurate information to choose among alternative paths, except when it is obvious that major differences exist. It may also imply threshold effects, in which the traveler is relatively indifferent to variations in travel times, as long as they do not exceed certain threshold limits. Finally, the choice of intermediate paths that tend to fall closer to minimum time paths than maximum time paths and the choice of maximum time paths only when they exceed minimum feasible times by relatively small amounts tend to indicate that path choice, if not a global optimization process, is perhaps at least a relatively rational, structured one, in which good choices are probably made far more often than bad ones.

In order to further elucidate overall patterns in the stop sequencing of multistop tours, the spatial ordering of these stops can be investigated. That is, if a trip maker visits n locations away from home, these locations can be ranked in order of their proximity to home, with distance rank 1 being allocated to the location nearest home. If the temporal ordering of the activities on the tour is given by the stop number i = 1, ..., n, the spatial ordering is given by the distance rank of the location at stop i, that is, R(i). The question to be investigated is then whether consistent patterns exist with respect to R(i), i = 1, ..., n.

Table 6 summarizes the distributions of stop-sequence and distance-rank combinations for two-stop through five-stop nonwork tours, respectively.

TABLE 6 Stop Ordering Versus Distance Ranking

	No. of Stops by Rank					
Stop No.	Rank 1	Rank 2	Tied for Rank 1	Rank 3	Rank 4	Rank 5
Two-Stop Trips						
1	917	1,203	446			
2	1,203	917	446			
Three-Stop Trips[a]						
1	571	299		246		
2	361	322		433		
3	629	297		190		
Four-Stop Trips[a]						
1	221	89		82	66	
2	99	118		107	134	
3	112	106		126	114	
4	232	104		68	54	
Five-Stop Trips[a]						
1	71	30		22	6	11
2	27	24		27	27	35
3	26	31		34	31	28
4	23	29		34	29	25
5	55	42		15	16	12

[a]Tieds ranks included.

Thus, for example, Table 6 indicates that for the two-stop tours in the sample, 917 had first stops that had distance rank 1 (i.e., were the closest to home of the stops visited on the tour), 1,203 had first stops that had distance rank 2 (i.e., were the second closest to home of the stops visited on the tour), and 446 had first stops that were tied for distance rank 1 (i.e., stops 1 and 2 were equidistant from home). Two major observations that emerge from consideration of this table are that the nearest stop is likely to be either the first or last stop on the tour and that increasingly distant stops are increasingly likely to be intermediate stops in the trip chain. This latter result is shown in Figure 1, in which it is seen that the percentage of intermediate stops increases both with distance rank (for a given number of stops) and with the number of stops (for a given distance rank).

These results are probably compatible with either a simultaneous or a sequential decision-making structure. They do, however, indicate that stop ordering tends to be at least partially a function of routing considerations (e.g., visit the nearest desired destination next) rather than an indication of activity priorities (e.g., visit the highest-priority destination next). This in turn implies that a priori categorization of tours according to their stop ordering (e.g., defining the tour purpose in terms of the purpose of the first stop) is probably not appropriate and probably does not represent a fruitful approach to conceptualizing multistop multipurpose tours.

SUMMARY OF RESULTS AND IMPLICATIONS FOR FURTHER WORK

The major results that emerge from the analyses presented in the previous two sections include the following:

1. Grocery shopping trips have dramatically different average link travel times for single-stop and multistop trips; the latter are typically nearly 40 percent larger than the former (4.33 min versus 3.10 min). This implies a definite tendency for two modes of grocery shopping to exist: localized single-stop shopping and more dispersed multistop shopping. Although the single-stop link times for nongrocery and SRO trips are the lowest link times for each of these other purposes as well, they are only 7 and 5 percent lower than the average multistop link time for each purpose, respectively.

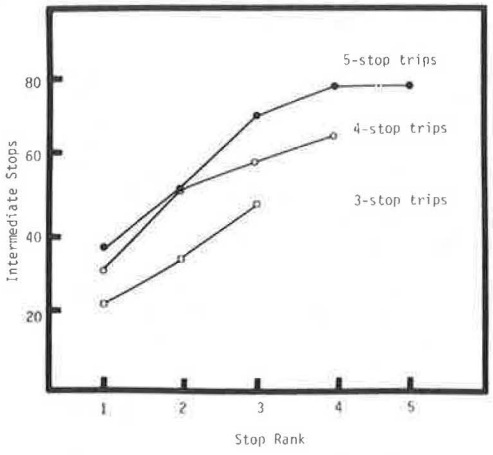

FIGURE 1 Percentage of intermediate stops by distance rank.

2. A consistent hierarchy exists in average link travel times with respect to purpose. Grocery stops consistently incur the shortest average link times, whereas SRO stops always incur the longest. This result holds regardless of whether one controls for the origin purpose of the link, the destination purpose, or both the origin and destination purposes.
3. No consistent pattern exists with respect to average link travel time and stop number. In particular, link travel times do not appear to grow either typically larger or smaller as the stop number increases.
4. A considerable number of multistop tours (between 46 and 76 percent, depending on the number of stops in the tour) do not exhibit minimum path stop sequences. This implies that temporal constraints exist that prevent the selection of minimum path routes, that within-tour dynamics exist that lead to the selection of nonminimum path routes, that information constraints typically exist that limit the traveler's ability to identify a minimum time path, or that tours are constructed through a sequential rather than simultaneous decision process.
5. Minimum time paths tend to be chosen when the differences between the minimum and maximum time paths are large, whereas maximum time paths tend to be chosen when this difference is small. This implies that people tend to make good path choices even if they are not ones that minimize travel time.
6. Stops ranked as being nearest to a trip maker's home tend to be either the first or last stop on multistop tours. If the locations visited on a tour are ranked according to distance, with the shortest distance ranked 1, places with high distance rank are unlikely to be either the first or the last stop on a trip. This result is consistent with the hypothesis that stop sequencing is at least partially determined by a path choice process (whether or not this process is a path-minimizing one), and it certainly implies that care should be taken in associating any priorities among a tour's purposes with the stop sequence.

Two major implications emerge from these results for the modeling of nonwork travel. First, it is clear that nonwork travel must be explicitly modeled as being multistop and multipurpose in nature. Ignoring multiple-stop tours will result in a serious underestimation of total travel as well as provide a poor conceptual starting point for behavioral modeling efforts. Significant variations in behavior exist among purposes (e.g., virtually a 100 percent difference in average link travel times between single-stop grocery links and multistop SRO links); thus it is unlikely that a sound behavioral model of nonwork travel that is capable of generating acceptable predictions can be constructed without its being explicitly multipurpose in nature.

Second, no strong evidence has been found to suggest that trip makers globally optimize their travel by means of a simultaneous choice process over all available destinations, purposes, and routes. Indeed, the results obtained are probably more consistent with a more sequential decision process. This is, on balance, a promising result for at least two reasons. First, sequential models provide a convenient mechanism for keeping the combinatorics or dimensionality of the choice process within practical limits (e.g., one might model two choices among n and m alternatives, respectively, rather than one choice among nm alternatives). Second, a more sequential decision process may well be more in keeping with actual human decision making than is a simultaneous or global-optimizing one.

To go beyond these preliminary speculations requires more work of both a theoretical and an empirical nature. With respect to the latter, logical extensions to the work presented here include the splitting of the analysis by mode of travel (a computerized transit network for the Hamilton-Wentworth region is currently under development that will enable this to be done), the analysis by means of computer-graphic displays of the shape of multistop travel patterns as well as the changes in the distribution of available alternatives for different purposes as the trip-maker moves from stop to stop, and the analysis of patterns in the generation of different tour purpose combinations (currently under way). In the longer run, the need exists to go beyond the revealed-preference travel diary data of the Hamilton-Wentworth survey and to question people directly concerning their choice process so as to achieve an improved understanding of this process.

ACKNOWLEDGMENT

The work reported in this paper was supported by an operating research grant from the Natural Sciences and Engineering Research Council of Canada. Computer time was provided by Ohio State University. The data used in the study were supplied by Michael Webber of the Department of Geography, McMaster University.

REFERENCES

1. S. Hanson. Urban Travel Linkages: A Review. In Behavioral Travel Modelling (D. Hensher and P. Stopher, eds.), Croom-Helm, London, 1979.
2. T. Adler and M. Ben-Akiva. A Theoretical and Empirical Model of Trip-Chaining Behavior. Transportation Research, Vol. 13B, 1979, pp. 243-257.
3. J. Horowitz. A Utility Maximizing Model of the Demand for Multidestination Non-Work Travel. Transportation Research, Vol. 14B, 1980, pp. 369-386.
4. G. Mulligan. Central Place Populations: Some Implications of Consumer Shopping Behavior. Annals of the Association of American Geographers, in preparation.
5. R. Kitamura. Sequential History-Dependent Approach to Trip-Chaining Behavior. In Transportation Research Record 944, TRB, National Research Council, Washington, D.C., 1983, pp. 13-22.
6. P. Jones. Identifying the Determinants of Destination Choice. In Identification and Evaluation of the Determinants of Travel Choice (D. Hensher, ed.), Saxon House, New York, 1977.
7. D. Hensher. The Structure of Journeys and the Nature of Travel Patterns. Environment and Planning A, Vol. 8, 1976, pp. 655-672.
8. G. Hemmens. Analysis and Simulation of Urban Activity Patterns. Socio-Economic Planning Sciences, Vol. 4, 1970, pp. 53-66.
9. F.E. Horton and W.E. Wagner. A Markovian Analysis of Urban Travel Behavior: Pattern Response by Socio-Economic Occupational Group. In Highway Research Record 283, HRB, National Research Council, Washington, D.C., 1969, pp. 19-29.
10. G. Bentley, A. Bruce, and D. Jones. Intra-Urban Journeys and Activity Linkages. Socio-Economic Planning Sciences, Vol. 11, 1977, pp. 213-220.
11. M. O'Kelly. A Model of the Demand for Retail Facilities Incorporating Multistop Multipurpose

Trips. Geographical Analysis, Vol. 13, 1981, pp. 134-148.

12. D. Damm and S. Lerman. A Theory of Activity Scheduling Behavior. Environment and Planning A, Vol. 13, 1981, pp. 703-718.
13. D. Damm. Towards a Model of Activity Scheduling Behavior. Ph.D. thesis. Department of Civil Engineering and Urban Studies, Massachusetts Institute of Technology, Cambridge, Mass., 1979.
14. S. Hanson. Spatial Diversification and Multipurpose Travel: Implications for Choice Theory. Geographical Analysis, Vol. 12, 1980, pp. 245-257.

Socioeconomic and Travel Forecasts for Alternatives Analysis in the Puget Sound Region

CATHY J. STROMBOM and G. SCOTT RUTHERFORD

ABSTRACT

The development and use of socioeconomic and travel forecasts for evaluation of major transit investments in the Puget Sound region of Washington State are described. Although procedures used to produce socioeconomic and travel forecasts may be considered standard relative to techniques used elsewhere, the analysis and interpretation of the results have had a substantial impact on the decisions of policy makers in the region. How the results are being used for decision making is the thrust of this paper. Highlighted is how forecasts have been used at each phase of the transportation planning process: for systems planning, for corridor analysis, and for project planning. First, forecasts played a key role in defining the nature of the future regional transportation system as contained in the Regional Transportation Plan. Predictions of levels of highway congestion and potential transit ridership were subsequently used to rank corridors as to priority for further analysis. In evaluating alternative transit projects within corridors, policy makers have given priority to those projected to generate additional transit patronage. Because billion-dollar decisions are being made today for tomorrow's transit capital and operating programs, the need for constant update of the regional data base and forecasting capabilities has been reinforced. Additional survey work and model refinements are planned to help ensure that adequate technical information is available as projects go into preliminary engineering.

In 1982 the Puget Sound Council of Governments (PSCOG) adopted a new Regional Transportation Plan (1). The plan constituted a major departure from earlier plans in that it contained an explicitly stated policy that there would be no new freeway corridors or major highway expansion in the region during the next 20 years. Yet the adopted population and employment forecasts used in preparing the plan implied that an almost 45 percent increase in daily person trips in the region would occur between 1980 and 2000. To help accommodate this growth in travel demand, the elected officials set as objectives of the plan to increase the market share of transit and of ridesharing over the next 20 years. These objectives were to be met through the development and implementation of aggressive transit and ridesharing programs.

Figure 1 shows the location of the Puget Sound region, which includes the cities of Everett, Seattle, and Tacoma. (The arrow indicates the corridor currently under study.) Population and transportation characteristics for the region are summarized in Table 1. As indicated in Table 1, use of transit for the work trip is forecast to increase from 9.6 percent to 11.7 percent during the 20-year period on a regional basis. Daily average vehicle occupancy is expected to increase from 1.38 in 1980 to 1.46 in 2000. Although this still is less than the average vehicle occupancy in 1960, a reversal of the downward trend that occurred between 1960 and 1980 is an objective of the plan. Figures 2 and 3 are graphs of transit use and average vehicle occupancy during the period 1960-2000.

Although a transit mode split for work trips of 11.7 percent in 2000 is forecast for the region as a whole, the proportion using transit for work trips destined for downtown Seattle is expected to increase from 40 percent in 1980 to 54 percent in 2000. In Table 2 downtown Seattle population, employment, and travel data are compared for 1980 and 2000. Given the large increases in employment projected for downtown Seattle and the high levels of transit use for trips to the central business district (CBD), the need for a higher-capacity transit system, such as light rail, seemed likely. PSCOG decided that the feasibility of a light rail transit

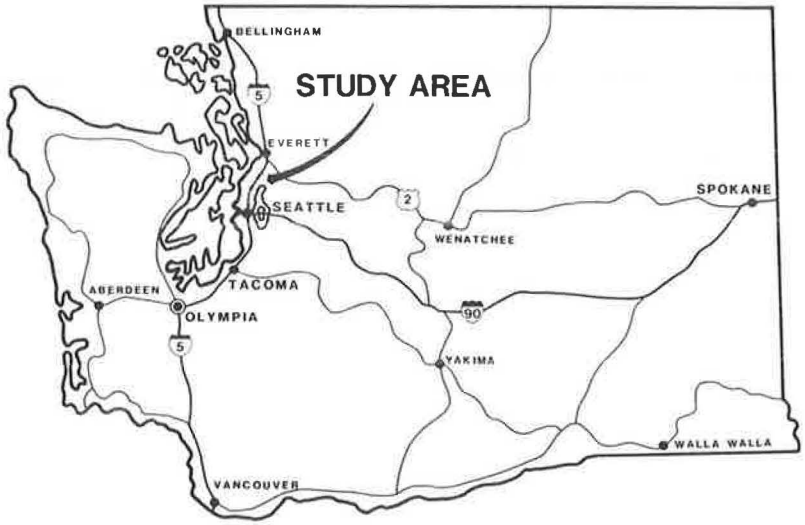

FIGURE 1 Location of Puget Sound region.

TABLE 1 Central Puget Sound Region: Population and Transportation Characteristics

DATA ITEM	YEAR 1960	1980	2000
Population	1,366,200	2,102,500	2,920,900
Transit Trips (Daily)	161,300	246,900	412,200
Percent total trips	5.2%	3.9%	4.6%
Percent work trips	9.5%	9.6%	11.7%
Vehicle Trips (Daily)	2,155,000	4,468,000	6,324,600
Average vehicle occupancy	1.54	1.38	1.46
Average trip length (miles)	5.90	7.73	7.98
Vehicle miles of travel	12,715,000	34,538,000	50,500,000
Per Capita Measures			
Transit trips	.12	.12	.14
Vehicle trips	1.58	2.13	2.17
Vehicle miles of travel	9.30	16.43	17.29

(LRT) system for Seattle should be assessed as part of the development of the Regional Transportation Plan. An outside consultant was asked to evaluate the potential for LRT in Seattle by 2000. The transit forecasts used for this assessment were those developed for the Regional Transportation Plan based on an all-bus system. In this preliminary study it was concluded that the transit demand in 2000 in at least two of the major travel corridors (the North Corridor and the East Corridor) would be well above the apparent decision threshold of 4,000 to 7,000 passengers per peak hour used by policy bodies in other cities choosing light rail and that this mode would be cost effective in Seattle (2). In fact, in the North Corridor, transit demand has already exceeded this threshold level; the evening peak-hour peak-direction transit demand was 7,000 in 1980.

On the basis of this study and as a result of the emphasis that the elected officials felt must be placed on transit development during the next 20 years, PSCOG and the Municipality of Metropolitan Seattle (METRO) decided to jointly evaluate alternative transit investments in major corridors.

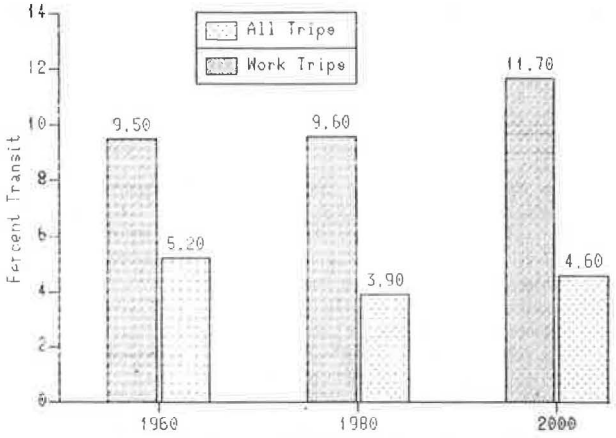

FIGURE 2 Percentage of daily trips by transit, Puget Sound region.

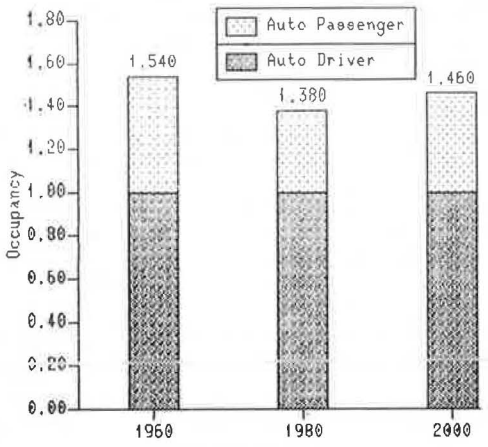

FIGURE 3 Automobile vehicle occupancy, Puget Sound region.

TABLE 2 Downtown Seattle: Population, Employment, and Transportation Characteristics

DATA ITEM	YEAR 1980	2000
Population	10,730	20,400
Employment	114,900	179,100
WORK TRIPS (DAILY)		
No. of Trips	175,400	265,200
Percent Transit	39.9%	53.6%
No. of Transit Trips	69,900	142,200
NON-WORK TRIPS (DAILY)		
No. of Trips	293,000	397,900
Percent Transit	14.5%	16.2%
No. of Transit Trips	42,400	64,600

RANKING CORRIDORS BY PRIORITY FOR ANALYSIS

In keeping with the guidelines laid out by UMTA for alternatives analysis (3), only one corridor could be evaluated at a time. Projected population and employment growth and travel forecasts were used to rank the major corridors by priority for analysis (4). These data are summarized in Table 3 for the North, East, South, and Eastside Corridors. The transit ridership estimates shown are based on an all-bus system.

The two most heavily traveled corridors today are the North and the East Corridors. Although a larger percentage of growth in population and employment was forecast for the East Corridor than for the North Corridor, the projected absolute number of transit riders was predicted to be considerably higher for the North Corridor.

TABLE 3 Central Puget Sound Region: Population, Employment, and Peak-Hour Transit Ridership for Major Travel Corridors

DATA ITEM	CORRIDOR NORTH[a]	EAST	SOUTH[a]	EASTSIDE
CORRIDOR POPULATION				
1980	441.0	227.0	187.0	236.0
2000	544.0	331.0	275.0	405.0
Percent Growth	23.4%	45.8%	47.0%	71.6%
CORRIDOR EMPLOYMENT				
1980	241.0	181.0	316.0	138.0
2000	357.0	308.0	446.0	239.0
Percent Growth	48.1%	70.2%	41.1%	73.2%
PEAK-HOUR RIDERSHIP				
1980	7.1	3.5	2.2	0.4
2000	14.0	8.7	4.6	0.7
Percent Growth	97%	149%	109%	75%

Note: Data are in thousands.
[a]Includes population and employment in the Seattle CBD.

Analysis of future vehicle trip demand for each of the corridors also revealed that the most severe congestion was likely to occur in the North Corridor. Figures 4 and 5 indicate the levels of congestion in the North Corridor in 1980 and 2000, respectively. In 1980, heavy congestion occurred along several segments of I-5 and adjacent arterials, whereas in 2000 severe congestion is forecast on both I-5 and Aurora Avenue (State Route 99).

Figures 6, 7, and 8 show what is likely to occur on I-5 and Aurora at 145th Street. In Figure 6 the existing distribution of traffic volumes in the afternoon and evening hours at 145th Street is represented. The reference line represents the capacity across this screen line. The peak period today is from approximately 4:00 p.m. to 6:00 p.m. In Figure 7 traffic volumes in 2000 are indicated. Given the same distribution of traffic over time as today, demand would well exceed capacity during today's 2-hr peak period. Figure 8 shows what might happen if all the traffic demand that has been forecast is to be accommodated: More than a 4-hr period of severe congestion would occur. Of course, other things might happen--trips might redistribute themselves or not occur at all. However, forecasts for the North Corridor indicate that an extension of today's peak-hour congestion is extremely likely.

On the basis of this analysis, the steering committee for the study (made up of elected officials from King and Snohomish Counties and a representative from the Washington State Department of Transportation) selected the North Corridor as the corridor with the highest priority for consideration of major transit improvements. The corridor stretches from downtown Seattle to south Snohomish County. Figure 9 shows the rankings for all corridors.

DEVELOPMENT AND SCREENING OF ALTERNATIVES FOR THE NORTH CORRIDOR

Following the selection of the North Corridor as the priority corridor for analysis, alternative alignments and transit technologies were evaluated. Several different technologies and up to eight different alignments were initially considered for the North Corridor. The technologies considered included conventional bus, advanced-technology bus (dual-propulsion vehicles that may or may not run on a guideway), LRT, and exclusive guideway. Figure 10 shows the initial set of alignments. On the basis of existing transit volumes and practical engineering considerations, these were quickly reduced to three alignments: I-5, Aurora Avenue (State Route 99), and a crossover alignment that used parts of each of these two major facilities.

The steering committee also chose at this time to eliminate consideration of exclusive guideway as a separate technology. Their decision was based on the following reasoning. First, preliminary transit forecasts did not appear to justify consideration of heavy rail as an alternative for the Seattle metropolitan region. Second, light rail operating on I-5 would essentially operate on an exclusive right-of-way for much of its length, and engineering considerations for light rail on this alignment would not be unlike those for heavy rail if the latter should come under consideration in the future.

Light rail and conventional bus were thus the technologies considered for each of these alignments, whereas the advanced-technology bus was considered only for I-5. The six basic alternatives endorsed by the steering committee for study included the following:

1. No build,
2. Transportation systems management (TSM),

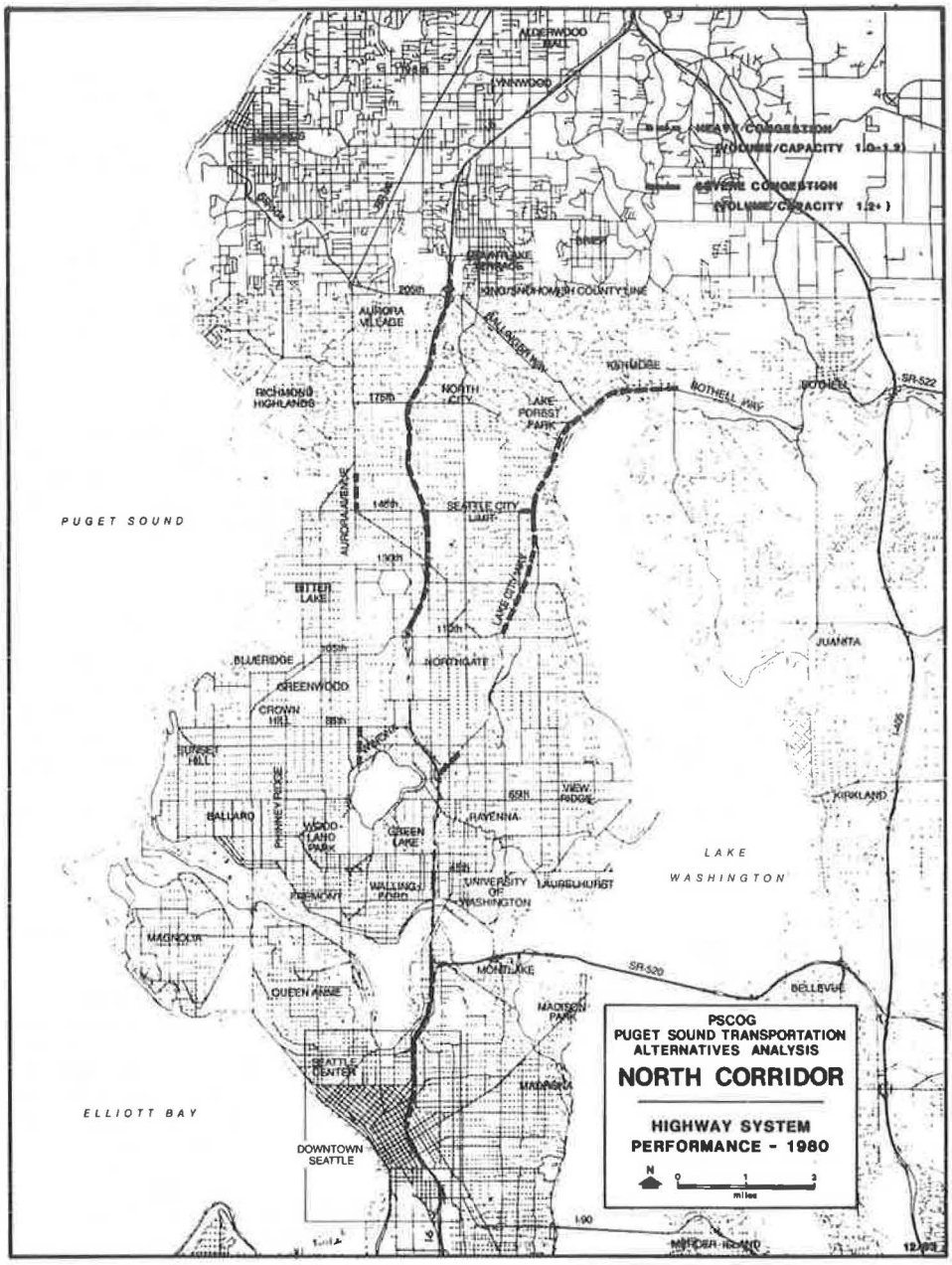

FIGURE 4 Highway system performance, 1980.

3. I-5 guided busway,
4. I-5 LRT,
5. Aurora/I-5 LRT, or
6. Aurora LRT.

Although there were six basic alternatives for the North Corridor, preliminary travel forecasts were prepared for more than 12 different variations. The 12 variations reflected the different combinations of alignments with technologies, surface, and subsurface options in downtown Seattle and different lengths for the light rail and guided-busway options. Figure 11 shows the 12 variations for which preliminary forecasts were prepared.

For the purpose of travel forecasting--and in keeping with UMTA guidelines for alternatives analysis--there were a number of technical and policy assumptions that were held the same for all alternatives. These included the following:

1. Population and employment forecasts (as adopted by PSCOG in 1982);
2. Trip generation by purpose;
3. Highway system (as adopted in the Regional Transportation Plan with exceptions as noted in the following);
4. Trip distribution;
5. Mode-split model coefficients;
6. Transfer penalties;
7. Highway operating costs, future parking costs, and transit fares; and
8. Level of service and geographic coverage of the North Corridor feeder-bus networks.

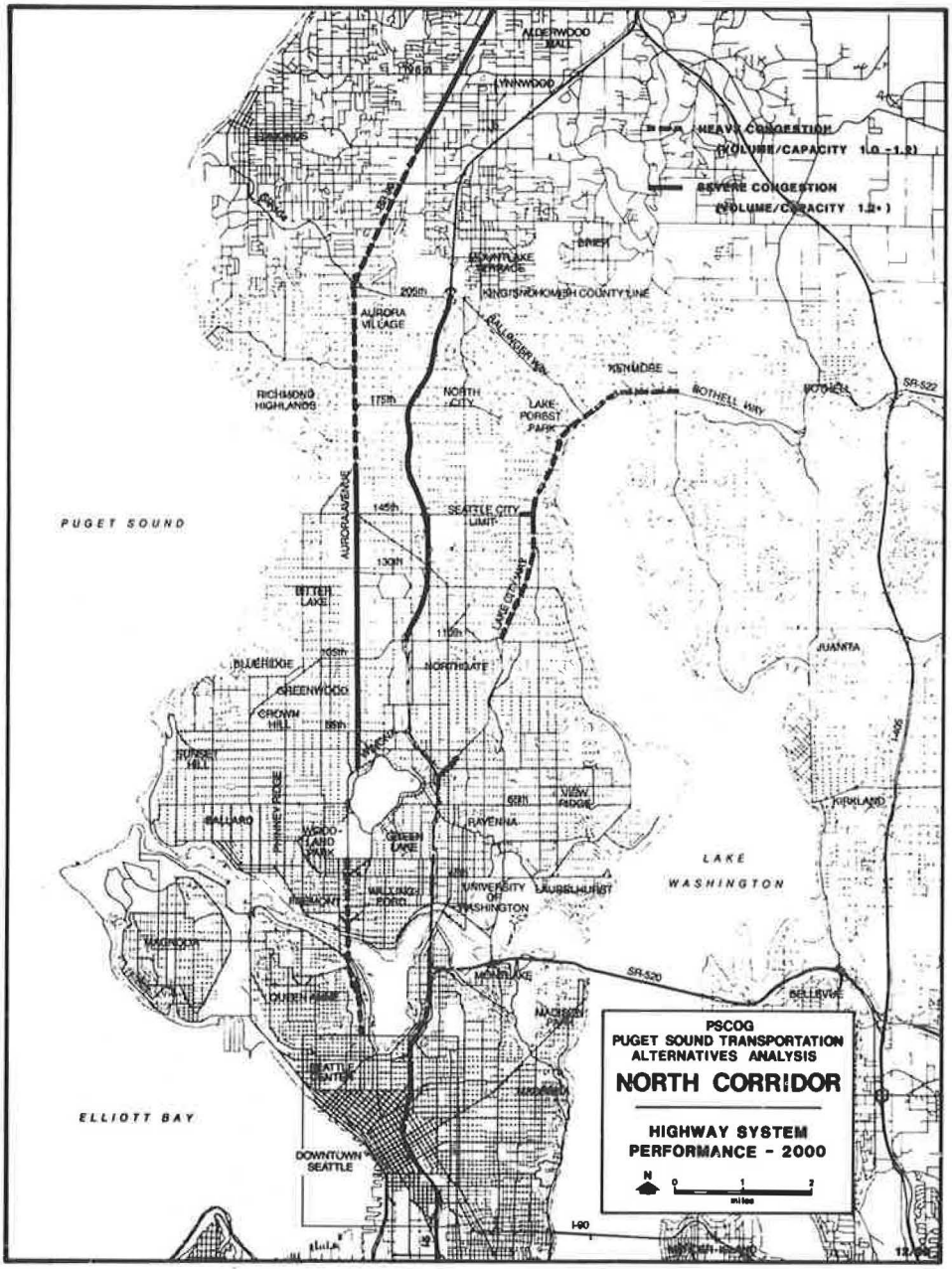

FIGURE 5 Highway system performance, 2000.

The assumptions that varied by alternative primarily affected the coding of the transit networks and included the following:

1. Alignment for the line-haul system,
2. Station locations,
3. Top operating speeds and average speeds in the CBD,
4. CBD stop times,
5. Park-and-ride lot locations,
6. Feeder-bus transfer points, and
7. Coded highway networks to reflect the taking of lanes for transit or the implementation of high-occupancy-vehicle facilities for the all-bus or TSM alternative.

Preliminary forecasts for the 12 variations were prepared over a 3-month period in spring 1983. The peak-hour peak-direction volumes on the line-haul system for the Aurora and Aurora/I-5 crossover alternatives ranged from 2,900 to 5,900, respectively, whereas the peak-hour peak-direction volumes on the I-5 line-haul facility (for either light rail or advanced-technology bus) were estimated at 9,000 to 10,200 passengers. These preliminary forecasts were presented to the steering committee in August 1983. The committee recommended that on the basis of these forecasts, the remaining work should focus on the I-5 alternatives only.

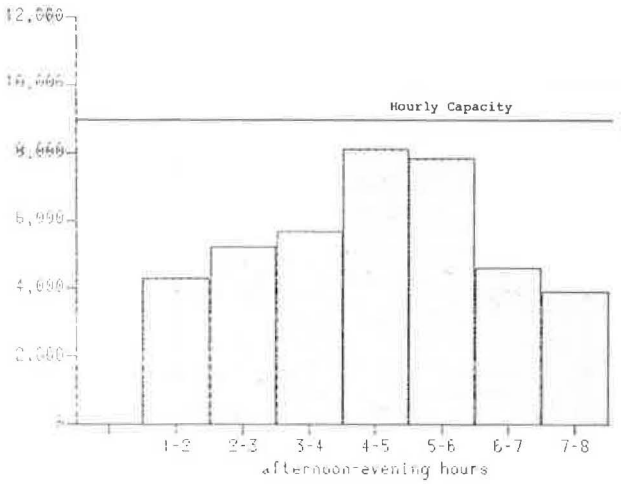

FIGURE 6 Existing traffic volumes, northbound I-5 and Aurora at 145th Street.

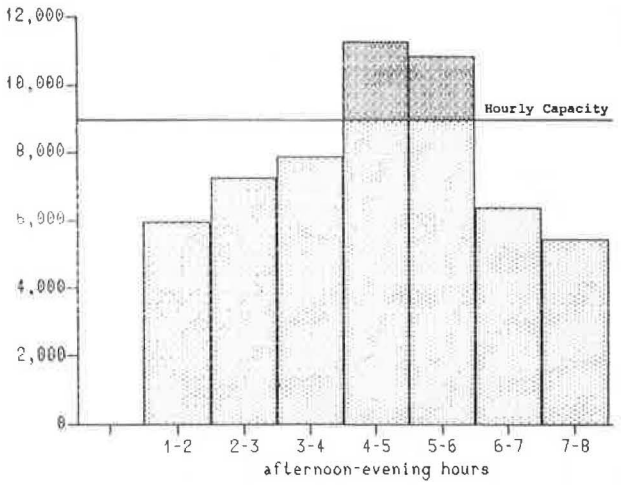

FIGURE 7 Traffic demand in 2000 with unconstrained time-of-day distribution, northbound I-5 and Aurora at 145th Street.

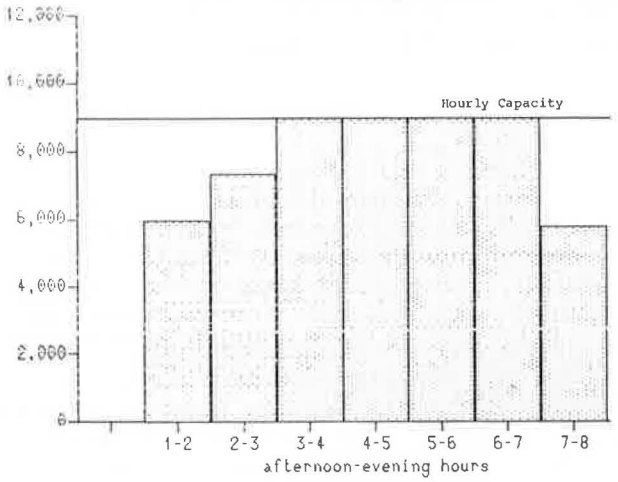

FIGURE 8 Traffic volumes in 2000 with peak spreading, northbound I-5 and Aurora at 145th Street.

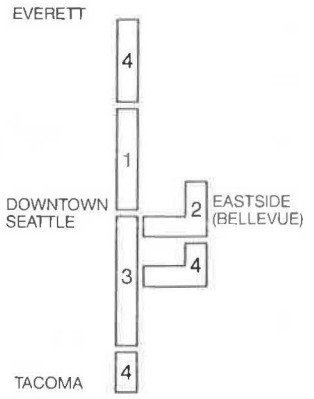

FIGURE 9 Corridor ranking.

REFINEMENT OF PATRONAGE FORECASTS FOR I-5 ALTERNATIVES

Additional refinement and analysis were carried out for five major alternatives on the I-5 alignment. These alternatives included

1. No build,
2. TSM,
3. Advanced-technology bus,
4. LRT in downtown tunnel, or
5. LRT on downtown mall.

Figure 12 shows the I-5 alignment for the LRT alternative. The advanced-technology bus would run on the surface out in the corridor and operate on a guideway in a tunnel in downtown Seattle. The buses would have dual-propulsion capabilities, using diesel power outside the CBD and running under electric power in the tunnel.

For the I-5 LRT alternatives, two lengths were evaluated. The maximum length would extend to Alderwood Mall in south Snohomish County, whereas the minimum length would end at 145th Street at the Seattle city limits. Refinements to the coded networks were made to reflect the final definition of alternatives. Coded speeds for the express line-haul transit service were adjusted to reflect an analysis of peak and off-peak speeds based on demand to capacity relationships on I-5 and at interchanges.

Table 4 summarizes existing and projected transit ridership for each of the alternatives. Both daily and annual figures are provided for the North Corridor and for the region as a whole. North Corridor transit trips are defined as those having at least one end of the trip within the North Corridor. As indicated in Table 4, the maximum-length LRT alternative operating in a tunnel in downtown Seattle would have the largest ridership, which would be 33,000 more daily trips than the no-build condition in 2000. The alternative with the next highest patronage estimate is the maximum-length surface LRT; there would be a difference of 29,000 trips daily relative to the no-build condition. The advanced-technology bus and the minimum-length LRT in a downtown tunnel produce only slightly different levels of transit patronage--27,000 and 26,000 more trips than those under the no-build condition, respectively.

The percentage of daily person trips in the North Corridor using transit in 2000 for each alternative is provided in Table 5. For trips to downtown

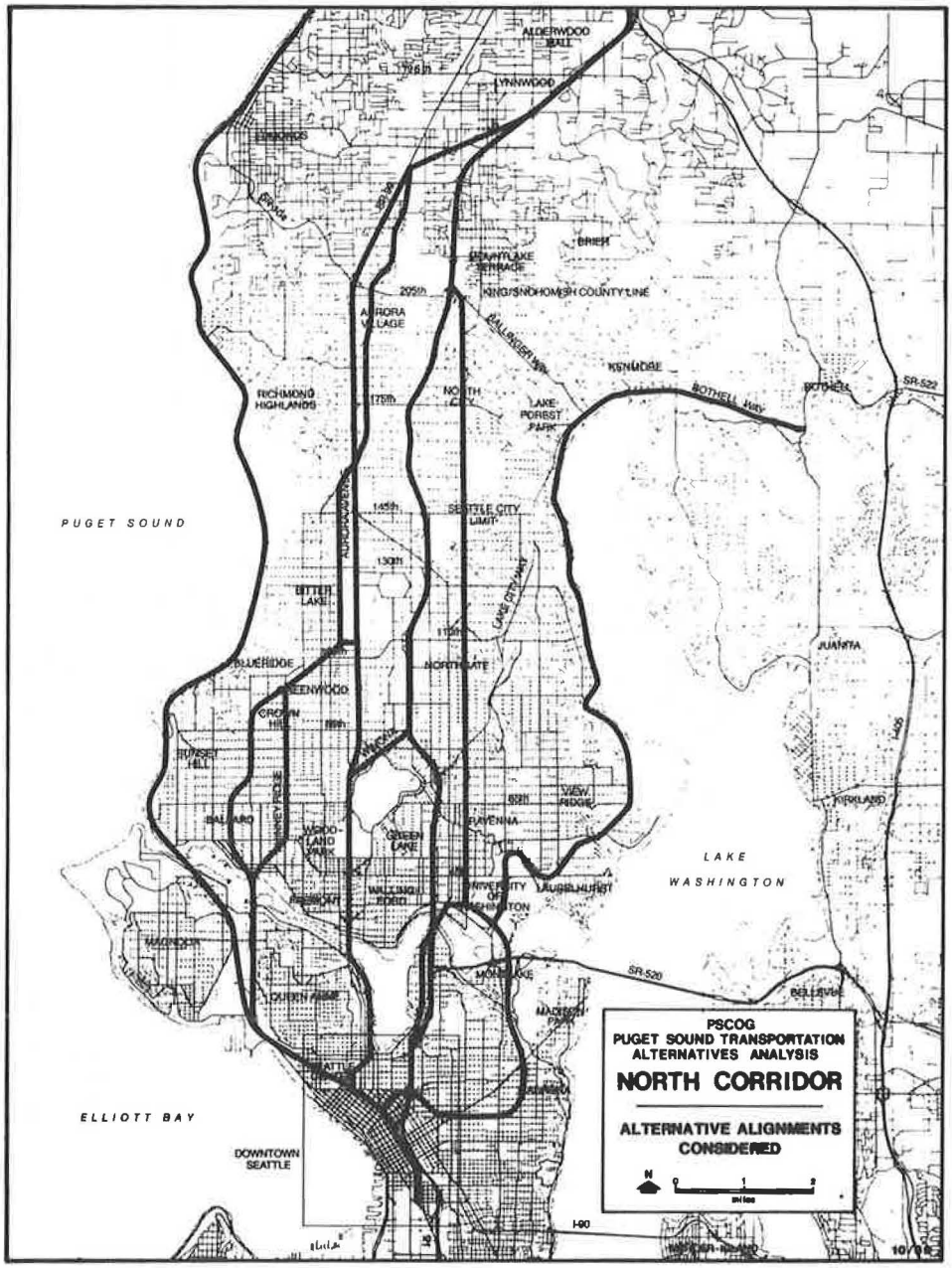

FIGURE 10 Alternative alignments considered.

Seattle, the percentage using transit ranges from 39.8 to 41.9 for the light rail alternatives. The advanced-technology bus captures 41.3 percent of the trips, and 40.8 percent of the trips are expected to use transit if the TSM alternative is implemented. Figure 13 is a graph showing the mode choice to the Seattle CBD on a daily basis. The build alternatives all increase the transit trips while they reduce the automobile trips relative to the no-build condition; the LRT alternative effects the greatest shift from automobile to transit.

In Figure 14 off-peak and peak-hour ridership are compared by alternative. The 4-hr peak shown combines the projected morning and afternoon ridership. One of the major differences between the LRT alternative and the advanced-technology bus alternative is demonstrated in Figure 14. The LRT alternative is a two-way operation throughout the day, improving the level of service not only for the commute to work but for midday trips as well. Trips to the University District would benefit from this service as would trips to south Snohomish County where employment is increasing rapidly. The advanced-technology bus uses the express lanes on I-5 inbound in the morning and outbound in the evening; although it improves operating speeds in downtown Seattle, it does not improve service in the off-peak direction.

EFFECT OF ALTERNATIVES ON VEHICULAR TRAFFIC VOLUMES

Tables 6 and 7 give daily and peak-hour traffic demand at selected locations in the North Corridor. Table 6 includes daily traffic volumes on all high-

	Alignment			Technology		Downtown		Length	
	I-5	I-5/A	AURORA	GB	LRT	SURFACE	TUNNEL	MAXIMUM	MINIMUM
1	●			●			●		●
2	●			●			●	●	
3	●				●		●		●
4	●				●		●	●	
5	●				●	●			●
6	●				●	●		●	
7		●			●		●		●
8		●			●		●	●	
9		●			●	●			●
10		●			●			●	
11			●		●	●			●
12			●		●			●	

FIGURE 11 Major investment alternatives considered.

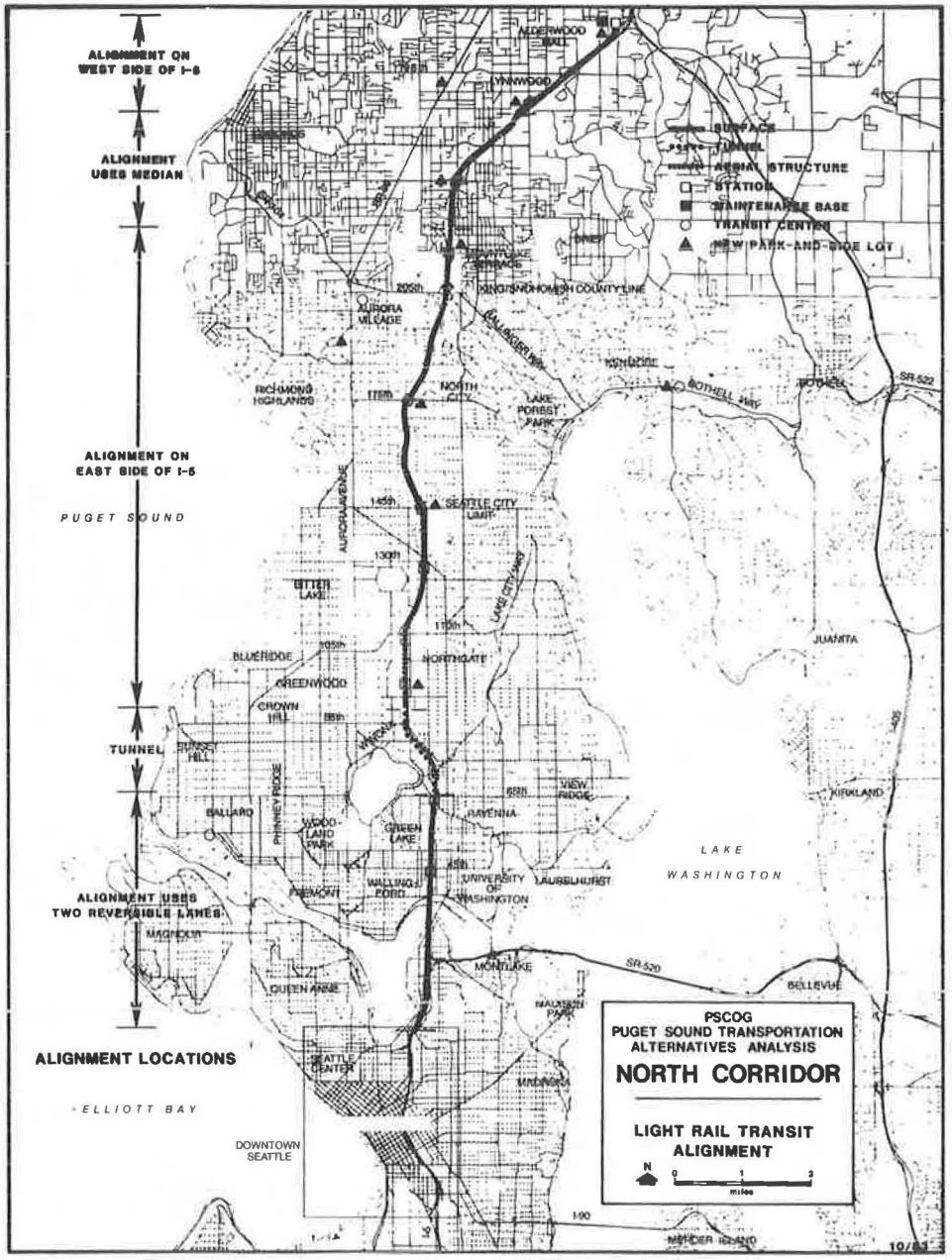

FIGURE 12 LRT alignment.

TABLE 4 North Corridor Alternatives Analysis: Transit Ridership in 1980 and 2000

AREA	EXISTING (1980)	YEAR 2000 NO-BUILD	TSM	A.T. BUS	I-5 LRT IN TUNNEL MIN.	IN TUNNEL MAX.	ON MALL SURFACE MIN.	ON MALL SURFACE MAX.
NORTH CORRIDOR								
Daily	142	215	234	242	241	248	238	244
Annual	44,700	67,700	73,700	76,200	75,900	78,100	75,000	76,900
Growth	NA	+52%	+65%	+71%	+70%	+75%	+68%	+72%
REGION								
Daily	210	325	352	360	359	366	356	362
Annual	66,100	102,400	110,900	113,400	113,100	115,300	112,100	114,000
Growth	NA	+55%	+68%	+72%	+71%	+74%	+70%	+73%

Note: Data are in thousands.

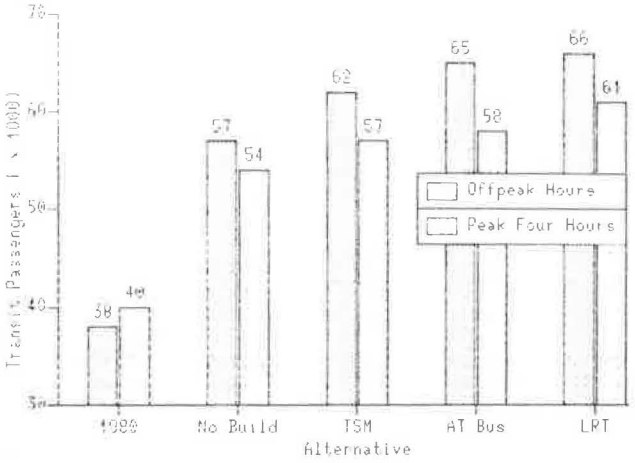

FIGURE 14 North Corridor transit volumes at ship canal screen line, 2000.

TABLE 5 North Corridor Alternatives Analysis: Percentage of Daily Person Trips on Transit in 2000

AREA	NO-BUILD	TSM	A.T. BUS	LRT DOWNTOWN TUNNEL MIN.	DOWNTOWN TUNNEL MAX.	DOWNTOWN SURFACE MIN.	DOWNTOWN SURFACE MAX.
NORTH CORRIDOR							
To Downtown	37.6%	40.8%	41.3%	40.2%	41.9%	39.8%	41.5%
Total	7.3%	8.2%	8.3%	8.5%	8.5%	8.4%	8.4%
REGION							
To Downtown	35.0%	37.9%	38.4%	37.9%	38.7%	37.7%	38.5%
Total	4.2%	4.5%	4.6%	4.6%	4.7%	4.6%	4.6%

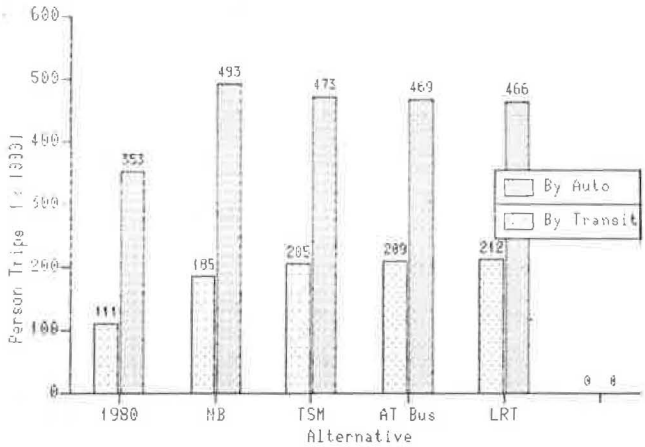

FIGURE 13 CBD mode choice, daily person trips in 2000.

ways and arterials across the ship-canal and 145th Street screen lines. Screen line 35 is at the ship canal, and screen line 41 is at 145th Street. At the ship canal, the advanced-technology bus and the TSM alternative are both expected to reduce traffic volumes from 432,000 (no-build condition) to 421,000 vehicles. A reduction to 418,000 vehicles is projected for the LRT alternative. Similar reductions

TABLE 6 North Corridor Alternatives Analysis: Daily Traffic Demand

SCREENLINE LOCATION	EXISTING	YEAR 2000 NO-BUILD	TSM	A.T. BUS	LRT TUNNEL	LRT SURFACE
Ship Canal	358	432	421	421	418	418
North of 145th	242	320	316	316	315	315

Note: Data are in thousands of vehicles.

TABLE 7 North Corridor Alternative Analysis: Peak-Hour Traffic Demand

SCREENLINE LOCATION		EXISTING	YEAR 2000 NO-BUILD	TSM	A.T. BUS	LRT TUNNEL	LRT SURFACE
Ship Canal	Volume	16	19	18	18	17	17
	Capacity	20	20	20	20	20	20
North of 145th[a]	Volume	8	11	11	11	10	10
	Capacity	9	9	9	9	9	9

Note: Data are in thousands of vehicles. Includes traffic on I-5 and Aurora Avenue only.
[a]Does not include the volume of capacity of diamond lanes reserved exclusively for transit or carpools.

in traffic volumes are forecast for the 145th Street screen line. Reductions in traffic at 145th Street are critical in 2000, because there are few alternative routes and congestion at this location causes severe bottlenecks throughout the corridor.

As indicated in Table 7, the TSM and advanced-technology bus alternatives are expected to remove 1,000 vehicles from peak-hour traffic crossing the ship canal on I-5 and Aurora, whereas 2,000 vehicles can be expected to be eliminated by the LRT alternative. Although the build alternatives do not eliminate traffic congestion during the peak period, they provide some relief. North of 145th Street on I-5 and Aurora Avenue, only the LRT alternative seems to eliminate a measurable number of vehicles from the road.

USE OF FORECASTS FOR DECISION MAKING

The patronage estimates and traffic impacts for each alternative are currently being reviewed by the steering committee, and during the coming months a preferred alternative will be selected. The steering committee has identified as the most important evaluation criterion the ability to maximize transit ridership in 2000 and afterwards. The ability to reduce traffic congestion ranks fourth. (The second- and third-ranked criteria were the ability to gain public support and the ability to expand the system into other corridors.) The committee members are studying the results of the travel forecasting process carefully and asking for tests of sensitivity of the forecasts to policy variables under their control. Additional analysis and refinement of the forecasts are likely to occur throughout the development of the draft environmental impact statement for the project.

The steering committee has requested information on patronage estimates for systems other than Seattle. Table 8 gives a summary of such information. The cities selected for inclusion in the table are those with light or heavy rail under study or development. Among the 10 cities, Seattle ranks second in percentage using transit to work, according to the 1980 census. (Among all U.S. cities with more than 1 million population, Seattle ranks 21st in population but 12th in percentage using transit for the journey to work.) Although it is difficult to determine whether the patronage estimates are directly comparable to one another, this survey of other studies indicates that peak-hour peak-direction volumes as well as daily volumes forecast for a major transit investment in Seattle's North Corridor equal or surpass those of other cities.

FUTURE TRAVEL FORECASTING EFFORTS IN SEATTLE

The North Corridor alternatives analysis has reinforced the need for constant review and update of the Puget Sound regional transportation data base and travel forecasting capabilities. Billion-dollar decisions are being made today concerning transit capital and operating programs for the next 20 years. The magnitude of these investments has brought policy-maker support for additional survey work and enhancements of the modeling process that should ensure the availability of adequate technical information as projects go into preliminary engineering.

ACKNOWLEDGMENT

Material presented in this paper is drawn from studies funded in part by grants from UMTA and FHWA administered through the Washington State Department of Transportation. The authors would like to thank Robert Shindler of PSCOG for his comments and suggestions.

TABLE 8 Comparison of Transit Patronage for Selected Cities

	1980 STATISTICS(a)			PATRONAGE FORECASTS(b)				
URBANIZED AREA(c)	POPULATION (millions)	% TRANSIT TO WORK	TRANSIT RIDERSHIP(d) (millions)	MODE UNDER STUDY OR DEVELOPMENT	FORECAST YEAR	CORRIDOR	PEAK-HOUR PEAK DIR.(e)	DAILY VOLUME(e)
Baltimore	1.91	10.6	114.8	Heavy Rail	1995	North/Metro Center	9,200	85,000
Seattle	1.38	10.1	85.0	Light Rail/ Guided Bus	2000	North	10,800	88,000
Portland	1.08	8.4	48.5	Light Rail	1990	Banfield Westside	6,400 6,300	47,800 53,100
Atlanta	1.62	7.4	99.0	Heavy Rail	1980 (Actual)	East Line	8,100	N.A.(f)
Denver	1.50	6.9	55.0	Light Rail	1990	West	8,700	50,000
Buffalo	1.06	6.6	37.0	Light Rail	2000	Amherst	8,200	68,000
Miami	1.61	6.4	76.6	Heavy Rail	2000	Systemwide	N.A.(f)	181,000
San Jose	1.25	3.2	31.7	Light Rail	1990	Guadalupe	7,800	40,000
San Diego	1.76	3.1	36.5	Light Rail	1995	"Tijuana Trolley"	1,600	45,000
Houston	2.61	2.8	42.8	Heavy Rail	1995	Westpark	12,000	141,900

(a) Sources: 1980 Census (5); APTA Transit System Operating Statistics Report (6).
(b) Sources: Draft Environmental Impact Statements; project staff.
(c) In order of percent transit to work.
(d) Unlinked passenger trips; all modes combined.
(e) At maximum load point.
(f) Not available.

REFERENCES

1. Regional Transportation Plan: Final Report. Puget Sound Council of Governments, Seattle, Wash., Sept. 1982.
2. Daniel, Mann, Johnson, and Mendenhall. Light Rail Transit Element: A Feasibility Assessment. Puget Sound Council of Governments, Seattle, Wash., March 1981.
3. Procedures for Alternatives Analysis: Detailed Outline (draft). Office of Methods and Support and Office of Planning Assistance, UMTA, U.S. Department of Transportation, April 14, 1983.
4. Background Report: North Corridor. Puget Sound Council of Governments, Seattle, Wash., Dec. 1982.
5. Census of Population and Housing: 1980. Summary Tape File 3-A. U.S. Bureau of the Census, 1981.
6. Transit System Operating Statistics for Calendar/Fiscal Year 1980. American Public Transit Association, Washington, D.C., 1981.

Predicting Travel Volumes for High-Occupancy-Vehicle Strategies: A Quick-Response Approach

THOMAS E. PARODY

ABSTRACT

The development of a set of demand and supply models that predict peak-hour travel volumes for high-occupancy-vehicle (HOV) strategies on freeways is described. The demand models were estimated by using a consistent series of before-and-after empirical data from a number of actual HOV facilities located across the United States. Supply models were developed on the basis of speed-volume relationships that estimate changes in running speeds and travel times on the general-purpose lanes for different volume levels and capacity configurations. These models have been incorporated into a set of easy-to-use worksheets to predict equilibrium travel flows of vehicles on the general-purpose freeway lanes and of carpools and buses on the HOV lane or lanes. The models forecast the net change in volume due to mode shift, time of day, trip generation, and route diversion behavior. Consequently, the models provide more information on anticipated travel impacts than can be obtained by using mode-choice models alone. Because the forecasting procedure is designed to provide quick-response results, data requirements are minimal and these data should be readily available to most planning agencies. The accuracy of the forecasting procedure should be interpreted as sketch-planning-level responses that, if conditions warranted, would be subjected to additional and possibly more refined analyses. However, test applications of the prediction procedures described yielded favorable results. Using only data collected before HOV facilities were established, average errors across the HOV sites were less than 4 percent for the nonpriority automobile and HOV bus modes and less than 14 percent for the priority automobile and carpool mode.

Priority treatments for high-occupancy vehicles (HOVs) are transportation system improvements that have proved to be highly cost-effective solutions to meeting urban transportation needs in selected cities across the United States. Given current constraints on constructing new highways in urban areas, such low-capital projects as HOV facilities could become even more popular during the next decade. Although there currently exist computerized models such as the Urban Transportation Planning System (UTPS) that can be used to forecast the travel impacts of alternative HOV treatments, there has historically been little development of approaches that can be used expressly for evaluating HOV strategies in a quick-response time frame.

In this paper the development and testing of a travel forecasting procedure designed specifically for predicting travel volumes resulting from the implementation of priority treatments for HOVs on freeways are described. The procedures developed and described in this paper are intended to be implementable in the face of severe constraints on turnaround time, data availability, and computational resources, while at the same time providing information that is both accurate and easy to obtain. To meet this quick-response capability, forecasts of peak-hour volumes (i.e., for nonpriority automobiles, carpools, and bus transit) can be made by using an ordinary hand-held calculator and a set of worksheets that contain the demand, supply, and equilibrium procedures that were developed.

A comprehensive review of current forecasting procedures revealed that no existing travel demand models have been estimated using actual before-and-after data from the broad cross section of HOV dem-

onstrations and projects supported by the U.S. Department of Transportation since about 1970. Consequently, a consistent set of empirical before-and-after data from HOV sites across the United States was collected and used to estimate models (or relationships) that were packaged in a set of worksheets. These worksheets can be used to make quick forecasts of mode shares and travel volume changes that may be expected from implementing various HOV strategies.

The initial plan was to develop one or more models that could be used to evaluate six different types of freeway or arterial HOV treatments. However, after the data available were examined, it was determined that models could only be estimated for freeway-based HOV sites. Consequently, the travel forecasting procedures reported here were developed to analyze the following four HOV freeway strategies:

1. Dedicate a new or existing lane for bus-only HOV operation;
2. Dedicate a new or existing lane for bus and carpool operation;
3. Allow carpools onto an existing bus-only HOV lane; and
4. Allow carpools with lower occupancy levels onto an existing bus and carpool HOV lane.

DATA COLLECTION FOR MODEL DEVELOPMENT

The principal sources of before-and-after data consisted of evaluation reports of HOV demonstrations that have been implemented during the past 15 years. The first of these large-scale demonstrations began in 1970 with the announcement of the Urban Corridor Demonstration Program. This program was directed toward reducing commuter corridor congestion through the implementation of projects that encourage transit or carpool ridership (e.g., reserved HOV lanes) or that increase the efficiency of existing street systems (1). Later, in 1974, FHWA established Research Project 2D, entitled Priority Techniques for High Occupancy Vehicles, as part of its Federally Coordinated Program (FCP). The objective of the 2D project was to increase the people-moving efficiency of the highway system by (a) applying a variety of techniques for the preferential treatment of HOVs (buses, vanpools, carpools); (b) thoroughly evaluating these techniques with respect to benefits, costs, environmental impacts, and institutional and public acceptance; and (c) providing all information necessary to facilitate wider implementation of the most promising techniques.

Finally, in 1974 UMTA began the Service and Methods Demonstration (SMD) Program to provide a consistent and comprehensive framework within which to formulate, implement, evaluate, and disseminate results of demonstrations including, among other techniques, HOV strategies (2). As a result of the many demonstrations sponsored by these programs, there exists a large body of quasi-experimental observations concerning the impacts of different HOV alternatives. These projects represent a prime source of data that, to date, has not been used in a comprehensive manner either to validate the efficacy of existing travel demand models or to develop new models for the prediction of travel flows resulting from a range of HOV alternatives.

To estimate the proposed HOV demand models, the following set of data was required for at least two time periods for each site: volume of vehicles (or persons) traveling on the general-purpose and priority lanes, travel speeds and times of vehicles on the general-purpose and priority lanes, HOV length, and roadway geometric descriptions (i.e., number of lanes or capacity or both). The first time period represents conditions before the implementation of the HOV treatment in question, whereas the second period reflects conditions approximately 1 year after the implementation of the HOV strategy. If an HOV treatment had been implemented in phases (e.g., bus only with the later inclusion of carpools), additional (before and after) periods were included. Typically, data were tabulated for conditions representing the peak hour in the peak direction. To the extent possible, the morning peak-hour period and direction were used for consistency across HOV sites.

The key before-and-after modal volumes and level-of-service characteristics that were obtained for 12 freeway HOV facilities (or phases) are given in Table 1 (3). Table 2 summarizes the HOV treatments that were already in operation and the new ones implemented. For each site, Table 1 presents the following information:

1. Nonpriority automobile volumes and capacity: Peak-hour (a.m.) volume of automobiles not eligible to use the HOV facilities and capacity of general-purpose lanes for both time periods.
2. Priority automobile volumes: Peak-hour (a.m.) volume of automobiles eligible to use the HOV facility for both time periods.
3. Transit ridership: The volume of bus riders in the morning peak hour who use the HOV lanes in both time periods.
4. Average total travel time: Average total travel time in minutes in general-purpose and HOV lanes for both time periods.
5. Length of HOV lane or lanes in miles.
6. HOV time advantage: The travel time saved by using the HOV facility compared with that by using the non-HOV lanes in the after period (i.e., nonpriority after time minus priority after time).
7. Change in nonpriority in-vehicle travel time (ΔIVTT) from the before to the after period for nonpriority automobiles.
8. Change in priority in-vehicle travel time (ΔIVTT) from the before to the after period for the priority-eligible vehicles (in two instances it includes the change in bus travel time while buses were already on the HOV lane in the before period).
9. Speed: Average speed on the HOV section of roadway for vehicles in the general-purpose and HOV lanes in both time periods.
10. Average trip length: Average trip distance in miles of all users on the HOV roadway.

Seven of the before-and-after sets of data represent the initial start of an HOV priority facility. Four others (Shirley Highway; San Bernardino, phase 2; US-101, phase 2; and I-95, phase 1) represent a change in the HOV facility from bus-only to mixed-mode (carpool and bus) operation. Finally, Banfield Freeway, phase 2, and Miami I-95, phase 2, involved allowing carpools of two or more onto an existing lane that previously allowed buses and carpools of three or more.

DATA LIMITATIONS

The data presented in the evaluation reports for the various HOV sites were not always consistent with the reporting format that was established. Therefore it was sometimes necessary to make adjustments to or estimates from the data presented if no other information was available. These types of calculations were typically required for the following three data items.

Travel Volume

Automobile and transit volumes were frequently presented in the evaluation reports for the full a.m. or p.m. peak periods or both. Therefore to obtain an estimate of the peak 1-hr volumes, the peak-period volumes were divided by typical peaking factors.

Average Total Trip Length

Average trip length was required in some instances to determine average total travel time. When information on trip length was not presented in a report, the agency or organization in charge of the facility was contacted to determine whether the information was available from other sources. In one instance, census information on trips made to the central business district from different zones in the study corridor was used.

Total Travel Time

For almost every site, information was available on the change in travel time in both the general-purpose and HOV lanes because of the implementation of the HOV project. This information was useful in determining the before or after total travel times or both if these data were not otherwise available. For example, if a total travel time estimate was available for the before period but not for the after period, the data on travel time change were used to estimate total travel time for the after period. When neither the before nor the after travel time was reported, an estimate of one value was made using average trip length and speed, and the value for the second time period was computed using the known change in travel time.

DEMAND MODEL DEVELOPMENT

The first consideration in estimating the various HOV demand models was the specification of the (dependent) variable being forecast, that is, whether travel volumes for each mode should be expressed in terms of person or vehicle volumes. The advantage of using persons is that one can examine directly the peak-hour person throughput of a given freeway for different types of HOV strategies. However, the major drawback of this approach is in travel equilibration on the general-purpose lanes, because highway supply relationships are expressed in terms of vehicles. Consequently, it was decided to use vehicles per hour as the measure of travel volume for various classifications of automobiles but to use person trips for bus transit, because equilibration is not an issue in predicting bus demand.

For each mode (nonpriority automobile, priority-eligible automobile or carpool, and HOV bus), models were estimated based on both linear and product specifications of the independent variables. In addition, volumes and level-of-service variables were entered by using either absolute differences or relative differences. Although one functional form did not dominate the others for all HOV sites, the model form that produced the most favorable results for all modes can be expressed as follows:

$$(V_1^m - V_0^m)/V_0^m = a_0 + \sum_i a_i[(T_1^i - T_0^i)/T_0^i] \quad (1)$$

where

V_0^m = peak-hour volume during before period for mode m,
V_1^m = peak-hour volume during after period for mode m,
T_0^i = travel times during before period for modes i to m,
T_1^i = travel times during after period for modes i to m, and
a_0, i = calibration coefficients.

In the following sections results are presented of travel demand models estimated using Equation 1 for the nonpriority automobile, priority automobile, and priority bus modes.

Nonpriority Automobile Model

In Table 3 the parameter estimates, t-statistics, and associated regression results for the nonpriority automobile model are presented. All signs for the parameter estimates are correct and all are significant at the appropriate levels. The R^2 for the entire model is 0.98, with an F-ratio significant at the 99 percent level.

A generalized least-squares estimation procedure was used for model estimation. Basically, this entailed multiplying all variables for each site by the square root of the sum of the travel volumes from the before and after periods. This procedure was used initially because it was suspected that variances in peak-hour volumes are less as travel volumes increase.

In Table 3 the percentage change in total travel time for nonpriority automobiles [i.e., $(T_1^{npa} - T_0^{npa})/T_0^{npa}$] is represented by NPA-TT. Similarly, the percentage change in travel time for two-person carpools is CP2-TT. If two-person carpools are not allowed onto the HOV lanes, this variable will take on the same value as that of NPA-TT. The percentage change in total travel time for carpools of three or four or more persons is CP3/4-TT. Again, the value of this variable will be the same as that of NPA-TT if these carpools are not allowed to use the HOV lanes; alternatively, they will have different values if carpools of three or four or more are already on, or will be allowed on, the HOV lanes. The percentage change in total travel time for buses that are already on, or will be allowed on, the HOV lanes is Bus-TT. The variable that reflects the percentage change in capacity on the general-purpose lanes made available in the after period for use by nonpriority automobiles is the eligibility factor (EFCTR), which is computed as follows:

$$EFCTR = (L_1^{GP}/L_0^{GP}) \cdot [(V_0^{npa} + V_0^{Pa} + 2B_0^{Peb})/V_0^{npa}] \quad (2)$$

where

L_0^{GP} = number of general-purpose lanes in the before period,
L_1^{GP} = number of general-purpose lanes in the after period,
V_0^{npa} = peak-hour volume of nonpriority automobiles in the before period,
V_0^{Pa} = peak-hour volume of priority-eligible automobiles in the before period, and
B_0^{Peb} = number of buses eligible to move to HOV lanes.

If no automobiles (carpools) or buses are allowed to move to the HOV lane and the number of general-purpose lanes does not change, EFCTR will equal 1.0. If, for example, 10 percent of the total number of

TABLE 1 Summary of Key Data for Freeway HOV Sites *(3)*

HOV Facility	Volume (vehicles per hour)						Transit Ridership (persons per hour)		Avg Travel Time by Lane (min)				HOV Length (miles)
	Nonpriority Automobiles				Priority Automobiles				General Purpose		HOV		
	Before	Capacity	After	Capacity	Before	After	Before	After	Before	After	Before	After	
Shirley Highway	4,896	5,880	5,126	5,880	195	758	7,900	8,756	56.2	58.3	37.5	38.3	9.
San Bernardino													
Phase 1	7,300	7,500	7,300	7,500	NPA	NPA	402	1,017	44.8	45.3	NPL	36.9	7.
Phase 2	7,067	7,500	7,277	7,500	299	576	2,490	2,708	45.7	46.1	34.0	35.3/ 34.0[a]	7/11[a]
US-101													
Phase 1	5,314	5,558	5,330	5,616	NPA	NPA	3,370	3,572	35	35	NPL	32.2	3.8
Phase 2	5,125	5,616	5,333	5,616	205	288	3,572	3,686	35	33.1	32.2	32.2	3.8
Banfield Freeway													
Phase 1	3,713	3,900	3,845	3,900	37	180	340	570	21.1	21.1	NPL	19.7	3.3
Phase 2	3,161	3,900	3,793	3,900	530/ 178[b]	1,107/ 163[b]	628	657	22.7	21.4	20.9	20.9	3.3
I-95, Miami													
Phase 1	6,145	7,200	6,416	7,200	170	309	274	314	36.9	34.5	33.4	31.3/ 32.9[a]	7.5
Phase 2	5,170	7,200	5,880[c]	7,200	1,246/ 309[b]	1,357/ 246[b]	314	352	34.5	34.7	31.3/ 32.9[a]	31.3/ 32.9[a]	7.5
Southeast Expressway													
1977	5,504	7,000	4,306	5,300	388	641	2,000	2,124	35	43	NPL	25.0	8.
1971	4,554	5,300	4,201	5,300	NPA	NPA	2,152	2,454	35	30.5	NPL	23.0	8.4
I-495, Lincoln Tunnel	2,324	5,200	3,227	5,400	NPA	NPA	21,868	26,092	70	68	NPL	60	2.5

Note: NPA = no priority automobiles included in HOV treatment. Numbers not rounded off to significant digits. NPL = no priority lane.
[a]Priority automobile or priority bus.
[b]Two-occupant or three-occupant carpool.
[c]High violation rate.

TABLE 2 HOV Treatment for Before and After Time Periods

HOV Facility	HOV Treatment	
	Before Period	After Period
Shirley Highway	Bus only	Carpools of four or more added
San Bernardino		
Phase 1	No priority	Bus only
Phase 2	Bus only	Carpools of three or more added
US-101		
Phase 1	No priority	Bus only
Phase 2	Bus only	Carpools of three or more added
Banfield Freeway		
Phase 1	No priority	Carpools of three or more and buses
Phase 2	Buses and carpools of three or more	Carpools of two or more added
I-95 Miami		
Phase 1	Bus (on NW 7th Avenue)	Buese and carpools of three or more added to I-95
Phase 2	Buses and carpools of three or more	Carpools of two or more added
Southeast Expressway		
1977	No priority	Buses and carpools of three or more
1971	No priority	Bus only
I-495, Lincoln Tunnel	No priority	Bus only

TABLE 3 Nonpriority Automobile Model: Regression Results

Variable	Parameter Estimate	t-Statistic	Level of Significance
Constant	-0.916	10.5	0.01
NPA-TT	-1.053	-3.3	0.01
CP2-TT	+1.190	+3.5	0.01
CP3/4-TT	+0.122	+1.4	0.10
Bus-TT	+0.278	+3.8	0.01
EFCTR	+0.949	+12.1	0.01

Note: NPA-TT = percentage change in total travel time for nonpriority automobiles, CP2-TT = percentage change in total travel for two-person carpools, CP3/4-TT = percentage change in total travel time for carpools of three or four or more, Bus-TT = percentage change in total travel time for buses, EFCTR = eligibility factor.

Summary statistics: F-ratio = 53.1, significance = 0.01, R^2 = 0.98, DFE = 6, MSE = 0.0007.

automobiles using the general-purpose lanes in the before period become eligible to use the HOV lanes, EFCTR will equal 1.11. If one of four general-purpose lanes is taken away for use by HOVs, the value of EFCTR will be reduced to 75 percent (i.e., 3 ÷ 4) of its value if the lane were not taken away. Thus, this variable controls for site-to-site differences in the percentage of vehicles from the before period that become eligible to use an HOV facility during the after period. In addition, the variable reflects the major supply effects due to using a general-purpose lane only for HOV vehicles.

Priority Automobile Model

Table 4 gives the parameter estimates, t-statistics, and associated regression results for the priority automobile models that were obtained by using a generalized least-squares estimation procedure. All parameter estimates have the correct sign and all are significant (PA2-TT is, however, only significant at the 0.12 or 88 percent level). The respective percent changes in total travel time for two- or three-or-four-or-more person automobiles that are already on or will be allowed on the HOV facility are PA2-TT and PA3/4-TT. Thus, the model or models shown in Table 4 can be used to forecast volumes for

TABLE 4 Priority Automobile Model: Regression Results

Variable	Parameter Estimate	t Statistic	Level of Significance
Constant	-0.2	-0.6	0.28
PA2-TT x [Q]	-6.7	-1.4	0.12
PA3/4-TT x [1 - Q]	-7.7	-4.1	0.01
Bus-TT	+4.8	+2.3	0.03

Note: PA2-TT = percentage change in total travel time for two-person priority automobiles, PA3/4-TT = percentage change in total travel time for three-or-four-or-more-person priority automobiles, Bus-TT = percentage change in total travel time for buses.

Summary statistics: F-ratio = 8.1, significance = 0.02, R^2 = 0.87, DFE = 5, MSE = 0.231, where Q = 1 for two-person priority automobiles and Q = 0 for three-or-four-or-more-person priority automobiles.

Travel Time (min)			Speed by Lane (mph)				One-Way Avg Trip Length (miles)	No. of Buses per Hour	
HOV Time Advantage (after)	Nonpriority ΔIVTT	Priority ΔIVTT	General Purpose		HOV				
			Before	After	Before	After		Before	After
20.0	+2.1	-17.9/ + 0.8[a]	19.0	17.7	55.5	51.5	12.4	176	194
8.4	+0.5	- 7.9	25.4	24.7	NPL	49	19.8	10	45
10.8	+0.4	-10.4	24.2	23.6	49	55/ 49[a]	19.8	81	81
2.8	0	- 2.8	30.0	30.0	NPL	47.1	16	86	94
0.9	-1.9	- 2.8	30.0	40.0	47.1	47.1	16	94	97
1.4	0	- 1.4	38	37.9	NPL	51.5	7	10	20
0.5	-1.3	- 1.8	33	42	47	46.7	7	20	22
3.2/ 1.6	-2.4	- 5.6/ - 0.5[a]	33.2	40.2	44.7	56.6/ 46.7	15.7	10	10
3.4/ 1.8	+0.2	- 3.2	40.2	39.4	56.6/ 46.7	56.6/ 46.7	15.7	10	10
18	+8.0	-10.0	21.0	15.5	NPL	37.2	15	50	54
7.5	-4.5	-12.0	23	29	NPL	50.4	15	57	65
8.0	-2.0	-10.0	10	11.5	NPL	30	25	497	597

carpools that are already on the HOV facility or that will become eligible to use the facility in the after period.

If the model is being used to forecast the number of carpools of three or more persons that will be using the HOV lanes, the variable PA2-TT x [Q] is deleted (or set equal to zero). Conversely, if the model will be used to forecast the volume of two-person priority-eligible automobiles, the variable PA3/4-TT is set equal to zero. (Note: This model cannot be used to forecast the volume of two- or three-or-four-or-more-person carpools that will be traveling on the general-purpose lanes in the after period.)

Because the coefficient of the three-or-four-or-more-person travel time variable is larger than that for two-person carpools, the model indicates that allow- ing carpools of three or more onto the HOV lanes will lead to a larger percentage increase in the volume of carpools of three or more relative to the percentage increase in carpools of two or more if these are granted access to the HOV facility. (Note that although this is true in percentage terms, it may not always be true in absolute terms, because the volume of two-person automobiles is usually much greater than the volume of three-person automobiles.)

The magnitudes of both carpool travel time coefficients (which are related to direct travel time elasticities) are much larger than those derived from traditional or contemporary mode-choice studies. The reason for this is that the priority automobile model is capturing the effects of trip generation, time of day, and route diversion changes as well as modal choice. Thus, a model that examined only mode-choice effects would seriously underpredict the actual volume of carpools on the HOV facility.

The variable Bus-TT is the percentage change in bus travel time between the before and after periods. Note that its magnitude is about two-thirds the size of the carpool travel time coefficients. If buses are already on the HOV facility and the policy being examined is to allow carpools onto the facility, the value of this variable will normally take on a value of zero (assuming, as is typically the case, no degradation in travel speeds on the HOV facility), and the percentage change in the volume of carpools will be a function of the percentage change in carpool travel times. However, if buses and carpools are being granted the use of the HOV facility at the same time (i.e., the facility did not exist in the before period), the model indicates that the percentage increase in carpool volume will be reduced by between one-third to one-half (depending on the size of the travel time savings) compared to the case in which carpools only were granted access. Again, this appears appropriate, because the bus mode will also be competing for some of the travelers who may wish to use the HOV facility.

Priority Bus Models

After alternative specifications for the priority bus model had been evaluated, it was determined that the most appropriate procedure for modeling changes in bus ridership was to use different variable specifications, depending on whether buses or carpools or both are allowed onto the HOV lanes and whether bus supply is determined exogenously or endogenously. A single model specification does not adequately explain the change in transit ridership for both bus-only and bus-or-carpool strategies.

In Table 5 the parameter estimates, t-statistics, and associated regression results are presented for the priority bus models that were estimated using a generalized least-squares estimation procedure. As indicated, model A is used when only buses will use the HOV lane and bus supply is determined endogenously or as a direct result of the HOV time savings. The one variable that was found to be significant was Bus-TT, the percentage change in bus travel time. In effect, the estimation process revealed that changes in nonpriority automobile travel time have little or no explanatory power compared with changes in bus travel time. Because of the small sample size, however, the coefficient estimate is only significant at the 83 percent level. Unfortunately, consistent data on other factors (besides travel times) that could affect bus ridership were not readily available.

Model B is used when only buses will use the HOV lane and supply is determined exogenously or apart from the ridership change expected just from the HOV

TABLE 5 Priority Bus Models: Regression Results

Model	Variable	Parameter Estimate	t-Statistic	Level of Significance
A	Bus-TT	-1.404	-1.1	0.17
B	Bus-TT	-0.308	-2.3	0.07
	Bus-No.	+0.422	26.7	0.01
C or D	Constant	+0.227	1.2	0.14
	PA2-TT x [Q]	+1.710	0.5	0.30
	PA3/4-TT x [1 - Q]	+0.435	0.5	0.30

Note: A = bus only on HOV lane (supply determined endogenously), B = bus only on HOV lane (supply determined exogenously), C = bus and carpools of three or more or four or more on HOV lane (Q = 0), D = bus and carpools of two or more on HOV lane (Q = 1). Bus-TT = percentage change in total travel time for buses, Bus-No. = percentage change in the number of peak-hour buses, PA2-TT = percentage change in total travel time for two-person priority automobiles, PA3/4-TT = percentage change in total travel time for three-or-four-or-more-person priority automobiles.

Summary statistics are tabulated as follows:

Parameter	Model A	Model B	Model C or D
F-ratio	1.2	505.1	1.1
Significance	0.35	0.01	0.46
R^2	0.28	0.99	0.44
DFE	3	2	4
MSE	0.587	0.002	0.066

time savings (e.g., for phase 1 of the San Bernardino project, the El Monte bus terminal was constructed, and in the after period, the number of buses per hour was increased by 350 percent, from 10 to 45). The two variables, percentage change in bus travel time and percentage change in the number of peak-hour buses, have the correct sign and are significant at appropriate levels. [The bus supply variable (Bus-No.) representing the percentage change in the number of peak-hour buses was not used in the other models because of concern for simultaneity. This occurs because bus supply is highly correlated with the dependent variable, bus passengers.]

Model C has a constant and a term for the percentage change in total travel time for priority automobiles with three or four or more persons. Consequently, this model is used when buses and three-or-four-or-more-person carpools are allowed onto the HOV lane. Model D has the same constant but uses the percentage change in travel time for two-person priority automobiles to forecast the volume of bus passengers when buses and two-person carpools will be using the HOV lane. Although the signs for all variables are correct, the significance levels are lower (70 percent) than those typically desired. Thus, the higher standard errors for these coefficients imply greater variances in the forecast of percentage change in bus riders. However, in many instances, the percentage change in bus ridership is relatively small (especially compared with changes in carpools); thus the effects of these larger variances are partially negated.

Unlike automobile and carpool volumes, changes in the volume of bus users are more likely to be dependent on many more site-specific characteristics in addition to changes in level of service (as represented by total travel time changes). Some of these other factors, which are difficult to incorporate in a sketch-planning model, would include average bus headways, average waiting and transfer times, characteristics of bus area coverage or route network, and provision of fringe parking lots for park-and-ride express bus service. Thus, the analyst is reminded that although the priority bus models may provide forecasts that reflect average conditions observed at other HOV sites, the results may not be the most applicable to the HOV facility being evaluated. In such instances, more complex procedures may be required to predict ridership on HOV buses.

SUPPLY MODEL DEVELOPMENT

Commensurate with the level of detail of the demand models, a supply model was developed to estimate average running speed and thus travel time changes for different volume levels (and possibly capacity changes) on the general-purpose lanes. The model was based on the Bureau of Public Roads and FHWA speed-volume relationship normally used in traffic assignment models (4). This relationship can be expressed in general terms as follows:

$$T_1 = T_0 \, [1 + a(V/C)^b] \quad (3)$$

where

T_1 = travel time in time period 1,
T_0 = travel time at zero volume (or under free-flow conditions),
V = highway traffic volume,
C = capacity of highway, and
a,b = model coefficients.

For the 12 HOV data sets used in developing the demand models, it was observed that traffic on the general-purpose lanes in the before period was operating either at or near capacity (service level E) or, more commonly, under force-flow conditions (service level F). One of the key questions is whether (and how) these service levels will change, given the implementation of a particular HOV strategy. By analyzing before-and-after service levels for various HOV freeway facilities, it was determined that force-flow conditions continued in the after period when (a) a general-purpose lane was taken away or (b) the number of general-purpose lanes did not change but a bus-only HOV lane was implemented. When the number of general-purpose lanes remained the same and carpools were allowed onto the HOV lane or lanes, traffic on the general-purpose lanes either continued operating under force-flow conditions or began operating under free-flow conditions. These observations, therefore, were incorporated into a straightforward procedure for computing supply changes due to various HOV strategies on freeways.

MODEL TESTING

The demand and supply models described in the foregoing have been incorporated into a set of seven worksheets that can be used in a sequential and if necessary iterative fashion to predict equilibrium travel volume resulting from any one of four types of HOV strategies. The flowchart in Figure 1 highlights the major activities for each worksheet. [Additional details on the use of the worksheets are presented in the User's Guide (5).]

With these worksheets, forecasts of peak-hour volumes were made by using as input data known service-level changes. For each site the forecast volumes are compared with the actual volumes for the after period, and a relative error or difference is computed and listed in Table 6. Also given in Table 6 are the average relative errors across all sites for the three modes. It is readily apparent that average errors and standard deviations are quite small for both the nonpriority automobile and the HOV bus modes. The largest errors occurred for the priority-granted carpool mode (-7.7 percent). Of course, more reliable forecasts are to be expected, because the model coefficients were estimated by using the same before-and-after data used in forecasting. Even so, the model is able to capture other

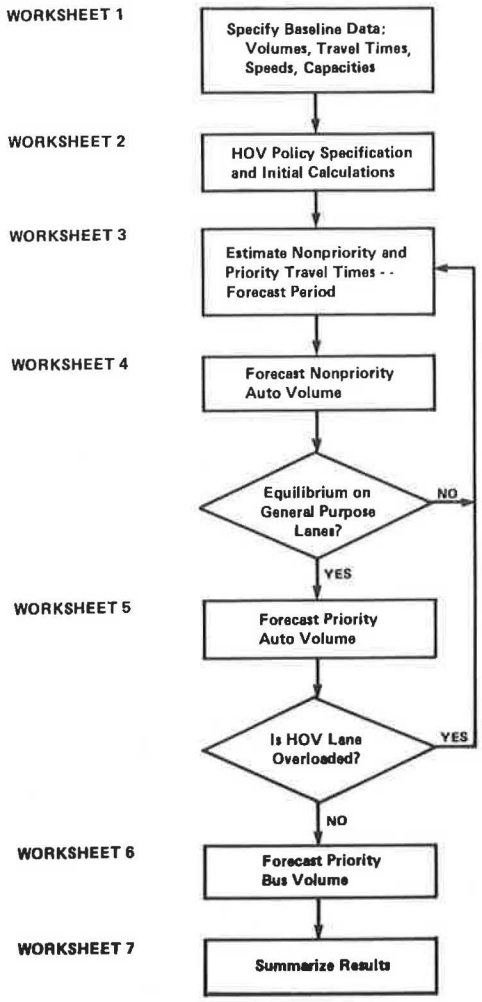

FIGURE 1 Activities undertaken in each worksheet.

effects in addition to shifts between modes. Clearly, this is a decided advantage compared with mode-choice (or general-share) models, given the importance of time-of-day and route diversion impacts. [See the report by Charles River Associates (3) for a comparison of the abilities of a pivot-point logit model to predict travel volumes for HOV facilities.]

In order to test the demand, supply, and equilibration components of the HOV worksheets, forecasts were also made by using only before data for each HOV site. (The one exception was to use the number of buses after the change in HOV facility for the San Bernardino, phase 1, site.) The forecasts and relative percentage errors are given in Table 7. The average errors increased slightly for the nonpriority automobile and priority bus modes but by a somewhat larger amount for the priority automobile mode.

The relatively higher standard deviations for the priority automobile or carpool forecasts indicate that other factors (either measurable, site-specific, or unobservable) in addition to those included in the model may influence the volume of carpools on the HOV facility. A lack of information describing the before-and-after characteristics of alternative highways in the HOV corridor prohibited a systematic examination of these effects within the context of this study. The tests reported earlier, however, clearly illustrate that the HOV models and worksheets can be used to provide a quick and reasonable examination of travel flows due to implementing alternative HOV strategies on freeways.

EPILOGUE

All the existing HOV sites that were used in developing the relationships embedded in the forecasting approach share some common characteristics that help define the type of HOV treatments that can best be analyzed with the procedures presented here. First, the HOV lanes operate on (or adjacent to) major radial freeways leading into a central city or central business district. Thus, the proposed HOV corridor or lane should have similar characteristics (i.e., the approach may not yield reliable results

TABLE 6 Comparison of HOV Worksheet Predictions to Actual Travel Volumes: Actual Level-of-Service Changes Used

	Nonpriority Automobile Volumes				Priority Automobile Volumes				HOV Bus Ridership			
	Actual			Relative Error[a]	Actual			Relative Error[b]	Actual			Relative Error[c]
HOV Facility	Before	After	Predicted After	[(P - A)/A]	Before	After	Predicted After	[(P - A)/A]	Before	After	Predicted After	[(P - A)/A]
Shirley Highway	4,896	5,126	5,105	-0.4	195	758	652	-13.9	7,900	8,756	8,595	-1.8
San Bernardino												
Phase 1	7,300	7,300	7,221	-1.1	NPA	NPA	NPA	–	402	1,017	1,018	+0.1
Phase 2	7,067	7,277	7,394	+1.6	299	576	761	+32.1	2,490	2,708	2,807	+3.7
US-101												
Phase 1	5,314	5,330	5,533	+3.8	NPA	NPA	NPA	–	3,370	3,572	3,748	+4.9
Phase 2	5,125	5,333	5,399	+1.2	205	288	289	+0.4	3,572	3,686	4,257	+15.5
Banfield Freeway												
Phase 1	3,713	3,845	3,790	-1.4	37	180	37	-79.4	340	570	407	-28.6
Phase 2	3,161	3,793	3,659	-3.5	708	1,180	846	-28.3	628	657	685	+4.3
I-95, Miami												
Phase 1	6,145	6,416	6,332	-1.3	170	309	321	+4.0	274	314	318	+1.3
Phase 2	5,170	5,880	5,919	+0.7	1,555	1,603	2,013	+25.6	314	352	335	-4.7
Southeast Expressway												
1977	5,504	4,306	4,269	-0.8	388	641	629	-1.9	2,000	2,124	2,205	+3.8
1971	4,554	4,201	4,226	+0.6	NPA	NPA	NPA	–	2,152	2,454	3,188	+29.9
I-495, Lincoln Tunnel	2,324	3,227	3,234	+0.2	NPA	NPA	NPA	–	21,868	26,092	26,254	+0.6

Note: NPA = no priority automobiles included in HOV treatment, A = actual, P = predicted. Data are from Charles River Associates.

[a]Average error = -0.03 percent, SD = 1.8 percent.
[b]Average error = -7.7 percent, SD = 34.9 percent.
[c]Average error = 2.4 percent, SD = 13.4 percent.

TABLE 7 Comparison of HOV Worksheet Predictions to Actual Travel Volumes: Only Before Data Used

HOV Facility	Nonpriority Automobile Volumes				Priority Automobile Volumes				HOV Bus Ridership			
	Actual		Predicted	Relative Error[a]	Actual		Predicted	Relative Error[b]	Actual		Predicted	Relative Error[c]
	Before	After	After	[(P - A)/A]	Before	After	After	[(P - A)/A]	Before	After	After	[(P - A)/A]
Shirley Highway	4,896	5,126	5,045	-1.6	195	758	654	-13.7	7,900	8,756	8,550	-2.4
San Bernardino												
Phase 1	7,300	7,300	7,195	-1.4	NPA	NPA	NPA	–	402	1,017	1,024	+0.7
Phase 2	7,067	7,277	7,415	+1.9	299	576	680	+18.0	2,490	2,708	2,846	+5.1
US-101												
Phase 1	5,314	5,330	5,512	+3.4	NPA	NPA	NPA	–	3,370	3,572	3,780	+5.8
Phase 2	5,125	5,333	5,438	+1.9	205	288	289	+0.3	3,572	3,686	4,258	+15.5
Banfield Freeway												
Phase 1	3,713	3,845	3,822	-0.6	37	180	46	-74.4	340	570	490	-14.0
Phase 2	3,161	3,793	3,662	-3.5	708	1,180	846	-28.3	628	657	685	+4.3
I-95, Miami												
Phase 1	6,145	6,416	6,864	+7.0	170	309	259	-16.2	274	314	325	+3.5
Phase 2	5,170	5,880[d]	6,347	+7.9	1,555	1,603	1,500[e]	–	314	352	335	-4.8
Southeast Expressway												
1977	5,504	4,306	3,909	-9.2	388	641	737	+15.0	2,000	2,124	2,124	0.0
1971	4,554	4,201	4,384	+4.4	NPA	NPA	NPA	–	2,152	2,454	3,170	+29.2
I-495, Lincoln Tunnel	2,324	3,227	3,253	+0.8	NPA	NPA	NPA	–	21,868	26,092	26,561	+1.8

Note: NPA = no priority automobiles included in HOV treatment, A = actual, P = predicted. Data are from Charles River Associates.

[a]Average error = -0.9 percent, SD = 4.7 percent.
[b]Average error = -13.9 percent, SD = 31.3 percent.
[c]Average error = 3.7 percent, SD = 10.6 percent.
[d]High violation rate.
[e]Model predicts saturated HOV-lane conditions.

for HOV lanes on surface arterials or HOV lanes on circumferential freeways). Second, the HOV lanes ranged from 2.5 to 9 miles in length. Third, all sites experienced force-flow or severe capacity constraint conditions on the general-purpose lanes in the before period during the morning peak hour. It appears, however, that the benefits would be slight (or even negative) if an HOV lane were instituted in a corridor that operated under relatively free-flow conditions during the morning peak hour.

Finally, among the HOV sites used in model estimation, many network conditions and alternative links (e.g., parallel freeways or arterials) exist, allowing different route diversion effects. The models and relationships that were developed reflect the average of these conditions. If a corridor being analyzed is especially atypical with respect to alternative routes, the models may not capture the full effects due to these alternative or competing links.

CONCLUSIONS

The development of a set of models for use in forecasting travel volumes resulting from implementing HOV strategies on freeways has been described. The models have been incorporated in worksheets that are described elsewhere (5).

To examine how well the models forecast, the procedures developed were used with known level-of-service changes to predict after peak-hour volumes of nonpriority automobiles, priority automobiles, and HOV bus passengers for each HOV site. More realistic tests were then made by using information only from the before period for each HOV facility.

Using known level-of-service changes, average relative errors were quite small for the nonpriority automobile (-0.03 percent) and HOV bus modes (+2.4 percent). The average error for the priority automobile mode was higher, although still quite acceptable (-7.7 percent). Using only the before data, average errors across the HOV sites increased slightly for the nonpriority automobile (+0.9 percent) and HOV bus (+3.7 percent) modes and by a somewhat larger amount for the priority automobile mode (-13.9 percent). In summary, these results demonstrate that the HOV models and procedures developed provide an effective way to estimate travel volumes that may result from implementing alternative HOV strategies on major freeways leading into the central area of cities.

ACKNOWLEDGMENT

The work described in this paper was performed by Charles River Associates under contract to FHWA. The author wishes to thank Howard Bissell, who, as contract manager, provided technical guidance during the course of the project.

The computer work was performed by Robert Hirschey, with econometric assistance provided by Lawrence Kolbe. Joan Solomon was involved in the initial data collection tasks. Useful supervision and direction were provided by Daniel Brand, with earlier assistance provided by William Tye and Fred Dunbar. Adolf D. May of the University of California at Berkeley contributed important input at various points throughout the course of this study. Finally, the author would like to thank the many individuals and agencies who cooperated with his requests for information on HOV evaluations.

Discussion

Olga J. Pendleton*

The objective of this discussion is to point out certain potential problems in least-squares regression analysis that could affect conclusions and resulting models. Whereas the author may have been aware of these problems and taken them into con-

*Texas Transportation Institute, Texas A&M University System, College Station, Texas.

sideration, I feel it would be constructive to discuss this issue because it may be instructive for others engaged in similar model building.

In applying least-squares regression techniques to a set of data, it is important that the necessary distributional assumptions on the model errors be validated; these assumptions are that the dependent variable in the model, for example, change in traffic volume and be normally distributed and that the errors be independent and have homogeneous variances. This frequently is not the case when the random variable is a ratio of count data. In this instance, the error variance often increases as the ratio increases. Transformations of the variable such as the log transformation can often be used to accommodate the nonhomogeneous error problem. In this study the author uses weighted least squares with the weights being the square root of the sum of the total traffic volume. Weighted regression is another means of achieving variance homogeneity because the weights will be inversely proportional to the variance of the dependent variable and hence stabilize the variance. The weight used in this study, the square root of the traffic volume, thus implies that the variance of the change in traffic volume is inversely proportional to traffic volume. It is not obvious that as the traffic volume increases, the variance in the change in traffic volume decreases. It would be of interest to know why the author believed that this relationship exists.

Given the impact of the distributional assumptions on the model development and subsequent conclusions, the residuals from such a model (errors) should be examined to verify that they do indeed satisfy these assumptions. Examination of the cumulative distribution of these residuals and plots of the residuals versus model variables is an effective means of doing this. In fact, the examination of residuals by using unweighted ordinary least squares can sometimes lead to the selection of the appropriate weighting factor, which may be the method the author used to determine the weight in this study. This type of information would be extremely useful and enlightening to the reader.

There is a potential problem in the definition of some of the independent variables in this study, which could affect the results. Because this method of defining variables is fairly common, I would like to take this opportunity to explain the potential consequences of this practice by using two of the variables from this study in a hypothetical example. Consider the definition of the variable percentage change in travel time for two-person priority automobiles (PA2-TT). This variable was assigned the value zero for sites where two-person priority automobiles were not allowed onto the HOV lanes and took on the percentage values for all other sites.

Figure 2 might represent the relationship between PA2-TT and percentage change in traffic volume. The slope of the line using this method of defining PA2-TT would be strongly positive due to the number of low-volume change sites that did not allow two-person priority automobiles. If separate lines were fit to the sites that allowed two-person priority automobiles, the slope would be lower. One can easily construct other situations where including the zero PA2-TT sites in the analysis in this way could affect the resulting model coefficients. A method that would circumvent this problem would be to include an indicator variable that assigned the sites that allowed two-person priority automobiles a separate slope and intercept from those sites that did not. This indicator variable would take on the value 1 or 0 and its coefficient would correspond to the intercept of a line fit to those sites allowing those vehicle types only. Correspondingly, the coefficient for PA2-TT would take on a new interpretation, for example, the slope of a line for those sites only. Furthermore, the difference in average percentage volume change for the two groups of sites would be reflected in the test of significance of the coefficient of the indicator variable.

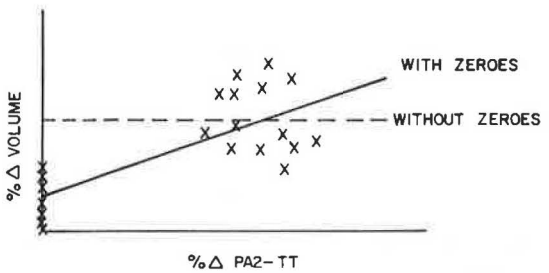

FIGURE 2 Relationship between PA2-TT and percentage change in traffic volume (Pendleton).

In reporting results from a regression model, information such as sample size and mean squared error (MSE) are as critical in model evaluation as R^2 or the F- and t-statistics. The latter statistics are difficult to interpret if the sample size is not reported. Whereas R^2 represents the percentage of the total variability in the data accounted for by the model, it can be unrealistically inflated for low sample sizes relative to the number of model parameters being fit. It also does not reflect the predictive ability of the model. The statistic that does this is the MSE, which is used in the computation of prediction and confidence intervals. If these prediction intervals had been reported in Table 6, they would have been useful in evaluating the predictive potential of the model. That is, if the predicted after volume for Shirley Highway is 5,105 ± 2,000, this would not be a good predictive model; however, if the model predicted within an interval of ±2, it would be extremely good. The author notes the danger in predicting the after volume by using the actual before volumes in the prediction. Why not report the predicted value of the expected percentage of change in total volume, that is, the dependent variable used to build the model, as well?

Table 5 gives results for several models, two of which do not appear to yield significant overall F-ratios. These values may be significant, but the reader is unable to verify this without knowing the sample sizes. An R^2 of 0.28 is surely too small to suggest any validity in the interpretations of these model parameters. Attempting to interpret signs and magnitudes of coefficients in such a model is like trying to make sense of random fluctuations about zero. If the overall F-ratio is not significant, this means that none of the model coefficients, tested simultaneously, are different from zero. To subsequently interpret the less sensitive t-statistics that test individually that each coefficient is different from zero corresponds to the error of interpreting some multiple range tests in a one-way analysis of variance when the overall F-ratio is not significant. As a point of clarification, in Table 5 why is the constant term omitted in models A and B? Omission of the constant term can have serious effects on the model fit and interpretation; thus it is important that there be evidence to justify its omission.

In conclusion, the problems addressed in this discussion could affect the conclusions based on least-squares regression analysis models. Because

many of the points raised in this study are frequently overlooked by researchers using this method, the objective of this discussion was to enlighten researchers about these potential problems. The author of this study may well have taken these points into consideration in this analysis. However, because many of these points are frequently ignored by the researcher, it would be beneficial if studies such as this one addressed these points in the literature.

Author's Closure

The discussion by Pendleton highlights various generic issues that one should be aware of in using least-squares regression analysis in model-building exercises. I share her concern that these issues may not always be properly examined by those using available regression packages, and I believe that her discussion and this response can provide useful information to others.

WEIGHTED LEAST SQUARES

As described by Pendleton, weighted least squares is used to correct for heteroscedasticity in order that efficient parameter estimates and unbiased estimates of the variances are obtained. Typically, prior information is not available to obtain estimates of the variances of the individual error terms in order to determine the appropriate weights. However, if it is suspected that the error variances are proportional to the size of an independent variable, for example, then a set of weights can be constructed. In this particular instance, the objective was to develop models to forecast after modal volumes as a function of before vehicle volumes and changes in levels of service (principally travel times for the own and competing modes). We also suspected that the variances in peak-hour volumes would decrease as travel volumes increased. In this instance the appropriate procedure was to weight the dependent and independent variables by the square root of the volumes. Subsequently, the dependent variable was redefined to be the percentage change in modal volume, but the weights were left unchanged. In effect, this presumes that the variances in the change in travel volume decrease as volume increases. To check this assumption, the priority automobile model in Table 4 was reestimated, first assuming that variances are inversely proportional to what was originally used and second that variances are constant, such that weighting is not necessary.

The results of the initial and two additional cases described earlier were nearly identical. The mean square errors were 0.231, 0.231, and 0.226, respectively. Because weighting is a procedure used to obtain better estimates of the parameter variances, the estimated values of the model coefficients were nearly identical for the three alternatives. We thus conclude that it is probably not necessary to use weights when ratios of travel volumes are used as the dependent variable but doing so did little to affect the model parameters and statistics.

DEFINITION OF VARIABLES

Pendleton discusses how alternative procedures for describing a variable (in this instance, PA2-TT) may affect the estimated value of the parameter. In general, it is likely that different variable definitions will produce different model results; the same is also true for different functional forms (e.g., linear versus log transformations). With respect to the latter point, alternative functional forms were evaluated and, as reported in the paper, the model form as described by Equation 1 was selected. With regard to the variable PA2-TT (the percentage change in travel time when two-person vehicles were allowed to use the HOV facility), different definitions were also examined, including ones essentially similar to those described by Pendleton.

Basically, given the limited sample size (because of the relatively low number of localities in the United States where two-person carpools were allowed onto a freeway HOV facility and adequate volume and level-of-service data were collected), we could not detect a consistent pattern between changes in travel times for vehicles on the nonpriority lanes (NPA-TT) and percentage change in priority automobile volumes. Consequently, if the variable PA2-TT was assigned the same value as NPA-TT when buses or carpools of three or more were allowed onto the HOV facility but took on the actual percentage change in travel time when carpools of two or more were allowed onto the HOV lanes, the significance of the parameter was greatly diminished because there are more of the former sites. Thus, the variables PA2-TT and PA3/4-TT (one or the other but not both) contribute information to the right-hand side of the equation if they have something of significance to explain (in that either two- or three-or-four-or-more-person carpools save time because of the introduction of the HOV strategy).

Although Figure 2 graphically demonstrates how one might view the consequences of restricting certain values to zero, it cannot be used in this instance to infer how the slope and the intercept may change. The reason is that in some instances, the percentage change in volume of two-person carpools may be negative if carpools of three or more persons are allowed on the HOV facility, or vice versa. This would be graphically depicted as shown in Figure 3.

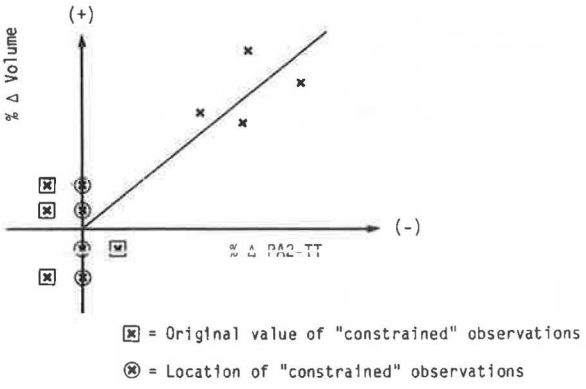

FIGURE 3 Relationship between PA2-TT and percentage change in traffic volume (Parody).

REGRESSION STATISTICS

I agree with Pendleton's remark with respect to reporting a full range of statistics from the model estimation process. It is a shortcoming I too have lamented in the work of others.

When weights were used in generalized least squares, the MSEs reported in the estimation package being used described characteristics of the newly transformed data and therefore did not provide correct information for the original data. Thus, they were not reported in the first draft of the paper. However, the MSEs for the unweighted data have now been calculated and, along with the degrees of freedom (DFE) (which equal the sample size less the number of variables in the model), are included in the tables. A note on their interpretation and application might be useful, however. The MSE values apply to the dependent variable (i.e., percentage change in a particular modal volume). A certain amount of mathematics is required to transform that statistic to the ultimate variable of interest, modal volumes after the introduction of the HOV facility.

The standard error of the estimate [i.e., the standard deviation ($\bar{\sigma}$) of the variable on the left-hand side] is equal to the square root of the MSE. Therefore, for the Shirley Highway example mentioned by Pendleton, the estimated percentage of change in volume for nonpriority automobiles ($\hat{PVNPA}$) can be calculated (using Tables 1 and 3) as follows:

$$\begin{aligned}\hat{PVNPA} = {} & -0.916 - 1.053[(58.3/56.2) - 1] \\ & + 1.190\ [(58.3/56.2) - 1] \\ & + 0.122\ [(38.3/56.2) - 1] \\ & + 0.278\ [(38.3/37.5) - 1] \\ & + 0.949\ \{(3/3) \cdot [(4,896 \\ & + 195 + 0)/4,896]\}\end{aligned}$$

$$\hat{PVNPA} = 0.043 \qquad (4)$$

Given that the formula for computing the after volume of nonpriority automobiles ($VNPA_1$) is

$$VNPA_1 = VNPA_0(1 + PVNPA) \qquad (5)$$

then

$$VNPA_1 = 4,896(1 + 0.043) = 5,105 \text{ vehicles per hour} \qquad (6)$$

The confidence interval for the forecast volume of 5,105 vehicles per hour on the nonpriority lanes of Shirley Highway is equal to the following expression:

$$\hat{VNPA}_1 = VNPA_0[1 + (PVNPA \pm t_{\alpha/2}\sqrt{h_i S^2})] \qquad (7)$$

where

S^2 = MSE = 0.0007,
h_i = the leverage = $x_i(X'X)^{-1}x_i'$, and
$t_{\alpha/2}$ = 2.45 for 95 percent confidence interval and 6 DOF.

Because the value of h is not readily available, it is set equal to 1 (which equals its value when intervals are constructed around the mean of the independent variables) such that the confidence interval of the forecast for Shirley Highway is approximately as follows:

$$\hat{VNPA}_1 = 5,105 \pm 317 \text{ vehicles per hour} \qquad (8)$$

Pendleton suggests reporting the predicted values of the percentage change in travel volumes. Although such a list would be useful in evaluating alternative models, it could tend to obscure what is happening on the highway. For example, the actual percentage increase in three-person automobiles for the San Bernardino, phase 2, case is 92 percent, whereas the priority automobile model forecasts an increase of 154 percent. Expressed in terms of vehicle volume, this difference amounts to only 185 vehicles per hour (see Table 6). The reader, however, can easily calculate both the actual and forecasted percentages from Tables 6 and 7.

Pendleton is correct in stating that one or two of the priority bus models are not highly significant. As described in the text, likely reasons relate both to the unavailability of information on fares, frequencies, availability of park-and-ride facilities, and route coverage and to the desire to produce a sketch-planning procedure using available data that can provide forecasts from beginning to end in one day or less. However, it is interesting to note that the after bus passenger volumes predicted by the models in Table 5 differed from actual volumes by an average absolute amount of 13.4 percent (see Table 6).

In Table 5 both models A and B were originally estimated with a constant term. In the case of model A, the inclusion of a constant reduced what little significance there was for the variable PTB to zero. Thus, the choice was between a model that would produce the same change in bus volumes irrespective of the change in bus level of service (i.e., even if there was no change in bus travel time), and one that was at least partially sensitive to alternative HOV configurations and local site conditions. Given that we expect the intercept to be basically zero when PTB is zero, the latter model was chosen. (With only one variable, tests of the F- and t-statistics produce the same results, which Pendleton could not ascertain without knowing whether or not a constant had been used.)

In the case of model B, the constant was estimated to have a value of zero. Because of the high explanatory power exhibited by the supply variable (percentage change in number of buses), likely because of its being highly correlated with bus ridership, there was not much left to be explained by a constant term.

In summary, the models presented in the paper quantitatively relate changes in volumes to changes in service levels for a dozen HOV facilities across the United States. The models were formulated to be consistent with the task of using actual before-and-after data to develop a set of forecasting procedures that could be used to examine changes in modal volumes for alternative HOV strategies in a quick-response time frame. As more data for these types of TSM options become available, the models may be refined. Revised model coefficients can easily be substituted in the worksheets presented in the User's Guide (5), and sensitivity tests can be performed.

I hope that the observations made by Pendleton relating to model estimation and the subsequent discussion on how they were addressed in the study are illuminating and will assist others when conducting and reporting on least-squares regression analysis models.

REFERENCES

1. Urban Corridor Demonstration Program. U.S. Department of Transportation, Oct. 1974.
2. P. Benjamin et al. Service and Methods Demonstration: Annual Report. UMTA, U.S. Department of Transportation, Nov. 1975.
3. Charles River Associates, Inc. Predicting Travel Volumes for HOV Priority Techniques: Technical Report. FHWA, U.S. Department of Transportation, April 1982. NTIS: PB82-231812.
4. Urban Transportation Planning: General Information. FHWA, U.S. Department of Transportation, March 1972.
5. T.E. Parody. Predicting Travel Volumes for HOV Priority Techniques: User's Guide. FHWA, U.S. Department of Transportation, April 1982. NTIS: PB82-247347.

The opinions and conclusions expressed in this paper are those of the author and do not necessarily reflect the views or policy of FHWA.