1120

TRANSPORTATION RESEARCH RECORD

Transportation Logistics

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL
WASHINGTON, D.C. 1987

Transportation Research Record 1120
Price: $10.50
Editor: Naomi Kassabian
Typesetter: Lucinda Reeder
Layout: Marion L. Ross

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.

Transportation logistics.

p. cm.—(Transportation research record, ISSN 0361–1981 ; 1120)
ISBN 0-309-04473-1
1. Transportation—Mathematical models—Congresses. I. National Research Council (U.S.). Transportation Research Board.
II. Series.
TE7.H5 no. 1120
[HE147.7]
380.5 s—dc19
[380.5'0724] 87–34803
CIP

Transportation Research Record 1120

The **Transportation Research Record** series consists of collections of papers on a given subject. Most of the papers in a Transportation Research Record were originally prepared for presentation at a TRB Annual Meeting. All papers (both Annual Meeting papers and those submitted solely for publication) have been reviewed and accepted for publication by TRB's peer review process according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The views expressed in these papers are those of the authors and do not necessarily reflect those of the sponsoring committee, the Transportation Research Board, the National Research Council, or the sponsors of TRB activities.

Transportation Research Records are issued irregularly; approximately 50 are released each year. Each is classified according to the modes and subject areas dealt with in the individual papers it contains. TRB publications are available on direct order from TRB, or they may 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.

Contents

Foreword

The papers contained in this Record represent the first publications sponsored by the Task Force on Transportation Supply Analysis. Lamar and Sheffi focus on the line-haul (intercity) operations of less-than-truckload (LTL) carriers rather than on local pickup-and-delivery operations. The optimal design of LTL motor carrier networks is cast as a fixed-charge network design problem. A recently developed lower bound is applied iteratively with a link inclusion heuristic in an implicit enumeration framework. Initial computational results for solving this class of problem are reported. The emphasis of the next paper, by Eckstein and Sheffi, is on group line-haul operations involving the daily movement of trailers between a central breakbulk terminal and a set of end-of-line satellite terminals. When each tractor can pull two trailers, there are many possibilities for creating tractor tours that accomplish the required pickup, delivery, and empty-balancing operations. The challenge is to create optimal tours that minimize transportation costs.

Increasing concern about accommodating rising truck traffic and associated size and weight trends led Mahmassani et al. to develop a general mechanism for the network representation of improvements in the highway system consisting not only of physical capacity expansion, but also corresponding operational strategies in the form of existing or new lane-access restrictions for either vehicle class. This mechanism allows the consideration, as a special case, of exclusive truck lanes or facilities contemplated by several agencies. The special requirements of the traffic assignment procedure in this context, including the need to explicitly consider the asymmetric interaction between cars and trucks, give rise to potentially serious methodological difficulties, which are addressed for specific types of applications. Kim and Mahmassani present the results of the empirical development and calibration of performance that capture the dependence of travel time on the respective volumes of passenger cars and trucks sharing the physical right-of-way on urban freeway sections.

Kuah and Perl present a network optimization methodology for the design of an integrated bus–rail rapid transit system. The Feeder-Bus Network Design Problem (FBNDP) is defined as that of designing a set of feeder-bus routes and determining the frequency on each route so as to minimize operation and user costs. An heuristic method is presented that generalizes the savings approach to consider operating frequency. The analysis presented illustrates the capabilities of the proposed model as a strategic planning tool for feeder-bus network design.

In two papers, Safwat presents the applicability of a Simultaneous Transportation Equilibrium Model (STEM) methodology from the computational as well as the behavioral points of view. In the first paper, the STEM approach is applied to the Egyptian intercity transportation system. STEM methodology is summarized as well as major issues related to intercity passenger travel in Egypt. In the second paper, a description is given of the computational experience with a globally convergent algorithm (SPND) that predicts trip generation, trip distribution, modal split, and traffic assignment simultaneously on a STEM applied to analyze intercity passenger travel in Egypt.

An Implicit Enumeration Method for LTL Network Design

BRUCE W. LAMAR AND YOSEF SHEFFI

The optimal design of less-than-truckload (LTL) motor carrier networks is cast as a fixed-charge network design problem. A recently developed lower bound is applied iteratively with a link inclusion heuristic in an implicit enumeration framework. Initial computational results for solving this class of problems are reported.

Less-than-truckload (LTL) motor carriers haul freight from many origins to many destinations. In this paper the emphasis is on the line-haul (i.e., intercity) operations of an LTL carrier rather than on the local or pickup-and-delivery operations. [Line-haul operations typically account for 40 percent of the operating expenses for an LTL motor carrier *(1)*.] The line-haul network design for LTL operations is frequently referred to as a load plan. In such a plan, shipments are considered to originate and terminate at the carrier's end-of-line (EOL) (i.e., city) terminals, and an origin-destination (OD) pair of EOL terminals is referred to as a market. Associated with each market is a shipment demand specifying the shipments that the carrier must haul from the origin to the destination. [Shipment demands are expressed in flow units such as hundredweight (CWT) per day or cubic feet per week.] Because the shipment demand for an individual market is typically much smaller than the capacity of a trailer (40,000 lb or 6,250 ft^3), it is economical to combine the demand for several markets at the origin EOL terminal and send this consolidated demand to an intermediate terminal, referred to as a breakbulk (BB) terminal. At the BB terminal, consolidated shipments on incoming trucks are unloaded, sorted, and reloaded onto outgoing trucks bound either for the final destination or another BB terminal.

In designing a line-haul network, a motor carrier must make the following two interrelated sets of decisions: (a) which links should be used and (b) how the shipment demand should be routed over the selected links. The motor carrier's criterion for making these decisions is to minimize its total transportation and handling costs while providing a high level of service. The handling costs, which occur at the BB and EOL terminals, are assumed to be proportional to the volume of freight sent through these terminals. Thus, each terminal in the carrier's line-haul network can be represented by a pair of nodes connected by a single arc, and the costs of these special arcs are simply the handling costs of the respective BB or EOL terminals.

B. W. Lamar, Graduate School of Management, University of California, Irvine, Calif. 92717. Y. Sheffi, Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

The representation of the transportation costs for an LTL motor carrier is somewhat more involved. The link cost function for a typical LTL line-haul link is shown in Figure 1. The daily link cost depends on the average daily volume of freight (i.e., flow) carried on that link. If the flow is zero, the link is not used and so naturally the cost is zero. If the flow is small, the carrier will send a prespecified number of trucks per day over that link even though these trucks will be only partly full. This is done in order to maintain a satisfactory level of service *(2)*. In Figure 1, this minimum frequency is set at two truckloads per day. If the link flow is below this minimum frequency, the carrier's transportation cost (fuel, maintenance, etc.) for this link is essentially set at a fixed value. On the other hand, if the link flow is above the minimum frequency, the carrier will implement a "go when full" truck-dispatching policy and the link cost will be proportional to the number of truckloads of freight carried over that link.

The highly nonlinear nature of the transportation costs experienced by an LTL motor carrier makes network design problems difficult to evaluate analytically. For this reason, the transportation cost function shown in Figure 1 is approximated by the fixed-charge cost function shown in Figure 2. This approximation captures both the cost for the minimum service requirement inherent in low flow levels and the cost proportional to freight volume inherent in high flow levels. As with any approximation, however, this simplified cost structure strikes a balance between a realistic representation of the problem and a mathematical form that is analytically tractable.

The mathematical model described in this paper provides two tools to aid an LTL motor carrier in designing and operating a line-haul network. First, the model generates a series of heuristic solutions that the carrier can implement to route and consolidate shipment demand. Second, the model provides a lower bound to the minimum total transportation and handling costs for the carrier's network. This lower bound is used to determine whether a heuristic solution is "close enough" to optimal. (It can also be used to evaluate the near-optimality of heuristics generated by other procedures.) The network design model can be solved whenever the carrier is updating the line-haul network. Depending on the volatility of the shipment demand in the OD markets, this updating could vary from a weekly to a quarterly basis. Moreover, the model can answer strategic "what if" questions concerning cities that the carrier is not currently serving and locations of new BB terminals. For each terminal configuration the model can suggest efficient strategies for including these markets in the carrier's network.

The paper is organized as follows: (a) the model formulation and its relaxation, (b) an implicit enumeration solution procedure, and (c) some computational results.

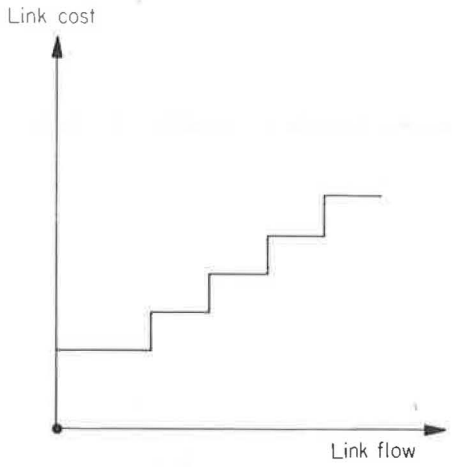

FIGURE 1 Typical link cost function.

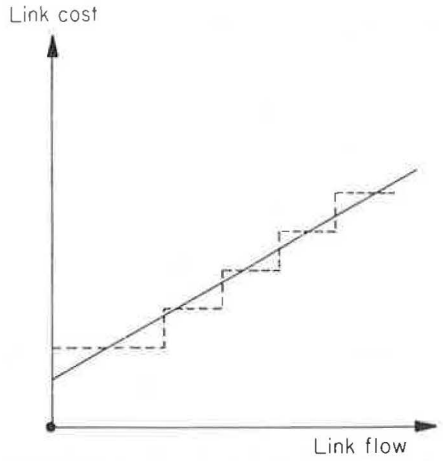

FIGURE 2 Approximation of LTL link cost function.

PROBLEM FORMULATION

This section is in two parts. First the fixed-charge network design problem addressed here as an integer program (IP) is formulated, and second, an LP relaxation of this formulation is discussed.

IP Formulation

The following notation is used to define the problem:

N = set of nodes with generic element n,
A = set of directed arcs with generic element a, and
M = set of OD pairs ("markets") with generic element m.

The shipment demand is characterized by a market flow vector, $\mathbf{q} = (. . ., q_m, . . .)$. The units of this demand are, for example, hundredweight per week or cubic feet per day.

All arcs in this problem are design elements. Each arc (a) is associated with a fixed charge (f_a) and a variable cost (c_a). In the application discussed here, f_a stems from the minimum service requirement, whereas c_a represents the (average) marginal cost of hauling freight over arc a (see Figure 3).

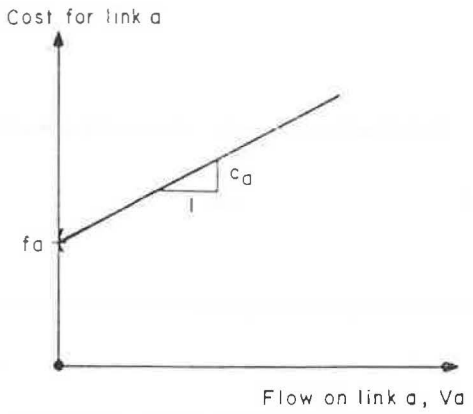

FIGURE 3 Fixed-charge link cost function.

To characterize the decision variables here, let $y_a = 1$ if link a is in the network, and $y_a = 0$ otherwise. Also let $x_{a,m} = 1$ if link a carries flow q_m (i.e., if it is used to carry flow between the origin and the destination associated with the mth OD pair), and $x_{a,m} = 0$ otherwise. Thus $\mathbf{y} = (. . ., y_a, . . .)$ is the network design vector, and $\mathbf{x} = (. . ., x_{a,m}, . . .)$ is the flow-routing vector.

The vector of link flows $\mathbf{v} = (. . ., v_a, . . .)$ can be expressed as

$$v_a = \sum_m q_m x_{a,m} \qquad \forall\, a \tag{1}$$

(When no other indication is given, the notations Σ and $\forall$ refer to all members of the relevant set. Thus in Equation 1, Σ_m is equivalent to $\Sigma_{m \in M}$, and $\forall\ a$ is equivalent to $\forall\ a \in A$.) A parameter used later in the program formulation is the capacity of each arc, denoted by the vector $\mathbf{u} = (. . ., u_a, . . .)$. Because the problem discussed here is not naturally capacitated, u_a can be defined as the largest possible value that v_a can attain. Depending on the network structure, this maximum may be equal to or less than the combined shipping demand for the entire network (i.e., $\Sigma_m q_m$).

To formulate the flow conservation constraints, let $I(a)$ and $J(a)$ denote the "tail" node and "head" node, respectively, of arc a. Similarly, let $O(m)$ and $D(m)$ define the origin and destination nodes, respectively, for market m. In addition, let A_n be the set of arcs whose tail node is n and let B_n be the set of arcs whose head node is n.

The multicommodity network flow balance constraints can now be written as

$$\sum_{a \in A_n} x_{a,m} - \sum_{a \in B_n} x_{a,m} = \left\{ \begin{array}{ll} 1 & \text{if } n = O(m) \\ -1 & \text{if } n = D(m) \\ 0 & \text{for all other } n \in N \end{array} \right\} \quad \forall\ n,m \tag{2a}$$

$$= x_{a,m} \geq 0 \qquad \forall\ a,m \tag{2b}$$

To avoid repeating this set of constraints throughout the paper, let X define the set of routing vectors $\mathbf{x} = (. . ., x_{a,m}, . . .)$ that comply with Equations 2.

The fixed-charge network design problem (denoted IP) can now be stated as the following integer program:

Program IP:

$$\min_{x \in X} \sum_m \sum_a c_a q_m x_{a,m} + \sum_a f_a y_a \tag{3a}$$

subject to

$$\sum_m q_m x_{a,m} \leq u_a y_a \qquad \forall\, a \tag{3b}$$

$$y_a \in \{0, 1\} \qquad \forall\, a \tag{3c}$$

Let (**x***, **y***) denote the optimal solution vectors, let **v*** denote the optimal link flow vector, and let z^*[IP] denote the optimal objective function value for program IP. The objective function 3a in this program minimizes the total system costs, including both variable and fixed charges, over all the links in the network. Carrying the minimization over X guarantees that the feasible solutions are restricted to directed paths over the network. Constraints 3b guarantee that a link carries flow only if it is in the network, and constraints 3c ensure the integrality of the decision variables. (Even without a specific constraint, it will always be the case that $x_{a,m} \in \{0, 1\}\ \forall\ a,m$ because the problem is uncapacitated.) Note that alternatively constraints 3b can be written in a disaggregate form:

$$x_{a,m} \leq y_a \qquad \forall\ a,m \tag{4}$$

Although this form generates a tighter LP relaxation than the LP relaxation of IP, the advantage of using constraints 3b is that the LP relaxation is then particularly easy to solve. This ease is exploited in the procedure that calculates the lower bound for z^*[IP], as shown by Lamar et al. *(3)*. For additional discussion of aggregate and disaggregate formulations, see work by Wong *(4)*, Rardin and Choe *(5)*, and Balakrishnan *(6)*. A general discussion of fixed-charge network design problems has been given by Wong *(4)* and Magnanti and Wong *(7)*.

LP Relaxation

The linear programming relaxation of IP is formed by replacing constraints 3c with the nonnegativity constraints

$$y_a \geq 0 \qquad \forall\, a \tag{5}$$

Let LP denote this relaxation of IP. Let **x***, **v***, and **z***[LP] denote the optimal values of the routing decision variables, the arc flows, and the objective function, respectively, in program LP. Note that constraints 5 can be omitted from LP because the nonnegativity of $\{y_a\}$ is ensured by constraints 3b and the nonnegativity of **x** (which is required in the set X).

The LP relaxation is then given by

Program LP:

$$\min_{x \in X} \sum_m \sum_a q_m (c_a + f_a/u_a) x_{a,m} \tag{6}$$

The equivalence between program LP as formulated in Equation 6 and its original formulation (Equations 2, 3a, 3b, and 5) can be shown on the basis of Balinski's observation *(8)* that constraints 3b will always be satisfied with equality in the optimal solution of LP. Because $u_a > 0\ \forall\ a$, constraints 3b can be solved explicitly for y_a and substituted in the objective function 3a to obtain the formulation of LP given in Equation 6. As mentioned in connection with Equation 4, program LP is easy to solve. The reason for this is that it decomposes by origin (and by OD pair) into a set of independent shortest-path problems (with flow assignment). Thus this program can be solved by a many-to-many algorithm [see, e.g., the paper by Floyd *(9)*] or by repeated application of a one-to-many algorithm [see, e.g., the discussion by Moore *(10)*]. These algorithms, coupled with recent list-processing techniques [see, e.g., the paper by Glover et al. *(11)*], are very efficient. Consequently, LP can be solved repeatedly in the course of finding both a lower bound and an heuristic solution of the integer program, IP.

The solution procedure for IP is described next.

SOLUTION PROCEDURE

The implicit enumeration (IE) procedure presented in this section determines an ε-optimal solution to the integer program, IP. That is, the algorithm determines a heuristic (i.e., feasible) solution to IP whose objective function value is always within $100 \cdot \varepsilon$ percent of z^*[IP] (or determines that IP is infeasible). The value of ε is assumed to be specified a priori.

A distinctive difference between the standard implicit enumeration method [see, e.g., work by Geoffrion and Marsten *(12)*] and the one presented in this section is that here the lower bound for each subproblem is not obtained directly from the LP relaxation of the integer problem. Instead, the lower bound is based on a "capacity improvement" procedure (described in the following discussion). This results in tighter lower bounds and a reduction in the computational effort required for the IE algorithm.

In the following procedure the basic "divide and conquer" strategy is used, which can be represented as a binary enumeration tree. Each point in the tree represents a subproblem of IP in which the binary decision variables, $\{y_a\}$, are partitioned into "out" and "in" subsets. The IE procedure determines a lower bound (to the optimal objective function value) and a heuristic solution for each feasible subproblem that it evaluates. The best (i.e., lowest-cost) heuristic solution (from among the subproblems that have been evaluated) is retained as the incumbent solution to IP. The process continues until the lower bound of all subproblems is within $100 \cdot \varepsilon$ percent of the objective function value of the incumbent.

The point in the binary enumeration tree being evaluated at any given time is referred to as the "current" point. This current point can be specified by a constraint set Y whose elements are of the form $\{y_a = 0\}$ or $\{y_a = 1\}$. The current integer subproblem, denoted IP_Y, is formed by adding the constraints in Y to IP; and the current linear subproblem, denoted LP_Y, is formed by adding the constraints in Y to problem LP. (Observe that LP_Y can be expressed as a shortest-path problem simply by modifying the objective function coefficients in LP rather than by adding constraints to this problem.) In addition, the IE procedure described here evaluates a

family of linear subproblems formed by parametrically altering an arc capacity vector, $\mathbf{u}_Y$. Let $LP_Y(\mathbf{u}_Y)$ denote this family of current linear subproblems.

Figure 4 is a flowchart of the basic steps in the IE algorithm. Step 0 initializes the algorithm by placing program IP in the candidate list. This list identifies the subproblems that are to be evaluated in the IE procedure. Along with the specification of IP, the capacity parameter vector $\mathbf{u}$ is also stored in the candidate list. [This vector is used later in the IE algorithm as the initial vector in the CI procedure (see discussion of Step 8).]

If a feasible solution to IP is known a priori, then in Step 0 this solution is used as the current incumbent solution to IP. Let h denote the objective function value of the current incumbent and let t denote a target value used in the CI algorithm. This target value is set at

$$t = \frac{h}{1 + \varepsilon} \tag{7}$$

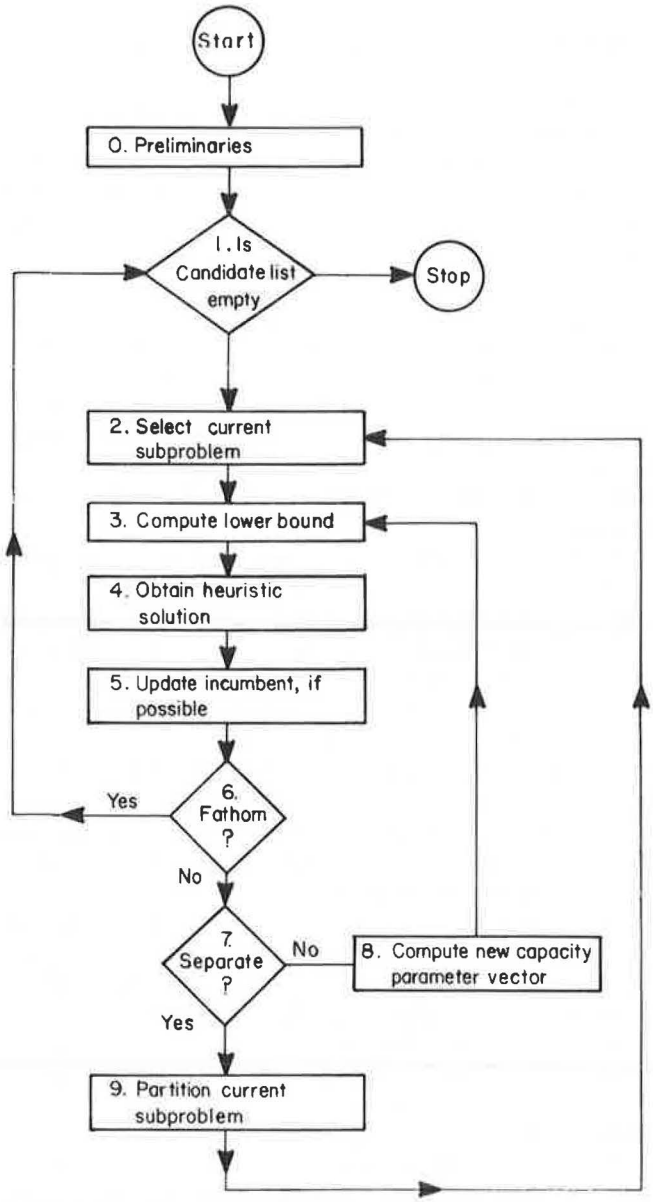

FIGURE 4 Flowchart of implicit enumeration algorithm.

where ε is the prespecified optimality measure. (If no incumbent feasible solution to IP is available, then h and t are set to ∞.)

In Steps 1 and 2, the subproblems in the candidate list are reviewed. If this list is empty, the IE algorithm terminates and the current incumbent is an ε-optimal solution to IP. (If the list is empty and there is no incumbent, then IP is infeasible.) If the candidate list is nonempty, a subproblem is selected on a last-in–first-out (LIFO) basis. Note that a LIFO selection rule generates a depth-first search in the IE enumeration tree, which in turn minimizes the number of subproblems contained in the candidate list at any one time. This is important because the candidate list stores an arc-length capacity vector along with each subproblem specification. Thus, the LIFO subproblem selection rule is used because it minimizes computer memory requirements associated with the candidate list.

In Step 3 a lower bound to $z^*[IP_Y]$ is computed. Here, the current linear subproblem is solved and l_Y, the lower bound to $z^*[IP_Y]$, is computed as follows:

$$l_Y = \min \{t, z^*[LP_Y(\mathbf{u}_Y)]\} \tag{8}$$

This lower bound is guaranteed to be at least as tight as $z^*[LP_Y]$ and computational experience has shown that l_Y can be considerably tighter than $z^*[LP_Y]$ [see work by Lamar et al. *(3)*]. Moreover, the lower bound given in Equation 8 is easy to compute. It involves only the solution of $LP_Y(\mathbf{u}_Y)$ (i.e., a set of shortest-path problems with flow assignment) and a comparison of $z^*[LP_Y(\mathbf{u}_Y)]$ with the target value t.

In Step 4 a new incumbent solution to IP is sought by obtaining a heuristic solution to the current integer subproblem, IP_Y. The rationale of this heuristic procedure is that because $LP_Y(\mathbf{u}_Y)$ is easy to solve, this current linear subproblem can be used repeatedly in determining a feasible solution to the current integer subproblem, IP_Y.

At each iteration of the heuristic procedure, $LP_Y(\mathbf{u}_Y)$ is solved and the links with large flow are selected to be part of the candidate network design of IP_Y. For the next iteration, the fixed charge on these links is assumed to be a sunk investment, and only the variable cost is used in solving $LP_Y(\mathbf{u}_Y)$ again. This iterative process of solving $LP_Y(\mathbf{u}_Y)$ and selecting additional arcs for inclusion in the heuristic network design is repeated until a feasible solution to IP_Y is reached. At this point, the arcs included in the heuristic network design form a path for each OD pair in the market set M. All arcs with zero flow can now be discarded and the procedure terminates. Note that if $LP_Y(\mathbf{u}_Y)$ is feasible, a feasible solution to IP_Y can always be obtained by this process. [For additional discussion of heuristic procedures for LTL network design problems, see work by Powell and Sheffi *(13)* and Sheffi and Powell *(14)*.]

In Step 5, h_Y, the cost of the current heuristic solution for IP_Y, is compared with h, the cost of the current incumbent solution for IP. If $h_Y < h$, the current heuristic solution is retained as the new incumbent solution and the value of h is updated. The target value (t) in Equation 7 is also updated.

In Steps 6 and 7 it is determined whether additional effort should be expended on the current subproblem. In Step 6 it is determined whether $l_Y = t$. If so, the objective function value of all feasible solutions to IP_Y is greater than or equal to $(1 - \varepsilon)h$. Thus, if $l_Y = t$, the current subproblem can be fathomed and the

IE algorithm transfers to Step 1 to review the candidate list. On the other hand, if $l_Y < t$, the algorithm goes to Step 7.

In Step 7 the improvement in l_Y, the lower bound to the current subproblem, is tested by using a relative improvement criterion. If there is sufficient improvement in l_Y, the IE algorithm goes to Step 8 to determine a new capacity parameter vector in an effort to fathom the current subproblem. On the other hand, if the improvement in l_Y is slight, it is more efficient to partition the current subproblem in Step 9 and evaluate each of these new subproblems separately.

In Step 8 the new capacity parameter vector for the current subproblem is determined. As shown by Lamar et al. *(3)*, this step is equivalent to solving a set of linearized knapsack problems, one for each arc in the network. This means that a new capacity parameter vector, $\mathbf{u}_Y$, can be computed very efficiently. After this step has been completed, the algorithm goes to Step 3 to compute a new lower bound for the current subproblem.

In Step 9 a "partitioning" arc (denoted as arc d) is selected and the current subproblem is separated into two new subproblems—one with $Y_d = 0$ and the other with $Y_d = 1$. Remember that the lower bound in this IE procedure is obtained from the CI procedure rather than just the LP relaxation of the subproblem. Thus, partitioning methods based on the LP solution of the new subproblems—such as "up and down penalty" methods [see, e.g., work by Driebeek *(15)* or Tomlin *(16)*]—do not necessarily identify the best partition. Therefore, the partition criterion used here is to select a link whose flow in the current linear subproblem is significantly greater than zero but less than the value of the current improved capacity parameter for that link. This choice of a partitioning arc has performed well in computational experiments.

Once the partitioning link (d) has been determined, two new subproblems—$Y + \{y_d = 0\}$ and $Y + \{y_d = 1\}$—are added to the candidate list in Step 9. The capacity parameter vector for the current subproblem is also placed in the candidate list. (It is used as the initial capacity parameter vector in the CI procedure when each of the new subproblems is evaluated.) The IE algorithm then goes to Step 2 to select the next problem from the candidate list.

The IE algorithm outlined earlier produces an ε-optimal solution to IP or determines that IP is infeasible. (Note that IP is feasible if and only if LP is feasible. Thus, the feasibility of IP is determined in the first pass of the algorithm.) As mentioned earlier, the optimality criterion (ε) must be specified a priori. In cases where there is no natural tolerance (due, for example, to data accuracy), the IE algorithm can be solved several times by starting with a large value of ε. This value is then decreased until the computational effort to obtain an incumbent solution outweighs the incremental value of such a solution.

This IE algorithm has performed well in computational experiments with small LTL networks as shown in the next section.

COMPUTATIONAL EXPERIENCE

In this section the computational performance of the implicit enumeration (IE) procedure is described. It is presented in three parts: the network generation procedure used in the computational tests, the procedure for a small prototype LTL line-haul network, and some computational experience for a series of LTL networks of various sizes and with several different levels of OD shipment demand.

Network Generation Procedure

The network and OD shipment demand generation procedure used in this section is outlined as follows.

For each network generated, the node set (N) is taken as a subset of a set of 50 metropolitan areas distributed throughout the continental United States. Set N is partitioned into two sets: N_E and N_B. Here, N_E contains the set of EOL terminals where shipments originate or terminate, or both, and N_B contains the set of BB terminals where transshipment occurs. The nodes in N_E are randomly selected from among the 50 cities. The nodes in N_B are selected using a p-median heuristic *(17)*, where the $|N_B|$ BB terminals are taken as the p-medians. The medians are selected from the set of 50 cities that are not contained in N_E. (Note that in this paper, the location of all terminals is considered fixed; it is the selection of network links and the flow on these links that is designed.)

The networks generated are fully connected; that is,

$$A = \{a: I(a) \in N, J(a) \in N, I(a) \neq J(a)\}$$

Each arc $a \in A$ is a design link with associated fixed charge f_a and variable cost c_a. In addition, the BB terminals are represented as arcs in order to represent handling costs at these terminals. There are no fixed charges associated with these links.

The set of OD markets (M) consists of all EOL city pairs; that is, $|M| = |N_E|^2 - |N_E|$. For each $m \in M$, the shipment demand (q_m) is generated using a gravity-type model *(18)*. Specifically,

$$q_m = \alpha_0 \frac{(r_m)^{\alpha_1} (s_m)^{\alpha_2}}{(d_m)^{\alpha_3}} \tag{9}$$

where

r_m = daily retail sales at city $O(m)$,
s_m = daily retail sales at city $D(m)$,
d_m = highway mileage from $O(m)$ to $D(m)$, and
$\alpha_0, \alpha_1, \alpha_2, \alpha_3$ = fixed parameters.

The numerical quantities used in the computational experiments are set to reflect a realistic LTL shipping environment. For each link $a \in A$, the fixed charge f_a (in dollars) is taken as the cost of running a single truck per day over that link at \$1.10/mi. The variable cost c_a (in dollars per pound) is taken as the cost of running a single truck over link a assuming an effective payload capacity of 300 CWT. The handling cost at each BB terminal is \$1/CWT. For each $m \in M$, the values for r_m and s_m are taken as the gross retail sales for the associated Basic Trading Area (BTA) adjusted to current dollars. The distances (d_m) are in statute miles. For all networks generated, the value of the parameters in Equation 9 is fixed at $\alpha_1 = 1.0$, $\alpha_2 = 0.5$, and $\alpha_3 = 0.5$. Note that α_1 does not equal α_2, so that nonsym-

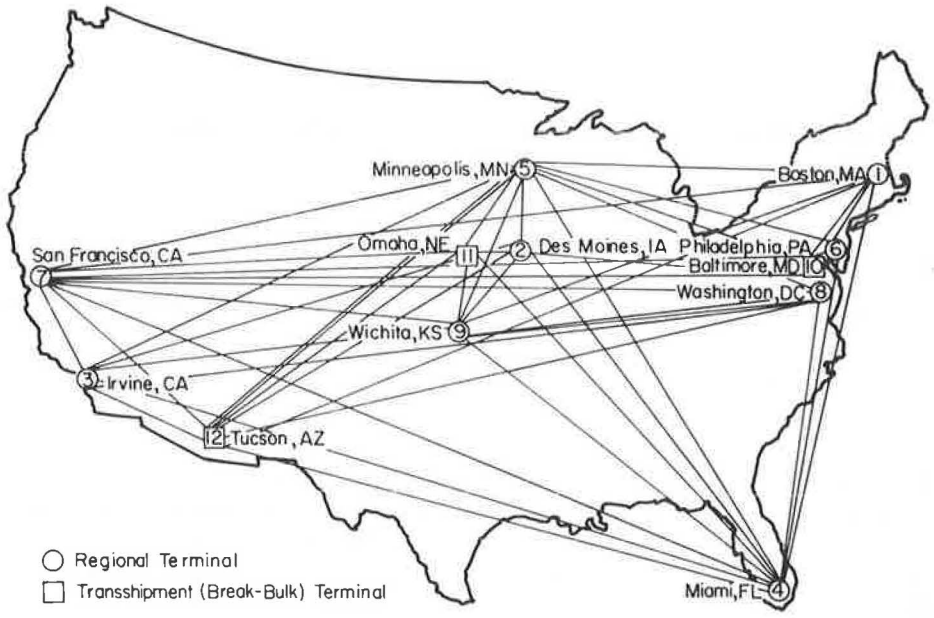

FIGURE 5 Terminals and links for prototype network.

metric shipping demands are modeled. The size of the networks is adjusted by varying $|N_E|$ and $|N_B|$ and the volume of shipment demand is adjusted by varying α_0.

Prototype LTL Load Plan

With the generation procedure just described, a small network was created to illustrate how the procedure described in this paper can aid an LTL motor carrier in designing and evaluating a line-haul network. The hypothetical network is shown in Figure 5. It consists of nine EOL terminals and three BB terminals. The network is completely connected (i.e., there is a link between every pair of terminals). Daily shipment demand between each pair of EOL terminals is given in Table 1.

As mentioned in the previous section, the IE procedure generates a heuristic network design whose objective function value is within $100 \cdot \varepsilon$ percent of the value of the optimal network design. For this example, ε was set at 0.01 (i.e., the heuristic solution is to be within 1 percent of optimal). This solution, shown in Figure 6, can be used by the LTL motor carrier in strategic studies that involve a load-plan design. Note that most of the freight in this example is routed through the BB terminals, although some shipment demand, such as that between San Francisco and Irvine, is sent directly. If the shipment demand between the EOL terminals or the number or location of the EOL or BB terminals is altered, the program can be rerun to suggest modifications in the load plan. It should be noted that such a procedure cannot be used for actual load planning because it ignores many important issues in LTL network design. These include actual trailer routing, equipment balancing, weekly demand cycles, terminal capacity, and many others. It may be applicable, however, to obtaining a general grasp of the costs of strategic alternatives.

The computational performance of the IE procedure is discussed next.

Performance Curves

In the following paragraphs an assessment is made of the tradeoffs between accuracy and computational effort that are associated with the IE procedure for several trial networks generated by the procedure described at the beginning of this

TABLE 1 SHIPMENT DEMAND FOR PROTOTYPE NETWORK

Origin Terminal[a]	Shipment Demand (CWT/day) by Destination Terminal[a]								
	1	2	3	4	5	6	7	8	9
1	–	55	12	101	101	273	91	195	40
2	25	–	3	20	48	29	24	25	16
3	2	1	–	2	2	2	6	2	1
4	89	39	10	–	69	107	70	96	32
5	84	89	11	65	–	99	80	86	44
6	294	69	15	131	127	–	109	417	51
7	105	62	44	92	110	117	–	100	51
8	178	51	11	100	93	354	80	–	38
9	15	13	2	13	19	17	16	15	–

[a]Terminal numbers refer to numbers on Figure 5.

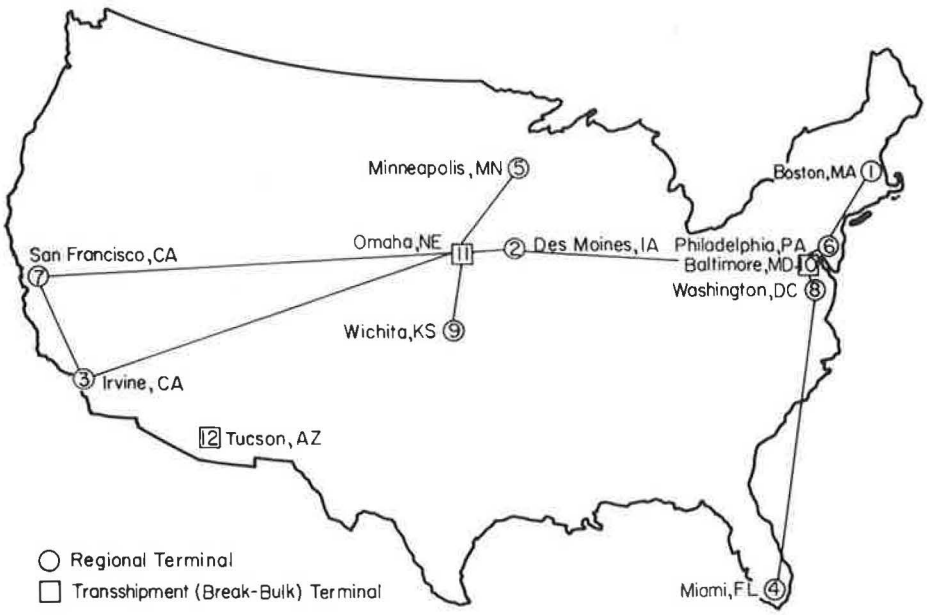

FIGURE 6 Heuristic design for prototype network.

section. For each of the networks tested, the computational time for the IE procedure was measured for various levels of the optimality criterion ε. The resulting performance curves are shown in Figures 7–13, in which ε is plotted versus the CPU seconds on a Digital Equipment Corporation VAX 780 minicomputer.

Figures 7 through 10 show the performance curves of the IE procedure for four networks ranging in size from $|N_E| = 10$, $|N_B| = 2$ [involving 90 OD pairs (markets) and 132 fixed-charge links] to $|N_E| = 40$, $|N_B| = 6$ (involving 1,560 OD pairs and 2,070 fixed-charge links). The dashed line in each figure shows the performance curve for a standard implicit enumeration (SIE) procedure in which the lower bound at each point in the enumeration tree is simply the LP relaxation of the corresponding subproblem. These curves show that the SIE produces relatively little improvement beyond the original LP relaxation (shown as the square symbol in the figures), even with significant computational effort.

The SIE procedure can be compared with the IE procedure presented in this paper. This latter procedure is shown as a solid

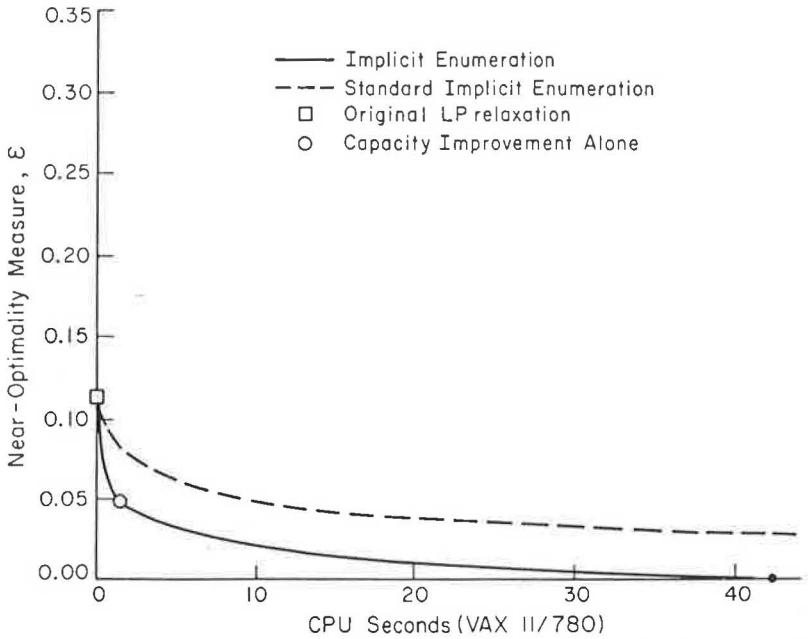

FIGURE 7 Near-optimality measure ε versus CPU time: $|N_E| = 10$, $|N_B| = 2$.

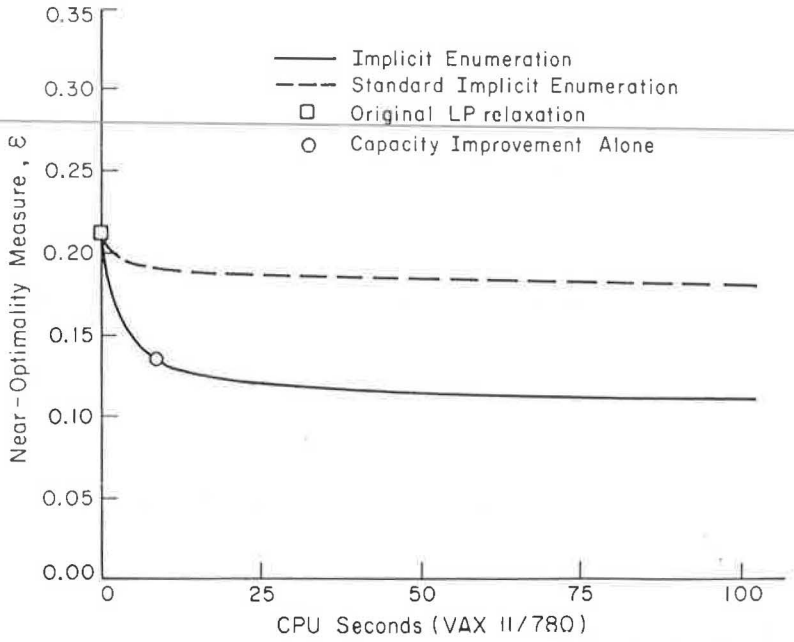

FIGURE 8 Near-optimality measure ε versus CPU time: $|N_E| = 10$, $|N_B| = 6$.

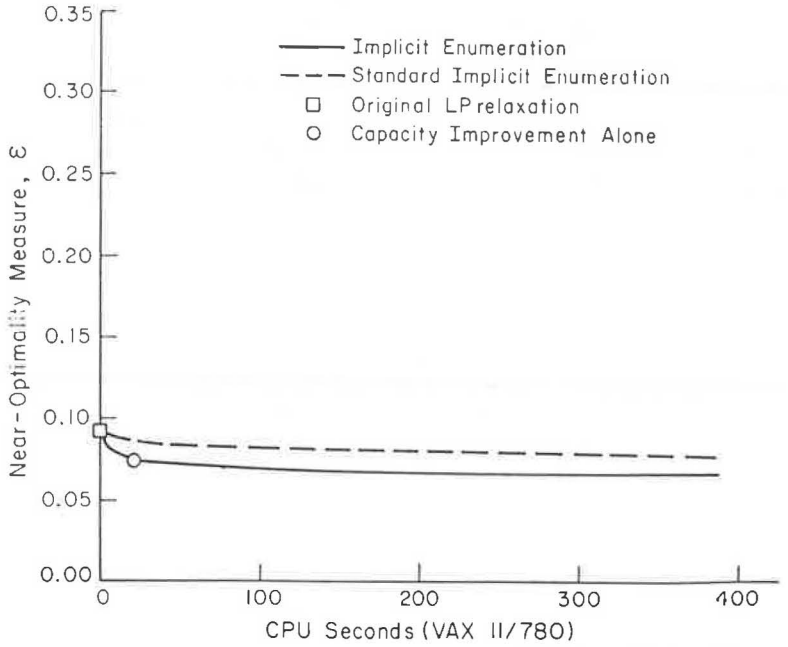

FIGURE 9 Near-optimality measure ε versus CPU time: $|N_E| = 40$, $|N_B| = 2$.

line in the figures. These curves point out the benefit obtained by incorporating the capacity improvement (CI) procedure into the IE algorithm. In fact, the majority of improvement of the IE over the SIE is due to application of the CI procedure to the root-node subproblem of the enumeration tree. This improvement is shown by the circle symbol on the curves. Beyond this point, although additional improvements in ε were obtained by the IE procedure, these improvements were usually achieved at substantial computational cost. Also, for the larger networks, although only small gains were achieved by the IE, this corresponds to essentially no improvement over the original LP relaxation for the SIE.

Note that increasing the number of OD markets increases the overall flow in the network. For the cases in which the flow

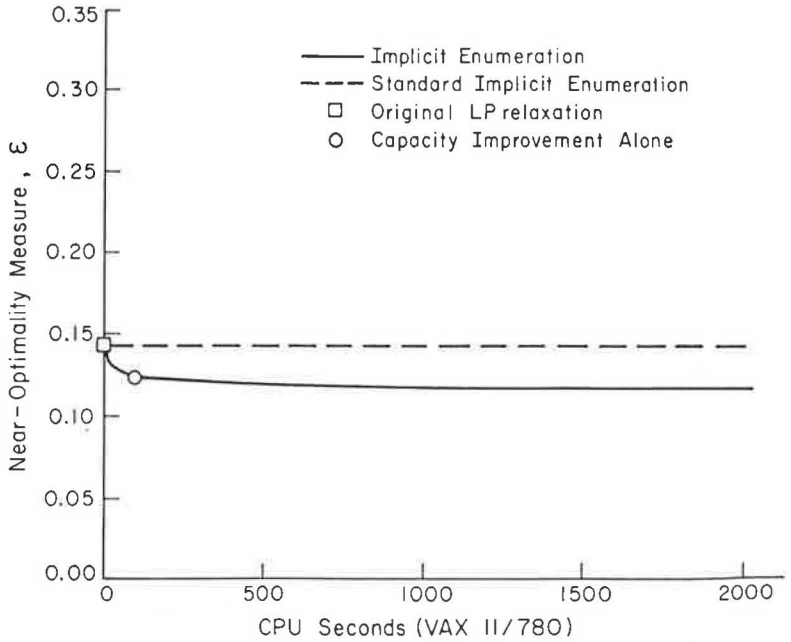

FIGURE 10 Near-optimality measure ε versus CPU time: $|N_E| = 40$, $|N_B| = 6$.

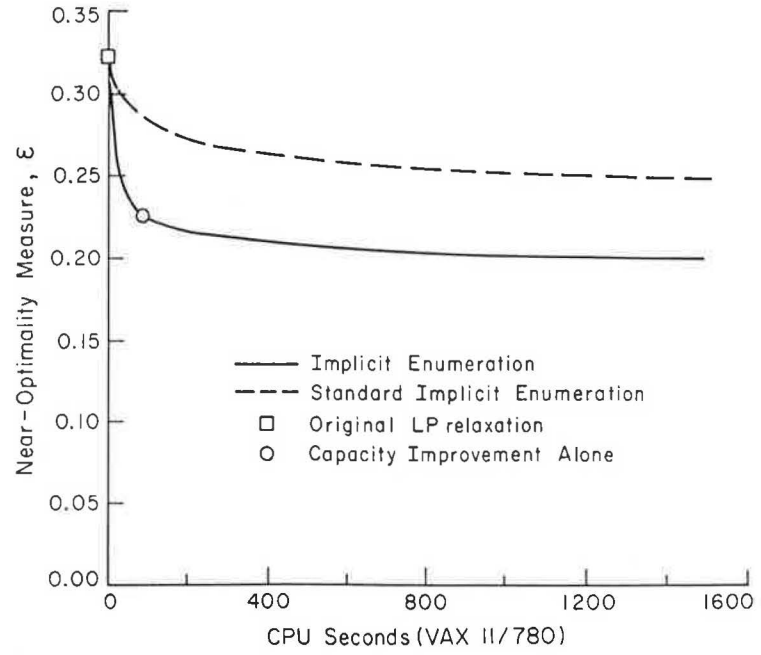

FIGURE 11 Near-optimality measure ε versus CPU time for largest trial network: low flow level.

level is relatively high, flow in an LP subproblem is more likely to be carried on direct links from the origin to destination rather than on links going to transshipment points. Thus, in this case, the CI procedure in the IE has a limited ability to tighten the lower bound to a subproblem. This inability, combined with the closeness of the original LP lower bound to z^*[IP], explains why the effect of the CI procedure in the IE diminishes as the size of the network increases in Figures 7–10.

To test the effects of flow level directly, the largest network ($|N_E| = 40$, $|N_B| = 6$) was optimized with three different flow levels: low, medium, and high. (The four networks exhibited in Figures 7–10 fall into the medium-flow category.) The volume

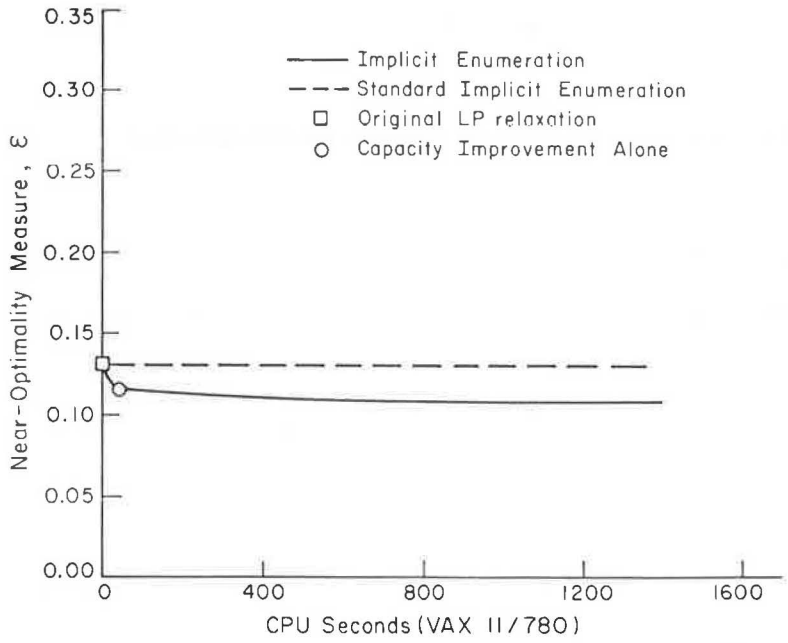

FIGURE 12 Near-optimality measure ε versus CPU time for largest trial network: medium flow level.

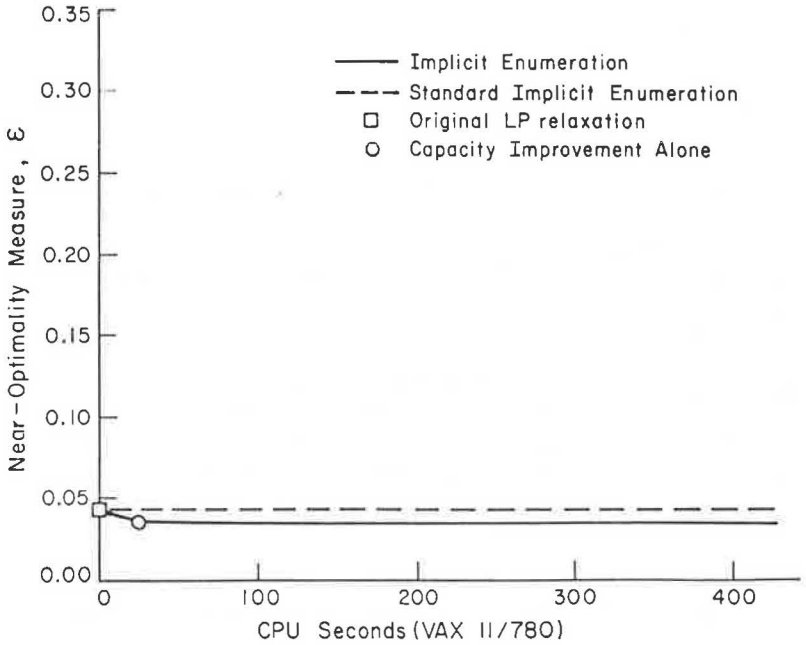

FIGURE 13 Near-optimality measure ε versus CPU time for largest trial network: high flow level.

of flow in the network can be characterized by the average link flow in the heuristic solution. For low, medium, and high flow, the average nonzero link flow in the heuristic solution was one-third, five, and eight truckloads, respectively. The fixed charge in all cases was equivalent to the cost of carrying one truckload. The results are shown in Figures 11–13. [Note that curves for the medium case (Figure 12) are not identical to those for the largest network (Figure 10) because the networks in each case were generated separately.] The designation of the SIE and IE curves is the same as that used in Figures 7–10. As expected, the IE always outperforms the SIE, but this improvement diminishes as the flow level increases.

In reviewing Figures 7–13, it seems that where optimality is not attained, the IE (and SIE) procedure cannot overcome

certain values of ε. That is, for small tolerances, the exponential growth of the enumeration tree predominates and the IE curves track those of the SIE. This overall pattern reflects the difficult nature of the problem.

The IE procedure presented in this paper appears to have strong merits, particularly for situations with relatively low levels of flow, as in LTL freight networks. Markets in LTL line-haul networks are characterized by flows that are at the low range of these computational experiments. The reason for this is that high link flow levels would typically indicate an opportunity for bypassing a transshipment point or opening a new terminal and consequently LTL carriers operate at the low flow range on most of their links (excluding links between break-bulks, which many operate with very high flows). Thus, the design problem has to do with the portions of the LTL motor carrier's network with low flow. In these cases, the fixed charges correspond to a minimum frequency of, say, one truck-load per day and arc flows are in the range of one-third to two truckloads per day. It is in this interesting range that the IE procedure yields the most improvement.

ACKNOWLEDGMENTS

This research was supported by I. U. International, Inc., through a research contract with Massachusetts Institute of Technology and Princeton University. The authors would like to thank John Terry and Al Pinkerton of I. U. International, Inc., and the personnel in the Industrial Engineering Department of Ryder/PIE truck lines for their assistance and support throughout this project.

REFERENCES

1. A. W. LaMond. *Competition in the General-Freight Motor-Carrier Industry.* Lexington Books, Lexington, Mass., 1980.
2. W. B. Powell and Y. Sheffi. The Load Planning Problem of Motor Carriers: Problem Description and a Proposed Solution Approach. *Transportation Research*, Vol. 17A, No. 6 (Nov.), 1983, pp. 471–480.
3. B. W. Lamar, Y. Sheffi, and W. B. Powell. *A Lower Bound Procedure for Uncapacitated, Multicommodity Fixed Charge Network Design Problems.* Report UCI-ITS-WP-86-4. Institute of Transportation Studies, University of California, Irvine, 1986.
4. R. T. Wong. *Accelerating Bender's Decomposition for Network Design.* Ph.D. dissertation. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, 1978.
5. R. L. Rardin and U. Choe. *Tighter Relaxations of Fixed Charge Network Flow Problems.* Report J-79-18. School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, 1979.
6. A. Balakrishnan. *Valid Inequalities and Algorithms for the Network Design Problem with an Application to LTL Consolidation.* Ph.D. dissertation. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 1984.
7. T. L. Magnanti and R. T. Wong. Network Design and Transportation Planning: Models and Algorithms. *Transportation Science,* Vol. 18, No. 1 (Feb.), 1984, pp. 1–55.
8. M. L. Balinski. Fixed-Cost Transportation Problems. *Naval Research Logistics Quarterly*, Vol. 8, No. 1 (March), 1961, pp. 41–54.
9. R. W. Floyd. Algorithm 97—Shortest Path. *ACM: Communications of the ACM,* Vol. 5, No. 6 (May), 1962, p. 345.
10 E. Moore. The Shortest Path Through a Maze. In *Proceedings of the International Symposium on the Theory of Switching,* Harvard University Press, Cambridge, Mass., 1957, pp. 285–292.
11. F. Glover, D. Karney, and D. Klingman. Implementation and Computational Comparisons of Primal, Dual, and Primal-Dual Computer Codes for Minimum Cost Network Flow Problems. *Networks*, Vol. 4, No. 3 (Fall), 1974, pp. 191–212.
12. A. M. Geoffrion and R. E. Marsten. Integer Programming Algorithms: A Framework and State-of-the-Art Survey. In *Perspectives on Optimization: A Collection of Expository Articles* (A. M. Geoffrion, ed.), Addison-Wesley, Reading, Mass., 1972.
13. W. B. Powell and Y. Sheffi. Design and Implementation of an Interactive Optimization System for Network Design in the Motor Carrier Industry. *Operations Research,* in press.
14. Y. Sheffi and W. B. Powell. Interactive Optimization for Motor Carrier Load Planning. *Journal of Business Logistics*, Vol. 7, No. 2, 1986.
15. N. J. Driebeek. An Algorithm for the Solution of Mixed Integer Programming Problems. *Management Science,* Vol. 12, No. 7 (March), 1966, pp. 576–587.
16. J. A. Tomlin. An Improved Branch-and-Bound Method for Integer Programming. *Operations Research,* Vol. 19, No. 4, (July–Aug.), 1971, pp. 1070–1075.
17. M. B. Teitz and P. Bart. Heuristic Methods for Estimating the Generalized Vertex Median of a Weighted Graph. *Operations Research,* Vol. 16, No. 5 (Sept.-Oct.), 1968, pp. 955–961.
18. P. R. Stopher and A. H. Meyburg. *Urban Transportation Modelling and Planning.* D.C. Heath and Co., Lexington, Mass., 1975.

Optimization of Group Line-Haul Operations for Motor Carriers Using Twin Trailers

JONATHAN ECKSTEIN AND YOSEF SHEFFI

Group line-haul operation involves the daily movement of trailers between a central breakbulk terminal and a set of end-of-line satellite terminals. When each tractor can pull two trailers, there are many possibilities for creating tractor tours that accomplish the required pickup, delivery, and empty-balancing operations. The challenge is to create optimal tours that minimize transportation costs. An optimization procedure is described that is based on a branch-and-bound framework. It involves Lagrangian relaxation for lower bounds and two upper-bound heuristics for solving this problem.

Less-than-truckload (LTL) motor carriers transport shipments (mostly ranging between a few hundred and a few thousand pounds) between many origins and destinations. Typically, large LTL carriers maintain a three-tiered network that operates as follows:

1. Shipments are picked up from individual customers and taken a short distance to a local end-of-line (EOL), or city, terminal.
2. Local shipments are consolidated in the EOL terminal and transferred to a regional breakbulk terminal, or "break," up to a few hundred miles away. There the incoming shipments are sorted and consolidated by destination into outbound trailers.
3. Each shipment then travels over a network of main-line routes (interconnecting breakbulk terminals) until it reaches the break serving its destination region.
4. In the reverse of Step 2, each shipment travels to the EOL terminal serving its destination city.
5. In the reverse of Step 1, shipments are delivered to individual consignees.

The focus of this paper is on the use of twin trailer trucks in the second step, known as group line-haul operations because LTL networks can be divided into groups, each containing a break and a set of EOL terminals connected to it.

Over the past few years, regulatory changes have allowed twin trailer trucks to be used much more widely on U.S. highways. In such combinations, a single tractor may haul two 28-ft trailers ("doubles" or "pups") instead of one 45- or 48-ft trailer (a van or "semi"). A tractor may also travel alone ("bobtail") or pull just one short trailer. Doubles have proved extremely popular with LTL carriers, particularly on main-line routes [see report by Sheffi and Powell *(1)*]. There they provide extra carrying capacity and improve the level of service without increasing costs.

Although doubles cannot be used in city pickup-and-delivery operations, they have the potential for reducing the costs of group line-haul operations between the EOL terminals and the regional break. The means of achieving such savings are not always obvious because of the combinatorial nature of the problem.

An optimization procedure is outlined for determining daily routes for twin trailer combinations in group line-haul operations. It is based on a multicommodity aggregate flow formulation [see paper by Magnanti *(2)* for classification of methods]. The solution method combines a Lagrangian relaxation for calculating lower bounds with two incumbent generation heuristics to calculate upper bounds, all imbedded in a branch-and-bound (B&B) framework. First the problem is presented in detail; then it is formulated as an integer program. The Lagrangian relaxation is discussed, and the heuristics for the upper bound are described. The details of the B&B procedure are outlined, and, last, some numerical results and a concluding section are presented.

PROBLEM STATEMENT

A group line-haul operation includes one break and a set (5 to 50) of EOL terminals; the latter are responsible for delivery and pickup of shipments at customer locations during business hours. Consequently, the group line-haul operation takes place, for the most part, during the night.

Each afternoon the line-haul planner, at the break, receives data on the number of trailers to be picked up from and delivered to each EOL terminal. Because the operation is repeated nightly, the planner must also consider trailer balancing—empty trailers should be removed from terminals when deliveries exceed pickups and supplied to terminals when pickups exceed deliveries. [In some cases (not considered here) empty trailers are balanced only on a weekly basis. In these cases the trailer inventory at each EOL terminal on each day is a decision variable.]

Trailer movements are accomplished by tractors that can pull up to two trailers at a time. At each terminal a tractor may drop off or pick up either one or two trailers (or do both). The tractor pool must also be balanced daily.

The objective is to cover all the required trailer movements with the minimum number of total tractor miles. The distances between all terminals (naturally) obey the triangle inequality.

Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

These distances are typically small enough (and the supply of tractors is large enough) so that the minimum mileage solution can be executed in one day.

Consider a group line-haul operation including one break and a set of EOL terminals numbered 1, 2, . . ., n. Let p_i denote the number of trailers to be picked up on a given day at EOL terminal i and let d_i denote the number of trailers to be delivered at i.

To get a feel for the options available to the planner, consider two EOL terminals (1 and 2), each requiring delivery and pickup of a single trailer (i.e., $p_1 = d_1 = p_2 = d_2 = 1$). If a tractor could haul only one trailer at a time, the solution would be simple. One tractor pulling outbound freight would be dispatched from the break to each EOL terminal. Each of these two tractors would then drop off its outbound load and pick up an inbound load to take back to the break (Figure 1).

If a tractor can haul two trailers, however, there is a lower-mileage solution using one tractor tour to visit both EOL terminals. The tractor is dispatched from the break pulling both outbound trailers. At the first EOL terminal, it exchanges one of these trailers for a trailer loaded with inbound shipments and then proceeds to the second EOL terminal. There it exchanges the second outbound trailer for an inbound one and then continues to the break. This solution is shown in Figure 2. By the triangle inequality, there are fewer tractor miles in the second solution than in the first.

Although this last example could be solved by inspection, larger problems are dramatically more complex. To see this, consider the four-EOL example shown in Figure 3.

One of the possible solutions resulting in minimum tractor mileage for these data is shown in Figure 4, in which each trailer is represented by a box labeled with the destination of its contents, where B stands for the break and empty trailers are unlabeled. Each set of two adjacent boxes represents a twin-trailer combination pulled by a tractor-trailer. Clearly, the combinatorics of the problem can get quite involved, even for a case with so few nodes.

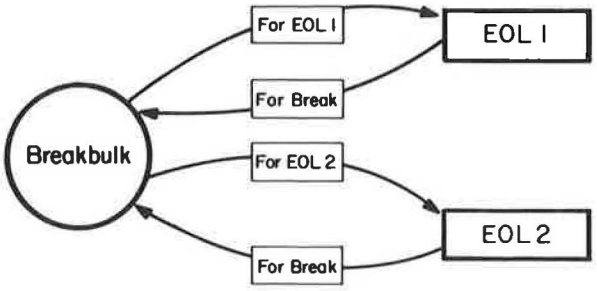

FIGURE 1 Group line-haul operation with two EOL terminals and one trailer per truck.

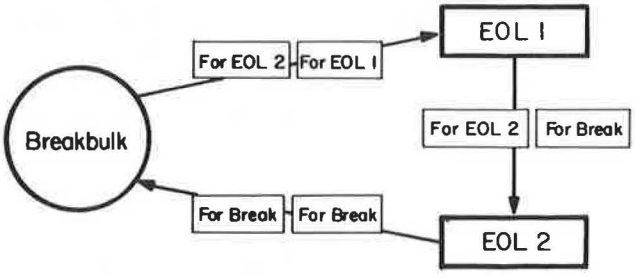

FIGURE 2 Group line-haul operation with two trailers per truck.

The group line-haul problem addressed here exhibits many elements of classical vehicle routing problems: the objective is to create tractor tours from a depot (the break) visiting all customers (EOL terminals) and picking up and dropping off shipments (trailers). A comprehensive review of vehicle routing and practice has been given by Bodin et al. *(3)*. Note, however, that the group line-haul problem addressed here differs from that body of literature in several important ways: (a) shipments (number of inbound and outbound trailers) are often larger than the vehicle size (trailer-pulling capacity), and (b) both pickups and deliveries as well as empty balancing are involved, rather than a single operation.

A better background for the work reported here is provided by Magnanti's survey *(2)*. The emphasis there is on formulations and mathematical programming issues related to basic vehicle-routing problems.

A related and somewhat more general problem is railroad blocking. There the question is which set of freight cars should make up blocks on particular trains. This is a generalization of the group line-haul problem because locomotives can carry more than two cars and because cars move between multiple origin and destination yards (rather than in and out of a central terminal). Assad *(4)* explores several methods for dealing with that problem.

FORMULATION

Consider a terminal group with one break, numbered 0, and n EOL terminals (numbered 1, . . ., n as before) with pickups and deliveries p_i and d_i, respectively, at each i. Define

$$d_0 \equiv -\sum_{i=1}^{n} d_i \qquad p_0 \equiv -\sum_{i=1}^{n} p_i \tag{1}$$

and

$$\mathbf{p} \equiv (p_0, \ldots, p_n)^T$$
$$\mathbf{d} \equiv (d_0, \ldots, d_n)^T \tag{2}$$

Further, define a complete network whose nodes are these $n + 1$ terminals. In this network let $\mathbf{A}$ be the node-arc incidence matrix and $\mathbf{c}$ be the vector of corresponding distances or costs (c_{ij}) of driving a tractor from node i to node j.

Now define four different flow vectors $\mathbf{t}$, $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{e}$ over the network given by $\mathbf{A}$. These vectors are all conformal to $\mathbf{c}$:

$\mathbf{t}$ is a vector of scalers t_{ij}, each giving the number of tractors on arc (i, j);

$\mathbf{x}$ is a vector of scalers x_{ij}, each giving the number of outbound trailers (from the break to the EOL terminals) on arc (i, j);

$\mathbf{y}$ is a vector of scalers y_{ij}, each giving the number of inbound trailers (to the break) on link (i, j); and

$\mathbf{e}$ is a vector of scalers e_{ij}, each giving the number of empty trailers moving on link (i, j).

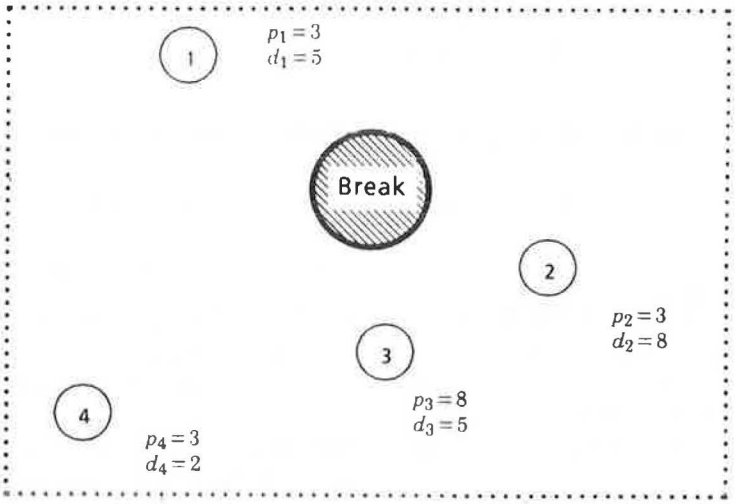

FIGURE 3 More complex example: four EOL terminals.

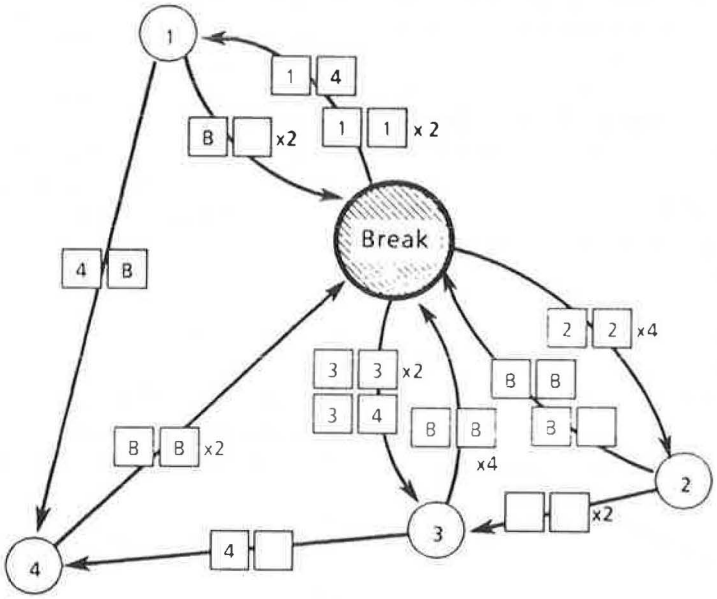

FIGURE 4 Solution for example in Figure 3.

The group line-haul problem can now be formulated as follows:

Minimize

$$\mathbf{c}^T\mathbf{t} \tag{3a}$$

such that

$$\mathbf{At} = 0 \tag{3b}$$

$$\mathbf{Ax} = -\mathbf{d} \tag{3c}$$

$$\mathbf{Ay} = \mathbf{p} \tag{3d}$$

$$\mathbf{Ae} = \mathbf{d} - \mathbf{p} \tag{3e}$$

$$2\mathbf{t} - \mathbf{x} - \mathbf{y} - \mathbf{e} \geq 0 \tag{3f}$$

$$\mathbf{t}, \mathbf{x}, \mathbf{y}, \mathbf{e} \geq 0 \tag{3g}$$

$$\mathbf{t}, \mathbf{x}, \mathbf{y}, \mathbf{e} \quad \text{integer} \tag{3h}$$

Constraint 3b guarantees the flow conservation for tractors, and constraints 3c and 3d guarantee that all pickups and deliveries are made. If $p_i - d_i > 0$, that number of empty trailers has to be supplied to terminal i, whereas if $p_i - d_i < 0$, then $|p_i - d_i|$ trailers must be removed from i. Thus Equation 3e requires the trailer inventory to be stationary. Constraint 3f couples the tractor to the trailers. It says that a tractor can pull at most two trailers; that is,

$$1/2(x_{ij} + y_{ij} + e_{ij}) \leq t_{ij} \qquad \forall (i, j) \tag{4}$$

Equation 3f is a rearrangement of Equation 4 in vector form.

Note that the optimum value of $\mathbf{t}$ is a flow vector and not an explicit set of driver instructions. [This formulation resembles

the one given by Gavish and Graves *(5)* for the traveling salesman problem.] To extract such instructions, the solution has to be decomposed into tours (typically all beginning and ending at the same terminal) and the trailers have to be assigned to the legs of all such tours. Such an extraction is clearly possible: because **t** is a balanced flow with no exogenous sources or sinks, it can be decomposed into tours by a process very similar to that for finding a Euler tour of an even-degree graph.

The solution method used here is based on a B&B method. The lower bound at every B&B subproblem is derived from a Lagrangian relaxation, whereas the incumbent generation is based on ad hoc heuristics. The details of the B&B implementation are explained two sections later, after the relaxation has been outlined and the heuristics have been explained.

THE LAGRANGIAN DUAL

Program 3 consists of four separate network problems linked by additional constraints. A standard technique for dealing with problems with such recognizable embedded structure is to dualize the linking constraints *(6, 7)*. This yields, for each nonnegative $\mathbf{u} \in \Re^{(n+1)n}$ [there are $(n+1)n$ linking constraints], the following Lagrangian relaxation:

$$L(\mathbf{u}) = \text{Min } \mathbf{c}^T\mathbf{t} - \mathbf{u}^T(2\mathbf{t} - \mathbf{x} - \mathbf{y} - \mathbf{e}) \tag{5a}$$

such that

$$\mathbf{At} = 0 \tag{5b}$$

$$\mathbf{Ax} = -\mathbf{d} \tag{5c}$$

$$\mathbf{Ay} = \mathbf{p} \tag{5d}$$

$$\mathbf{Ae} = \mathbf{d} - \mathbf{p} \tag{5e}$$

$$\mathbf{t, x, y, e} \geq 0 \tag{5f}$$

$$\mathbf{t, x, y, e} \quad \text{integer} \tag{5g}$$

$L(\mathbf{u})$ in Equation 5a is a lower bound on the optimum objective value for the original problem for all $\mathbf{u} \geq 0$. The cost coefficients in Equation 5a can then be rearranged to emphasize the structure of four independent network problems. Furthermore, because network problems have integer optima if their right-hand sides are integer, constraint 5g can be dropped from the formulation, which becomes

$$L(\mathbf{u}) = \text{Min } (\mathbf{c} - 2\mathbf{u})^T\mathbf{t} + \mathbf{u}^T\mathbf{x} + \mathbf{u}^T\mathbf{y} + \mathbf{u}^T\mathbf{e} \tag{6a}$$

such that

$$\mathbf{At} = 0 \tag{6b}$$

$$\mathbf{Ax} = -\mathbf{d} \tag{6c}$$

$$\mathbf{Ay} = \mathbf{p} \tag{6d}$$

$$\mathbf{Ae} = \mathbf{d} - \mathbf{p} \tag{6e}$$

$$\mathbf{t, x, y, e} \geq 0 \tag{6f}$$

The best lower bound (D) would be obtained as

$$D \equiv \text{Max}_{\mathbf{u}\geq 0} L(u) \tag{7}$$

By the strong duality theory of linear programming, $D = Z_{LP}$, where Z_{LP} is the optimum of the LP relaxation of program 3. Thus, the Lagrangian relaxation can produce a lower bound to program 3 that is only as tight as the LP relaxation of program 3 (given the correct choice of **u**).

As it turns out, it is possible to pick a priori the values of the multipliers **u** that yield the largest possible value of $L(\mathbf{u})$. The appropriate choice is

$$\mathbf{u}^* \equiv (1/2)\,\mathbf{c} \tag{8}$$

This gives each arc (i, j) an imputed tractor cost of 0 and an imputed trailer cost of $c_{ij}/2$. Essentially, one can think of this as attaching half a tractor to each trailer. The constraint $2\mathbf{t} - \mathbf{x} - \mathbf{y} - \mathbf{e} \geq 0$ is then automatically satisfied with zero slack, and trailers circulate in a shortest-path manner subject to a lowest-cost repositioning of empties.

In order to strengthen relaxation 6, consider the original integer program (3). Note that if d_i trailers must be delivered at EOL terminal i, at least $\lceil d_i/2 \rceil$ tractors must visit that terminal (where $\lceil \cdot \rceil$ denotes the upwards integer rounding function). Similarly, at least $\lceil d_0/2 \rceil$ tractor trips must be dispatched from the break. An analogous argument applies to pickups, and so one can conclude that the minimum number of tractors to pass through terminal i is given by

$$v_i \equiv \text{Max } \{\lceil |p_i|/2 \rceil, \lceil |d_i|/2 \rceil\} \quad \text{for } i = 0, 1, \ldots, n \tag{9}$$

This observation can be used to add the following set of $(n+1)$ constraints to program 3:

$$\sum_{j \neq i} t_{ij} \geq v_i \quad \text{for } i = 0, \ldots, n \tag{10}$$

Such "node-activity" constraints are redundant and may be added to the basic aggregate flow formulation without altering it. However, when the linking constraints are dualized, the node-activity constraints are no longer redundant: they alter the tractor part of the computation of $L(\mathbf{u})$ considerably. With the addition of these constraints this part of the problem becomes

$$\text{Min } (\mathbf{c} - 2\mathbf{u})^T\mathbf{t} \tag{11a}$$

such that

$$\mathbf{At} = 0 \tag{11b}$$

$$\sum_{j \neq i} t_{ij} \geq v_i \qquad \text{for } i = 0, \ldots, n \tag{11c}$$

$$\mathbf{t} \geq 0 \tag{11d}$$

Constraint 11c can be viewed as cutting planes that disallow the previous LP optimum, $\mathbf{t}^* = 1/2(\mathbf{x}^* + \mathbf{y}^* + \mathbf{e}^*)$, unless it is integer (and hence optimal). The Lagrangian relaxation is thus strengthened by the addition of these constraints. By using a

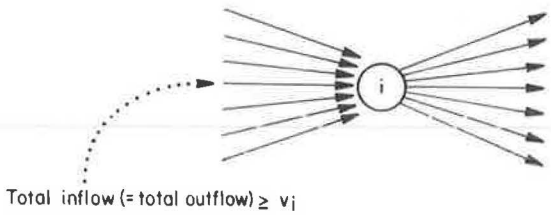

FIGURE 5 Before node splitting.

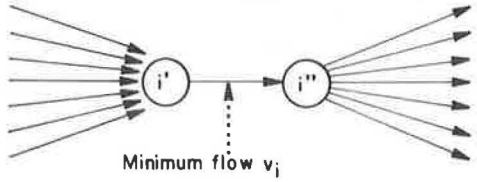

FIGURE 6 After node splitting.

standard "node-splitting" technique *(8)*, constraints 11c may be handled without destroying the network structure of the tractor subproblem (Figures 5 and 6). In practice, the addition of the node-activity constraints improves the efficiency of the Lagrangian B&B method by a factor of approximately 5.

CALCULATING THE LOWER BOUND

At any given point in the B&B tree, Lagrangian 5 (along with constraints 11c and additional separation constraints) has to be solved. Note that the separation constraints include only upper and lower bounds on arc flows, and consequently they do not disturb the network structure of this program. Given $\mathbf{u}$, $L(\mathbf{u})$ can be computed by applying the primal network simplex four times (with the appropriate network modification for the tractor part to enforce the node-activity constraints).

Given a solution to the program consisting of Equations 5 plus constraints 11c plus separation constraints, $L(\mathbf{u})$ is a lower bound to the program consisting of Equations 3 plus separation constraints. To get a set of multipliers that would give a better lower bound, one can use a subgradient method. With this method the new value of the multipliers is

$$\mathbf{u} := \mathbf{u} + s\mathbf{g}, \tag{12a}$$

where the direction $\mathbf{g}$ is given by the subgradient

$$\mathbf{g} \equiv \mathbf{x} + \mathbf{y} + \mathbf{e} - 2\mathbf{t}, \tag{12b}$$

and the step s is given by

$$s = a\frac{Z_{INC} - L(\mathbf{u})}{||\mathbf{g}||^2} \tag{12c}$$

This is a standard subgradient step, where a is the step size and the "target" value of $L(\mathbf{u})$ is taken to be Z_{INC} (the objective function value of the incumbent). This is the highest possible value one may choose for the target. It is reasonable only because the incumbents are usually very good.

The initial value of $\mathbf{u}$ in a given subproblem of the B&B tree is taken to be the terminal $\mathbf{u}$ in the predecessor ("parent") subproblem. In the initial problem, best results were obtained with $\mathbf{u} = \mathbf{c}/4$. This is the center of the "box" $\{\mathbf{u}|0 \le \mathbf{u} \le \mathbf{c}/2\}$ which, as shown in the following, approximates the dual feasible region.

If the convergence of the subgradient algorithm for a given subproblem is slow, it may be best to separate the subproblem. The rule used to terminate is the following:

If for iteration $m > K$,

$$\frac{L(\mathbf{u}) - Z_{INC}}{L(\mathbf{u}_0) - Z_{INC}} \ge 1 - m\delta, \text{ STOP} \tag{13}$$

where $\mathbf{u}_0$ is the first multiplier vector used for the subproblem, and K and δ are adjustable parameters. In other words, the algorithm has to go through at least K iterations. It then splits if the average improvement is no more than a fraction δ of the way to the incumbent value per step.

The reason that a minimum of K iterations must be performed before separation is that one cannot rely upon $L(\mathbf{u})$ to increase at every step, even when it is far from its maximum value for the subproblem at hand. Without the $m > K$ restriction, a single "bad" step at the outset of subproblem analysis would cause an immediate, unnecessary separation. Every such mishap would double the work the method must do to fathom a particular branch of the enumeration tree. Separations are thus costly and should be avoided unless they are absolutely necessary. On small test problems, values of $K = 3$ and $\delta = 0.02$ proved efficient.

Note that if at some iteration, as the result of a subgradient step, $u_{ij} > c_{ij}/2$ for some (i, j), the imputed cost $c_{ij} - 2u_{ij}$ in the Lagrangian tractor network will be negative. Because this entails a high risk of creating a negative cycle in the imputed tractor costs, $\mathbf{u}$ should be restricted to the dual feasible region:

$$U \equiv \{\mathbf{u} \ge 0 | \mathbf{c} - 2\mathbf{u} \text{ has no negative cycles}\} \tag{14}$$

This region is a polytype, but it has an exponentially growing number of constraints. It is therefore easier to make the approximate restriction

$$0 \le \mathbf{u} \le \mathbf{c}/2 \tag{15}$$

which assures, more strongly, that $\mathbf{c} - 2\mathbf{u}$ has no negative cost arcs. After each subgradient step, the resulting new value of $\mathbf{u}$ is projected onto this subset of the feasible region. Because this region is box-shaped, the projection is simply

$$u_{ij} := \text{Max}\{0, \text{Min}\{u_{ij}, c_{ij}/2\}\} \qquad \forall(i, j) \tag{16}$$

Strategies other than projection for enforcing that $0 \le \mathbf{u} \le \mathbf{c}/2$ (such as truncating the step) appear not to allow $L(\mathbf{u})$ to grow as quickly.

INCUMBENT GENERATION

Given a solution to a relaxed subproblem, one needs to find an incumbent. With the Lagrangian-based lower bound, one ob-

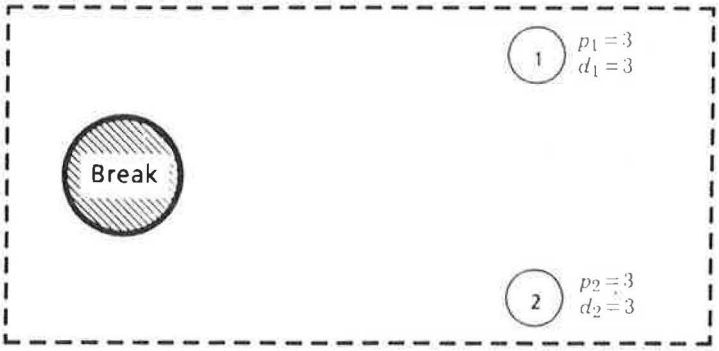

FIGURE 7 Roundup fails—problem data.

vious strategy presents itself. Given the (**t**, **x**, **y**, **e**) optimal in $L(\mathbf{u})$, **t** can be replaced with the minimum-cost integer tractor flow vector $\mathbf{t}_z$ required to support the trailer flows **x**, **y**, and **e**. On each link (i, j), this flow must be at least $(x_{ij} + y_{ij} + e_{ij})/2$, but also integer, hence at least $\lceil (x_{ij} + y_{ij} + e_{ij})/2 \rceil$. However, the tractor flow must also be balanced and consequently t_z should be the optimal solution to the problem

$$\text{Min } \mathbf{c}^T\mathbf{t} \tag{17a}$$

such that

$$\mathbf{At} = 0 \tag{17b}$$

$$\mathbf{t} \geq \lceil 1/2\ (\mathbf{x} + \mathbf{y} + \mathbf{e}) \rceil \tag{17c}$$

Program 17 is a network circulation problem and may be solved by network simplex. The quadruplet $(\mathbf{t}_z, \mathbf{x}, \mathbf{y}, \mathbf{e})$ is then a feasible solution to the integer program 3. Unfortunately, this simple roundup incumbent generation strategy is inadequate. To demonstrate this, consider the simple example problem in Figure 7. One of the two optimal solutions to this group line-haul problem is given in Figure 8. (In the other optimum, the roles of EOL terminals 1 and 2 are interchanged.) Because the algorithm separates only on the t_{ij} (tractor) variables, as $L(\mathbf{u})$ is computed, all trailers will still always be free to take the shortest path to their final destinations. For instance, all outbound traffic for Terminal 2 could take the route shown in Figure 9 or, with a different **u**, the route shown in Figure 10. However, the IP optimum includes routings in which two different paths are used, such as the one shown in Figure 11.

Splitting only on the **t** variables, the roundup heuristic never changes any trailer routes, so it can never discover the optimum. Two courses of action can be used to detect the optimum: separating on trailer variables and creating an incumbent finder that intelligently alters trailer routes. The former

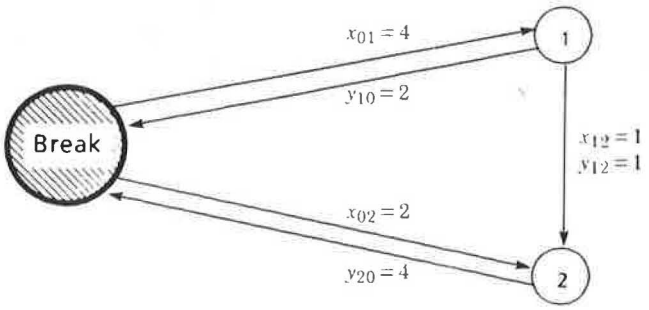

FIGURE 8 Roundup fails—integer optimum.

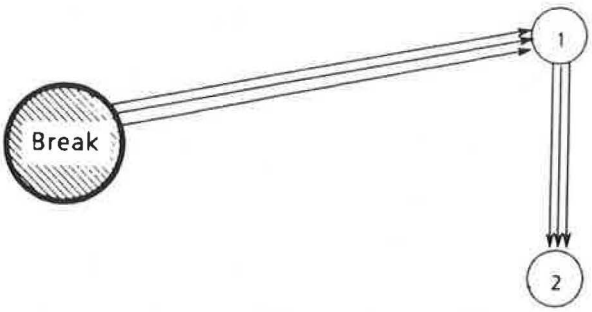

FIGURE 10 Roundup fails—another possible routing from break to 2 in $L(\mathbf{u})$.

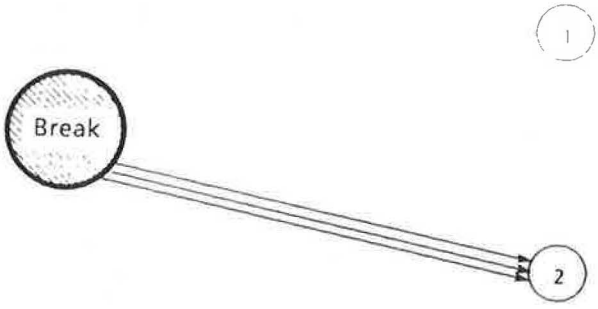

FIGURE 9 Roundup fails—one possible routing from break to 2 in $L(\mathbf{u})$.

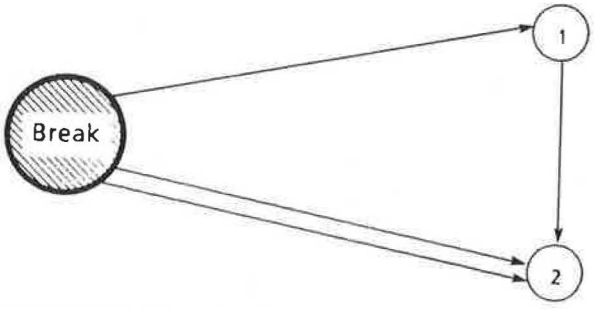

FIGURE 11 Roundup fails—a routing that cannot occur in $L(\mathbf{u})$.

strategy puts most of the burden of finding a solution on the enumeration component of the algorithm—which may lead to a very large B&B tree. Consequently the second course of action was chosen.

In the following sections two incumbent generation schemes based on local improvement heuristics are described. The focus in the first one is on developing a method that will run quickly, whereas the focus of the second is on getting accurate solutions. These methods are explained separately.

Method 1: Simple Local Improvements

The first heuristic takes the $L(\mathbf{u})$ solution and makes some local modifications to it. The trailer routings are slightly perturbed [ignoring the $\mathbf{t}$ part of the $L(\mathbf{u})$ solution] so that the minimum integer tractor movements needed to "cover" them can be reduced. In the combined $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{e}$ solution to $L(\mathbf{u})$, the method detects all patterns similar to those shown in Figure 12, where solid arrows represent arcs with an odd total number of trailers.

The local improvements shown in Figures 13–15 are applied, respectively, to each of the three patterns in Figure 12 (dashed lines represent arcs whose trailer flows have been increased from odd to even values, and boxes represent particular trailers). The modifications are then ranked by their total savings, that is, c_{0b} in the first case, c_{a0} in the second, and

$$(c_{0a} + c_{a0} + c_{0b} + c_{b0}) - (c_{0a} + c_{b0} + c_{ab}) = c_{a0} + c_{0b} - c_{ab} \geq 0 \qquad (18)$$

in the third. These improvements are then implemented in a greedy manner, highest savings first, until no more can be implemented. (Note that some of the detected patterns might overlap so that performing one might preclude applying others.) After these modifications have been made, the flow-based roundup procedure (program 17) is performed to compute the $\mathbf{t}$ part of the candidate incumbent, this time using the modified trailer flows as input.

As shown by Eckstein *(9)* the method described earlier cannot identify all the possible patterns that can be improved to generate a lower-cost tractor flow. More improvement patterns can be found if the minimum tractor covering is calculated first (given a solution of the Lagrangian relaxation) and then the resulting slack capacity is investigated for promising patterns. This is the basis of the second method for generating incumbents.

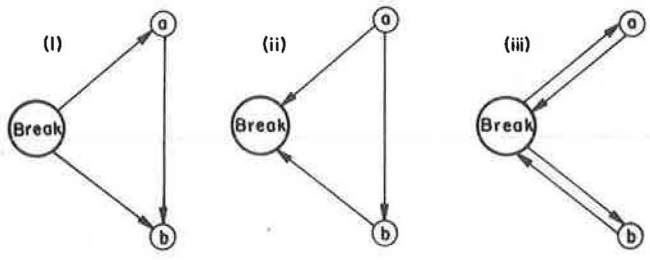

FIGURE 12 Patterns recognized by simple local improvement heuristic.

Method 2: Cycle Splicing

Given a "round-up solution" $(\mathbf{t}_z, \mathbf{x}, \mathbf{y}, \mathbf{e})$, let the slack vector $\mathbf{s}$ be defined as

$$\mathbf{s} \equiv 2\mathbf{t}_z - \mathbf{x} - \mathbf{y} - \mathbf{e} \qquad (19)$$

Since $\mathbf{A}\cdot\mathbf{t} = 0$ (see program 11) and $\mathbf{s} \geq 0$, $\mathbf{a}\cdot\mathbf{s} = 0$. To see this, note the following:

$$\begin{aligned} \mathbf{As} &= 2\mathbf{At}_z - \mathbf{Ax} - \mathbf{Ay} - \mathbf{Ae} \\ &= 0 + \mathbf{d} - \mathbf{p} - (\mathbf{d} - \mathbf{p}) \\ &= 0 \end{aligned} \qquad (20)$$

Thus, the vector of slack capacities $\mathbf{s}$ may also be considered a balanced "flow" in the interterminal network. Now any balanced flow can be expressed as a sum of flows around simple directed cycles. (A simple cycle is defined as one that repeats neither arcs nor nodes.)

Definition: A cycle decomposition of some balanced flow $\mathbf{w}$ ($\mathbf{w} \geq 0$, $\mathbf{Aw} = 0$) is a sequence $H_1, \ldots, H_J$ of (not necessarily distinct) simple directed cycles such that

$$\sum_{l=1}^{J} \mathbf{f}(H_l) = \mathbf{w} \qquad (21)$$

where $\mathbf{f}(H_l)$ denotes a flow vector that has a 1 in each position (i, j) corresponding to an arc of H_l, and zeroes in all other positions.

Note that (if $\mathbf{c}$ has no zero-cost cycles), $\mathbf{s}$ will always decompose into a sequence of distinct simple cycles. Moreover, if $\mathbf{u}$ is strictly triangular (which it is when $\mathbf{u}$ is proportional to $\mathbf{c}$), this decomposition is unique.

Consider two simple cycles, R and S in the decomposition of $\mathbf{s}$, having the property that they cross only at the break. The two cycles can be combined in a splicing operation that inserts one in a given arc of the other. As an example, consider Figure 16, which shows two such cycles. Cycle S can be inserted into arc (i, j) to create the combined cycle shown in Figure 17 to create a substantial cost saving. In this example, the decrease in tractor costs is

$$\Delta Z = c_{0a} + c_{ij} + c_{b0} - c_{ia} - c_{bj} \qquad (22)$$

which, in general, may be positive, negative, or zero.

Note that there are a variety of different splicing opportunities because the roles of R and S may be interchanged, and the insertion arc (i, j) may be varied. However, if the cycles are both short, there will only be a handful of possibilities, some of which will produce savings, whereas others will not.

Note also that this method is applicable to cases in which R and S share some arcs (but there is a slack of at least 2 on all shared arcs). Essentially, each nonshared part of one cycle must be spliced into some nonshared arc of the other, and the number of tractors on all shared arcs is reduced by 1. For example, consider the problem shown in Figure 18. The LP solution (ignoring node-activity constraints) is shown in Figure 19, and the corresponding slack is shown in Figure 20. This slack can be broken down into the cycles S = 0-1-0 and R = 0-2-1-0.

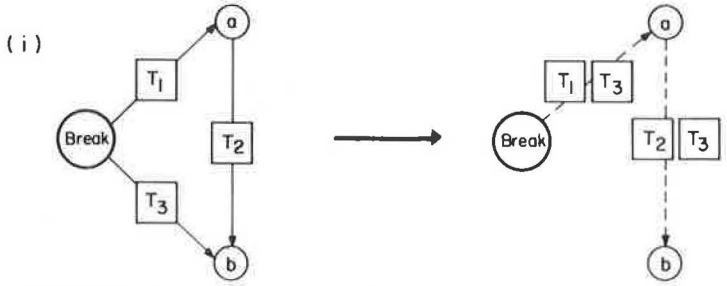

FIGURE 13 Local improvement for pattern (i).

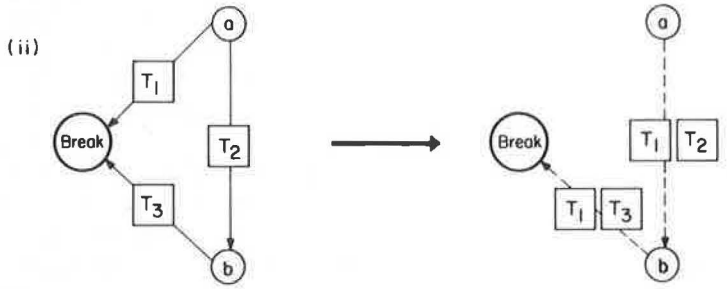

FIGURE 14 Local improvement for pattern (ii).

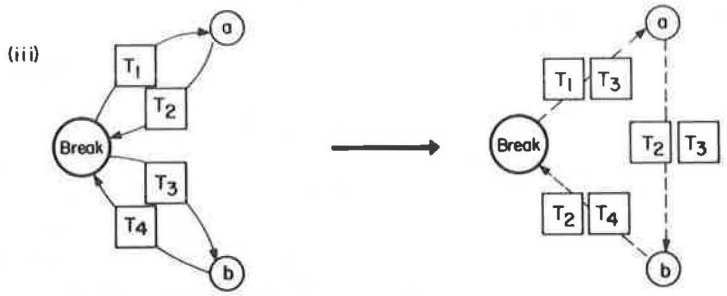

FIGURE 15 Local improvement for pattern (iii).

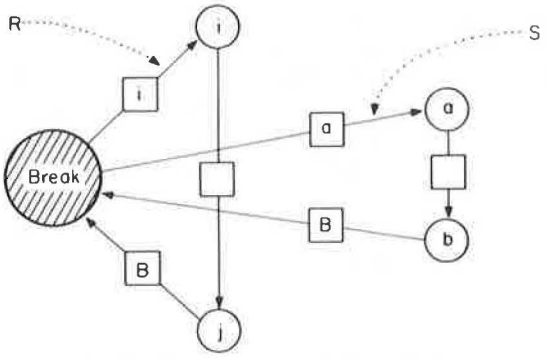

FIGURE 16 Two cycles ready for splicing.

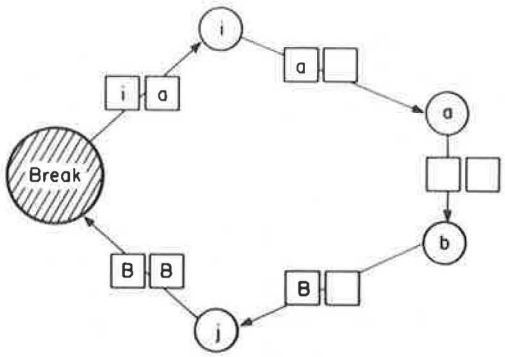

FIGURE 17 Outcome of the splicing operation.

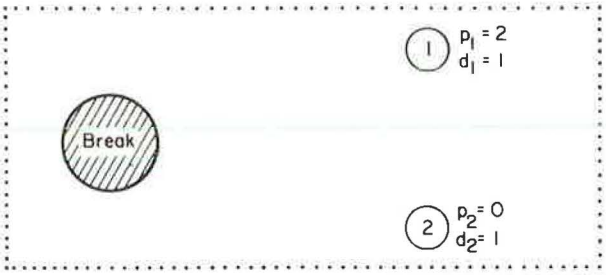

FIGURE 18 Splicing with shared arcs—problem data.

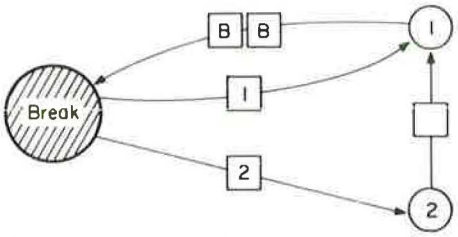

FIGURE 19 Shared-arc splicing—optimum trailer flows.

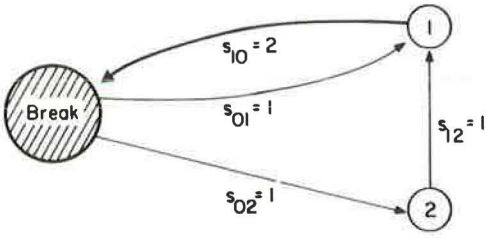

FIGURE 20 Slacks for shared-arc splicing example.

Splicing the nonshared part of R, 0-2-1, into the nonshared arc (0, 1) of S, the IP optimum shown in Figure 21 is obtained.

The cycle-splicing method works as follows:

1. Given **x, y,** and **e** optimal in $L(\mathbf{u})$, it solves the constrained roundup problem (Equations 11) and computes the resulting slack $\mathbf{s}$.
2. It decomposes **s** into a sum of simple cycles.
3. For each pair of such cycles, it examines all the possible ways of splicing them together. For each pair, the heuristic identifies the most attractive splicing opportunity and records it in a list.
4. It performs the recorded splicing operations in a greedy manner, starting with the one with the greatest savings and proceeding to the next most profitable one that is still allowable until the list is exhausted.

Both the simple local improvement heuristic and the cycle-splicing heuristic take as input the **x, y,** and **e** arising from a solution of $L(\mathbf{u})$. They then attempt to juggle some of the routings in order to reduce the cost of covering these trailer movements with tractors. Neither heuristic ever really changes some fundamental aspects of the incoming solution, in particular the assignment of empty trailers. If some EOL terminal i receives a given number of empties from EOL j in the solution

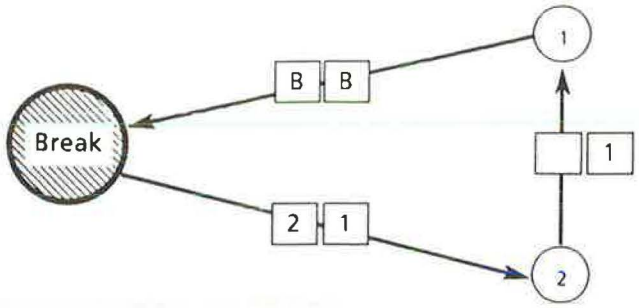

FIGURE 21 Outcome of shared-arc splice (also IP optimum).

to $L(\mathbf{u})$, it will still do so in the perturbed solution output from either heuristic. The only difference will be that certain detours may be inserted in paths taken by these trailers.

In practice, the cycle-splicing heuristic works very well, as shown by Eckstein *(9)*; under certain general conditions the cycles that arise in the decomposition of **s** are always short, allowing for efficient implementation. It is, however, somewhat slow, and implementing it for every subproblem actually slows down the overall run time of the Lagrangian B&B procedure (as compared with using Method 1). A more efficient approach is to use the cycle-splicing heuristic just once, at the very beginning of the algorithm, and the much faster simple local improvement heuristic (Method 1) at every iteration of every subsequent subproblem processed. In this way, one gets the advantage of a good initial incumbent without the computational burden of running the cycle-splicing heuristic repetitively.

LAGRANGIAN B&B PROCEDURE

As mentioned in the first section, the Lagrangian relaxation and the heuristic incumbent generation methods are used within a B&B procedure. Figure 22 shows a general flowchart of the calculation involved in Lagrangian-based B&B methods following Fisher *(6)*. A description has been given of how a **u** vector is chosen for each subproblem (block 2 in Figure 22), how $L(\mathbf{u})$ is calculated for a given **u** (block 3), and how the dual iterations are performed (block 8) and when they are terminated (block 7). The generation of a new incumbent solution (block 5) was described in the previous section. Here, some of the implementation details of the B&B procedure are given.

At each point in the solution procedure the B&B procedure deals with one subproblem (a point on the enumeration tree), that is, the original integer program with some added separation constraints. An active subproblem is one that has not been further separated and is thus an end point of the tree. At any given time, each subproblem q has some best lower bound β_q on its LP-relaxed objective value. This lower bound is the highest value of $L(\mathbf{u})$ found for the subproblem, or any of its ancestors, for all the **u**'s tried so far. The lowest value the global integer optimum could possibly have is

$$\beta \equiv \text{Min } \{\beta_q | q \text{ an active subproblem}\} \tag{23}$$

The B&B method tries to increase β and decrease the upper-

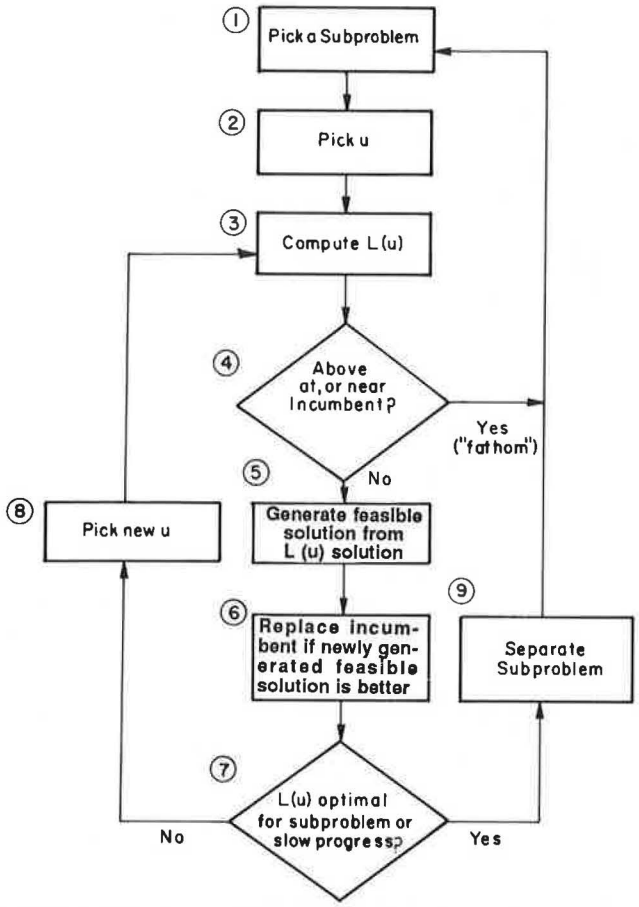

FIGURE 22 General Lagranglan branch and bound.

bound Z_{INC} (the objective value for the incumbent) until they are near enough to conclude that Z_{INC} is acceptably close to optimality.

Fathoming

A subproblem is fathomed (removed from further consideration) if a lower bound for its objective value provided by an $L(\mathbf{u})$ computation is within P percent of the upper bound on the global optimum provided by the incumbent. Such a strategy guarantees that the final solution is within P percent of optimality. The fathoming condition is the following:

$$\lceil L^*\mathbf{u}) \rceil \geq \left(1 - \frac{P}{100}\right) Z_{INC} \tag{24}$$

The user-specified parameter P can be used to explicitly trade off between the accuracy of the solution and the speed of obtaining it.

Separation

If slow improvement is detected in solving the Lagrangian relaxation of a given subproblem, it is split in two. In practice, it seems most efficient to split on the tractor variables t_{ij} rather than on the trailer variables x_{ij}, y_{ij}, or e_{ij}. Each split requires that t_{ij} be less than or equal to some integer m in one offspring subproblem and greater than or equal to $m + 1$ in the other.

Good computational experience was obtained by splitting on the t_{ij} for the longest arc (i, j) for which the number of trailers in the current $L(\mathbf{u})$ solution is odd. Thus the following constraints are added to the offspring subproblem:

$$t_{ij} \leq m \tag{25a}$$

$$t_{ij} \geq m + 1 \tag{25b}$$

The value of m is chosen to be $\lfloor x_{ij} + y_{ij} + e_{ij} \rfloor$ unless t_{ij} has been otherwise constrained ($\lfloor \cdot \rfloor$ denotes the floor or integer round-down function).

The separation scheme used in practice is somewhat more complicated; a "scoring" method is used for choosing the splitting arc. Each arc with odd trailer flow is assigned a score based on its length, whether it emanates from or terminates at the break, and whether it was separated on before. The algorithm then splits on the arc with the highest score. This method provided only about a 10 percent run time improvement over the simpler "longest odd arc" method mentioned before.

For the "lower" ($t_{ij} \leq m$) offspring, where at most m tractors are allowed on some arc (i, j), it is clear that there can be at most $2m$ trailers of each kind on (i, j). Thus the following redundant constraints can be enforced:

$$x_{ij} \leq 2m \tag{26a}$$

$$y_{ij} \leq 2m \tag{26b}$$

$$e_{ij} \leq 2m \tag{26c}$$

These simple upper bounds strengthen the Lagrangian relaxation and can be added without breaking network structure.

Subproblem Selection

After the algorithm has fathomed or separated a subproblem, it is faced with the decision of which subproblem to try next (unless there are none left unfathomed, in which case it terminates). The procedure uses here a simple rule of processing the problem, q, with the lowest βq.

The algorithm also includes a feature that allows it to switch to high-β_p subproblems in the event that there are so many active subproblems that the program is close to exhausting its virtual memory allocation. The intent is that these subproblems may be fathomed relatively quickly, freeing up memory for the more important ones. This feature allows the algorithm to handle larger problems with a given amount of memory, albeit with some speed penalty.

Basis Preservation

To improve run times the procedure uses information from earlier network simplex bases to speed up calls to the network simplex code. When $L(\mathbf{u})$ is computed, four network simplex

optimizations must be performed. At each subgradient iteration, the previous four optimal bases for the subproblem are used as the four starting bases. Because the optimal bases for two consecutive values of **u** should resemble one another, fewer pivots may be needed than if all initial bases are constructed from scratch.

This idea is carried one step further: when $L(\mathbf{u})$ is computed first for a subproblem, the procedure starts essentially with the four bases that were optimal in the last iteration of its parent. The slight difficulty here is that the added separation constraint may make one or more of the old optimal bases infeasible for the offspring. To see how this is handled, assume, for example, that a constraint $t_{ij} \leq m$ was added to a parent problem in which $t_{ij} = r > m$, rendering the parent basis infeasible in the offspring. Now, all the network representations contain a "super transshipment" node connected to all other nodes by "artificial" arcs of very high ("big M") cost. To maintain feasibility in the child, the flows are perturbed as shown in Figure 23.

By also setting the maximum flow capacity of the two artificial arcs to $r - m$, the procedure avoids having to add them to the basis (although either of them could already be in the basis in a degenerate manner), and the original spanning tree of the parent basis remains valid in the subproblem's perturbed network. Of course, when the network simplex routine is called, the artificial arcs will immediately have flow removed from them, because they have such high costs. Analogous techniques can be used for the addition of constraints of the form $t_{ij} \leq m + 1$, and also for the **x**, **y**, and **e** bases.

Old basis information could have been retained without perturbations by using a dual method to reoptimize the offspring of a subproblem, as in standard B&B methods, but this would have required the implementation of both primal and dual network simplex algorithms.

The drawback to using basis-preservation methods is that information must be stored for all active subproblems, increasing program memory requirements. This can be alleviated by using a depth-first tree exploration strategy in which the parent basis for one of the offspring of a subproblem is used without having to allocate any more memory, as long as that one offspring is analyzed immediately after separation.

COMPUTATIONAL PERFORMANCE

The Lagrangian B&B procedure described earlier was implemented on a VAX 11/780 minicomputer using VAX FORTRAN. (The computational power of this machine, at least for compute-bound numerical tasks, is now roughly matched by some of the recently released work stations and "high-end" personal computers.)

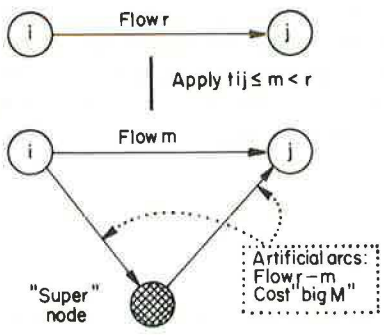

FIGURE 23 Basis modification for a violated upper bound.

The code was run on three sets of 20 randomly generated problems, one set with 4 EOL terminals per instance, one with 8, and one with 16. For these problems, EOL terminals were uniformly distributed over the interstices of a 40-by-40 grid centered on the break, with all p_i's and d_i's drawn from the probability mass function (PMF)

$$p(x) = \begin{cases} 0.4, & x = 1 \\ 0.4, & x = 2 \\ 0.1, & x = 3 \\ 0.1, & x = 4 \end{cases} \tag{27}$$

This PMF resembles values that might arise in practice. Distances were Euclidean, rounded up to integer values, with small adjustments and "stop-off" costs added to ensure that the triangle inequality was always strictly met.

Each of the 4-, 8-, and 16-node problem sets was run at varying levels of the percentage-fathoming parameter P. For budgetary reasons, each individual problem run was limited to roughly 5 megabytes of virtual memory and 15 min of CPU time. In summary, results were excellent for the 4-node problems, acceptable for 8 nodes, and somewhat disappointing for the 16-node examples.

Table 1 gives the results for selected combinations of numbers of EOL terminals and levels of P. It should be noted that runs were simply cut off if the available time or memory limits had to be exceeded. The method is quite memory-intensive, because basis information is retained from subproblem to subproblem. On the 8- and 16-node problems, upon reaching about 300 active subproblems, the algorithm would often switch to the high-β_q mode in which it attempted to fathom unpromising subproblems first in an effort to free up space. This procedure tended to slow down convergence and was sometimes unsuccessful in holding down memory requirements (when high-β_q subproblems could not be fathomed), in which case the program simply halted upon reaching its 5-megabyte storage limit.

As the results given in Table 1 suggest, the B&B procedure can easily solve small problems but does not perform well for large ones. In particular, the 16-node runs produce a duality gap only marginally smaller than that one would get by simply running the cycle-splicing heuristic and the LP relaxation and then comparing the two. None of the twenty 16-node cases run showed any improvement in the incumbent after the first iteration of the first subproblem, and the improvement in the lower bound attributable to enumeration was not very great. Thus, the best course in practice is to use the cycle-splicing heuristic as a stand-alone procedure and to optionally employ the rest of the algorithm as a means of assessing nearness to optimality.

In a detailed analysis of the smaller problems, there appeared to be a strong dependence of run time on the initial value of **u**. On average, the best results were obtained with $\mathbf{u} = \mathbf{c}/4$. However, for individual problems, different values would sometimes work better. This phenomenon, once understood, could perhaps be exploited to improve the updating of **u**.

TABLE 1 PERFORMANCE OF LAGRANGIAN B&B METHOD

No. of EOL Terminals	Fathoming Percentage P	Run Time (sec)		Duality Gap (%)	
		Mean	Median	Mean	Median
4	0	16.7	4.3	0.1	0.0
4	5	5.6	0.9	5.0	5.0
8	0	439.0	348.9	1.6	0.0
8	5	166.7	32.9	5.2	5.0
16	5	900+	900+	15.8	16.5

CONCLUSIONS

A method for optimizing tractor-trailer and twin-trailer movements in a group line-haul operation has been described. The method is based on a B&B procedure in which the lower bounds are calculated by a Lagrangian relaxation. The first incumbent is generated by a cycle-splicing heuristic and the incumbents at every subproblem are generated by a simple local improvement heuristic.

For small problems the B&B procedure worked well, but for larger ones the cycle-splicing heuristic should be used as a stand-alone method. The shortcomings of the B&B procedure may be alleviated by improving the dual step and thus setting the multiplier variables **u** in a more efficient way.

Other solution methods for this problem may be based on extensions of the cycle-splicing heuristics in which a matching problem is solved in order to decide on the best splicing combinations *(9)*. Alternatively one can think of a set covering formulations based on tractor tours *(10, 11)*.

ACKNOWLEDGMENT

Several individuals from Consolidated Freightways, Inc., helped in the understanding of group line-haul operations and the problems involved. They include Lee Frazee, Robert McDonald, Philip Seeley, Robert Blackburn, Toby Asarese, and Al Delara. In addition, Gene Hughs and Tom Meyers of United Parcel Service helped in the understanding of the (somewhat different) UPS twin-trailer matching problem. Thanks are expressed to all these individuals.

REFERENCES

1. Y. Sheffi and W. Powell. *Interactive Optimization for LTL Network Design*. Center for Transportation Studies Report 85-17. Massachusetts Institute of Technology, Cambridge, 1985.
2. T. L. Magnanti. Combinatorial Optimization and Vehicle Fleet Planning: Perspectives and Prospects. *Networks*, Vol. 11, 1987, pp. 179–213.
3. L. Bodin et al. Routing and Scheduling of Vehicles and Crews: The State of the Art. *Computers and Operations Research*, Vol. 10, 1983, pp. 63–211.
4. A. A. Assad. *Modeling Rail Freight Management*. Ph.D. thesis. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 1978.
5. B. Gavish and S. C. Graves. *The Traveling Salesman and Related Problems*. Working Paper OR 078-78. Operations Research Center, Massachusetts Institute of Technology, Cambridge, 1978.
6. M. L. Fisher. The Lagrangian Relaxation for Solving Integer Programming Problems. *Management Science*, Vol. 27, 1981, pp. 1–18.
7. M. L. Fisher. An Applications Oriented Guide to Lagrangian Relaxation. *Interfaces*, Vol. 15, No. 2, 1985, pp. 10–21.
8. B. L. Golden and T. L. Magnanti. Course Notes, Course 15.082. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, 1985.
9. J. Eckstein. *Routing Methods for Twin-Trailer Trucks*. Master's thesis. Operations Research Center, Massachusetts Institute of Technology, Cambridge, 1986.
10. M. L. Balinksi and R. E. Quandt. On an Integer Program for a Delivery Problem. *Operations Research*, Vol. 12, 1964, pp. 300–304.
11. F. H. Cullen, J. J. Travis, and H. D. Ratliff. Set Partitioning Heuristics for Interactive Optimization. *Networks*, Vol. 11, 1981, pp. 125–143.

Application and Testing of the Diagonalization Algorithm for the Evaluation of Truck-Related Highway Improvements

HANI S. MAHMASSANI, KYRIACOS C. MOUSKOS, AND C. MICHAEL WALTON

Highway and transportation officials are increasingly concerned about accommodating rising truck traffic and the associated size and weight trends. A network traffic assignment procedure is an essential component of the methodological support for the identification, evaluation, and selection of truck-related physical and operational improvements in a highway system. A general mechanism is presented for the network representation of improvements consisting not only of physical capacity expansion but also corresponding operational strategies in the form of (existing or new) lane-access restrictions to either vehicle class; this mechanism allows the consideration, as a special case, of exclusive truck lanes or facilities contemplated by several agencies. The special requirements of the traffic assignment procedure in this context, including the need to explicitly consider the asymmetric interaction between cars and trucks, give rise to potentially serious methodological difficulties that must be addressed for specific types of applications. The applicability of the diagonalization algorithm to such problems is investigated by using numerical experiments on three test networks under varying conditions. The three test networks include an abstracted condensed representation as well as a full-scale version of the Texas highway network, thus providing a realistic case application. The main aspects of the algorithm's performance addressed in these experiments are its convergence characteristics as well as the effectiveness of some computational streamlining strategies. Although convergence is not guaranteed a priori, it was actually achieved in all test cases. Furthermore, it is shown that shortcut strategies can considerably reduce the algorithm's computational requirements.

During the past 30 years, there has been a considerable increase in the size of the fleet of passenger cars and trucks, with an increasingly diverse mix of vehicles in the traffic stream. Different types of vehicles are entering the highway system, and they have different physical and performance characteristics. Recent trends toward less stringent regulations have allowed larger and heaver trucks in the highway system, jeopardizing geometric and capacity considerations in some parts of the system and resulting in increased pavement deterioration. Furthermore the interaction of vehicles with different sizes and performance characteristics, such as large combination trucks, on the one hand, and subcompact passenger cars, on the other, may have resulted in more hazardous driving conditions, with increased potential severity of collisions.

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

The foregoing concerns have led the appropriate agencies to consider improvements in the highway infrastructure, such as the construction of exclusive facilities for different classes of users, as well as operational measures involving the restriction of access to existing selected lanes for certain vehicle types. The current work was motivated by the need for a network-modeling methodology to support the design and evaluation of such truck-related improvements in a highway system *(1)*. A central component in this methodology is the assignment of traffic flows to various parts of the network in response to contemplated improvements. These flows are essential in the calculation of costs and benefits incurred with a particular set of improvements. Specifically, the function of the network traffic assignment procedure is to determine the link flow patterns resulting from the allocation of origin-destination trip matrices (for cars and trucks, respectively) corresponding to present or future conditions to a given highway network. By changing the configuration of the latter or by modifying the characteristics of some of its links, the traffic assignment procedure allows the assessment of the impact in all parts of the network, and on various user groups, of selected truck-related improvements, such as special truck lanes or facilities in designated corridors or highway sections.

To be useful in the evaluation of truck-related improvements, which arise in urban as well as intercity contexts, the network traffic assignment model should (a) yield separate estimates of truck flows and passenger-car flows on every network link, (b) capture the nonlinear dependence of the travel time (or cost) incurred by link users on the total flow using that link, (c) recognize the interaction between vehicle classes (cars and trucks) sharing the same right-of-way, and (d) be policy sensitive in that it should allow the representation of the contemplated truck-related countermeasures, particularly because these are not limited to the construction of additional facilities but also include operational strategies affecting both existing and proposed facilities.

Items b and c are generally captured in the link performance functions, which yield the travel time as a function of the flow of each vehicle class on that link, or more generally as a function of the entire link flow pattern. The resulting network user equilibrium problem requires the simultaneous solution of the link flows (which determine link travel times) and the link travel times (which in turn affect the routes chosen by motorists in the network). Furthermore, when interactions among vehicle

classes are explicitly represented, they are effectively equivalent to interactions among links (where different conceptual links are defined for each user class, as discussed in the next section). Such interactions are asymmetric, meaning that the marginal contribution of a vehicle belonging to a given category to the travel time of other classes is different from the marginal contribution of a vehicle in the latter category to the former's travel time. For this type of problem, interactions between cars and trucks are generally asymmetric, thereby giving rise to a user equilibrium network assignment problem with asymmetric link interactions.

A more detailed presentation of the mathematical and algorithmic background of this problem as it arises in this context and of its solution approach can be found in the report by Mouskos et al. *(2)*. Essentially, there are no guaranteed procedures to solve the network equilibrium problem with asymmetric interactions. However, an approach known as the diagonalization algorithm has emerged *(3–5)* as a promising one to solve for such general equilibrium problems. Other methods, particularly linearization methods [of which the projection method proposed by Dafermos *(6)* is a special case] and simplicial decomposition *(7)*, have also been suggested. By and large, the diagonalization approach is the most easily accessible to practitioners and researchers, because it can be implemented with relatively simple modification of widely available packages for the Frank-Wolfe (F-W) solution algorithm for the standard single-class network equilibrium problem.

The necessary conditions for the diagonalization algorithm's convergence to the desired equilibrium solution are not well understood in the transportation science literature. However, known sufficient conditions *(3)* are recognized as being too strict and often far from necessary *(5)*. Therefore, it is necessary to test the approach in the specific context in which it is to be employed. Furthermore, in its complete version, the algorithm is rather demanding computationally. Fortunately, some shortcuts have been suggested to improve its performance in this regard *(3)*. However, these approaches remain to be tested, because numerical experience to date appears to have been limited to small unrealistic networks. A major objective of this effort is therefore to test these shortcut strategies and develop computational experience in realistic networks in order to assess their usefulness as operational tools in the analysis of truck-related improvements in a highway network.

The principal objective of this work is thus to assess the practical applicability of the diagonalization algorithm for the evaluation of truck-related improvements. This involves two principal tasks: (a) the development of a mechanism for representing the improvements of interest and (b) testing the performance of the diagonalization algorithm for this type of application. Two questions are of concern in this regard: (a) convergence of the algorithm, which, as noted earlier, is not guaranteed, and (b) streamlining or shortcut strategies in the implementation of the algorithm in order to reduce its otherwise heavy computational cost.

This paper is organized as follows. In the next section, the improvement options are defined and their network representation is described. This is followed by a brief presentation of the steps of the algorithm and the shortcut strategies. Next, the numerical experiments conducted to examine the convergence of the algorithm and to compare the effectiveness of the various shortcut strategies are presented. The results of these experiments are then summarized, followed by concluding comments.

NETWORK REPRESENTATION OF IMPROVEMENT OPTIONS

As noted previously, each highway link is used by two classes of vehicles, cars and heavy trucks, which interact, through their use of the common shared right-of-way, in determining the travel time incurred by vehicles of both classes using that link. Naturally, it is possible to further disaggregate the heavy vehicles into various vehicle categories, but the dichotomy between cars and trucks is sufficient for most purposes. The interaction between the two classes on a highway link is represented through the use of identical networks (referred to as "copies" of each other) for each class. Each physical highway link is thus decomposed into two "conceptual" links. Each link has its own performance (or travel cost) function, and the flow on any given link consists of one or the other designated class only. Interaction among the various classes using a particular physical link thus translates into interaction among links in this network representation.

In order to test the effect of truck-related improvements to a particular highway link, it is necessary to devise a general-purpose mechanism that allows the representation of this improvement not only in terms of lane addition, but also in terms of how this new lane might be operated in conjunction with the existing lanes (e.g., access restriction of a given lane to certain vehicle classes). In particular, the new lane can be used by trucks only (exclusive truck lane), passenger cars only (restricted lane), or all traffic. Similarly, the existing lanes can be used by either all traffic or car traffic only. The following four types of improvement options, defined in terms of different mutually exclusive combinations of these factors, are of particular interest in this study:

Option 1: Expand the link by one lane and allow all traffic on entire link.

Option 2: Expand the link by one lane but allow only truck traffic on new lane with all traffic allowed on old lanes.

Option 3: Expand the link by one lane, but all truck traffic must use new lane with all car traffic allowed on old lanes only.

Option 4: Expand the link by one lane, but allow only car traffic on new lane with all traffic allowed on old lanes. Note that this option is equivalent to building a new lane that would be open to both cars and trucks and at the same time restricting trucks from using the left-most lane.

Figure 1 shows the implementation of each of the four options on a link that currently has three lanes.

The general mechanism for representing the foregoing improvements is as follows. With each given physical highway section, with start node i and end node j, it has been shown that two conceptual link copies are defined, one for trucks and one for cars, with the "coupling" accomplished through a special numbering scheme as well as through the respective link performance functions, as described by Mouskos et al. *(2)*. One additional node (dummy node) and two additional links are

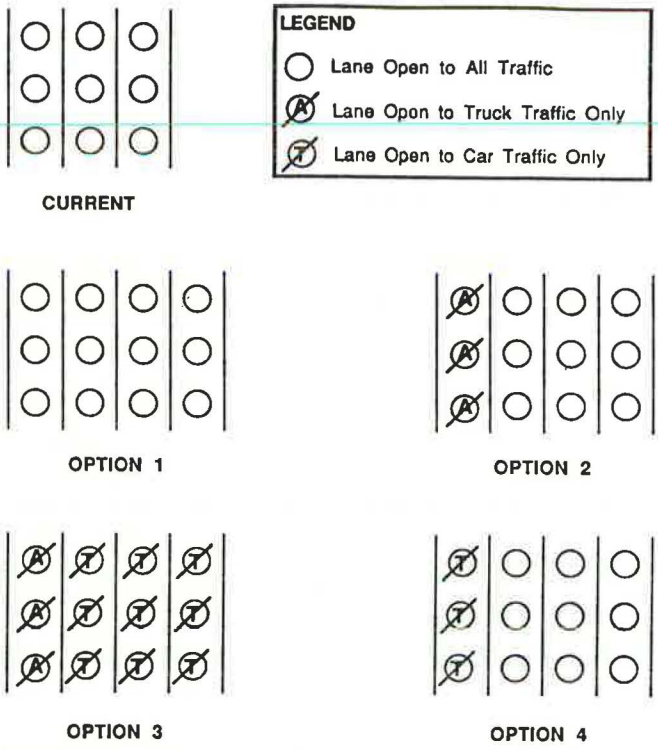

FIGURE 1 Link improvement options.

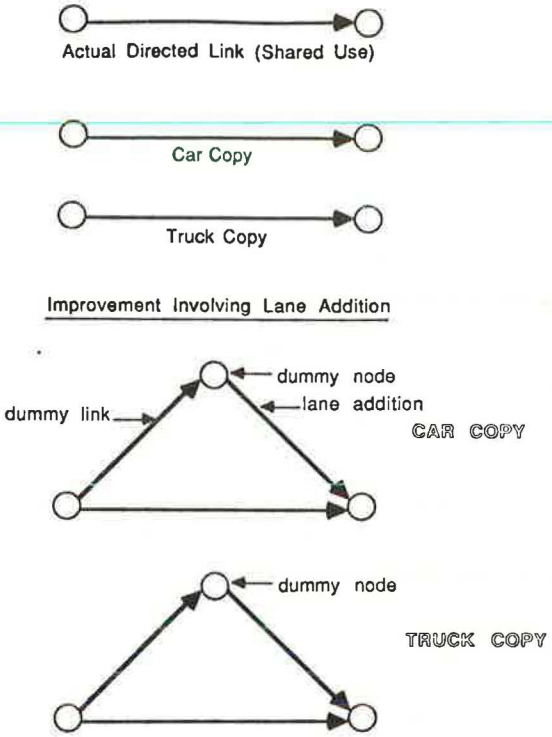

FIGURE 2 Special network representation of truck-lane addition.

also defined for each of the car and truck copies, as shown in Figure 2. The additional links are a dummy link from the start node i to the dummy node and a link from the dummy node to the end node j that represents the actual lane addition, which is included in each copy to allow either cars or trucks (or both) to use it. However, if it is desired to restrict its use to trucks only, the preceding dummy link in the car network copy will be associated with a very large positive cost (travel time), which effectively prohibits cars from using it (and thereby from getting onto the added lane). If, on the other hand, it is desired to allow cars to use this new lane, the cost is set equal to zero, and cars are therefore allowed to consider using the additional lane in their route choice, because no penalty is associated with the dummy link. Of course, the travel time that they will experience on that link (representing the new lane) is given by the associated performance function, which properly accounts for the interaction with the existing lanes as well as with the truck flows on the existing and new lanes (represented by the links defining the truck network copy).

To illustrate the foregoing mechanism, its application to the four improvement options of interest is described next. Consider existing directed link a and let a_A denote the copy of link a for car traffic; a_T, the copy of link a for truck traffic; a_A', the potential lane addition link for cars; a_T', the potential lane addition link for trucks; and a_{Ad} and a_{Td}, the corresponding dummy links in the car and truck copies, respectively.

Furthermore, define the link travel times t_{aA}, t_{aT}, $t_{a'A}$, $t_{a'T}$, and t_{aTd} on the corresponding links (see Figure 3). The travel times are, in the general case, related to the flow vector $\mathbf{X}_a = \{\mathbf{X}_{aA}, \mathbf{X}_{aT}, \mathbf{X}_{a'A}, \mathbf{X}_{a'T}\}$ comprising the respective flows on the just-defined links. Note that t_{aAd} and t_{aTd} are equal to either M (a

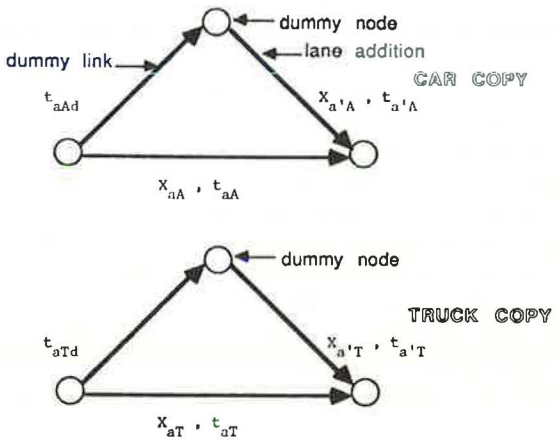

FIGURE 3 Notation for link improvement representation.

very large positive number) or 0, depending on the access rule for the additional lane, as seen hereafter. The performance functions for the nondummy links are denoted by $t_{aA}(.)$, $t_{a'A}(.)$, $t_{aT}(.)$, and $t_{a'T}(.)$, respectively. The operational schemes associated with each improvement option are translated through the specific dependence of the above link travel times on the components of the flow vector $\mathbf{X}_a$.

Option 1

Under Option 1, all traffic is allowed on the new lane, with no changes (i.e., still all traffic) on the existing lanes. Therefore,

$$t_{aA} = t_{a'A} = t^*_{aA}(X_{aA} + X_{a'A}, X_{aT} + X_{a'T})$$
$$t_{aT} = t_{a'T} = t^*_{aT}(X_{aA} + X_{a'A}, X_{aT} + X_{a'T})$$
$$t_{aAd} = t_{aTd} = 0$$

In other words, the average travel time on links a_A and a_A' is effectively the same (i.e., there is no basis for distinguishing between the performance of these lanes) and is dependent on the total automobile and total truck flows, respectively, on the upgraded highway link a; similarly for a_T and a_T'. The functions $t^*_{aA}(.)$ and $t^*_{aT}(.)$ denote the modified performance functions for the upgraded facility.

Option 2

Under Option 2, the new lane is an exclusive truck lane, with no car traffic allowed on that lane. No other restrictions apply on the existing lanes. This translates into

$$t_{aA} = t_{aA}(X_{aA}, X_{aT})$$
$$t_{a'A} = t_{aAd} = M \text{ (very large positive number)}$$

$$t_{aT} = t_{aT}(X_{aA}, X_{aT})$$
$$t_{a'T} = t_{a'T}(X_{a'T}) \quad t_{aTd} = 0$$

Effectively then, the new exclusive truck facility is assumed to operate virtually independently from the existing lanes. Therefore, travel time for trucks on the truck facility depends only on the flow of trucks using that facility. Naturally, travel time for cars on that facility (as well as on the corresponding dummy link) is set to a very large positive number to prohibit its use by cars.

Option 3

Under Option 3, the new lane operates as an exclusive truck facility. However, this is coupled with the restriction of truck traffic from using the other (existing) lanes of the highway. The corresponding relationships are

$$t_{aA} = t_{aA}(X_{aA}) \quad t_{aT} = 0$$
$$t_{a'A} = t_{aAd} = M$$
$$t_{a'T} = t_{a'T}(X_{a'T}) \quad t_{aTd} = 0$$

Here, the two types of facilities (existing lanes and new exclusive truck lane) operate virtually independently from one another, thus the absence of cross-link effects in the corresponding performance functions.

Option 4

Under Option 4, an exclusive new car-only lane is built; no other restrictions apply. This is represented by

$$t_{aA} = t_{aA}(X_{aA}, X_{aT})$$
$$t_{a'A} = t_{a'A}(X_{a'A}) \quad t_{aAd} = 0$$
$$t_{aT} = t_{aT}(X_{aA}, X_{aT})$$
$$t_{a'T} = t_{aTd} = M$$

In essence, under Option 4, the new exclusive car lane is assumed to operate virtually independently of the existing lanes, with no truck interference, whereas the same relationships remain in effect in the existing, shared-use lanes.

The foregoing equations illustrate the functional dependence between the various travel time and flow components associated with the particular network configuration introduced in this study to represent the truck-related link improvements of interest. The specific functional forms and parameter values of the link performance functions for different facility types should ideally be determined from actual field observations. The calibration of such functions was not within the scope of the present study. In the numerical experiments conducted to test the algorithm, performance functions of the well-known Bureau of Public Roads *(8)* form were modified to reflect the interaction between cars and trucks. For example, for a common link a shared by cars and trucks, the basic functional form is

$$t_{aA} = t^o_{aA}\ \{1 + ALPHA \cdot [(X_{aA} + h \cdot X_{aT})/C_a]^{BETA}\}$$

where

t^o_{aA} = free-flow time of passenger cars on link a,
$ALPHA, BETA$ = parameters capturing the sensitivity of travel time (cost) to flow,
C_a = parameter that captures the "capacity" of link a in passenger-car equivalents per unit time, and
h = parameter that transforms the effect of trucks into passenger-car equivalents.

The parameter values used are adapted from published values reported in the literature and ensure that travel cost increases monotonically with the flow components. In the absence of observationally developed performance functions that explicitly address car–truck interactions, it is believed that the previous well-known functional form and the standard traffic engineering approach for treating trucks are adequate for the purpose of implementing the assignment methodology and performing the desired tests of the diagonalization algorithm for this type of problem.

STEPS OF THE DIAGONALIZATION ALGORITHM

Before the results of the numerical tests are described, it is useful to present a conceptual overview of the diagonalization algorithm to define the shortcut or streamlining strategies tested

in this study. The difficulty with the network equilibrium problem with asymmetric link interactions is that, unlike problems with symmetric or no interactions, it does not admit an equivalent mathematical programming formulation that can be readily solved to obtain the desired equilibrium flow pattern. The diagonalization algorithm is a so-called "direct" algorithm that is not guided by the minimization of some global objective function in its attempt to converge on a solution. Proofs exist that it will only converge on a solution that satisfies the desired equilibrium conditions *(3, 9)*.

Essentially, the diagonalization algorithm is an iterative procedure that involves solving a series of tractable single-class user equilibrium programs. Letting $t_a^n = f_a(X_a^n, \mathbf{X}_{\bar{a}}^n)$ denote the performance function for link a, where $t_{\bar{a}}^n$ is the travel time and X_a^n is the flow on link a at the nth interaction of the algorithm, whereas $\mathbf{X}_a^n$ is a vector of flows on all links other than a that affect the travel time on that link. The vector $\mathbf{X}_{\bar{a}}^n$ will typically contain at least the flow on the link corresponding to the "other" vehicle class sharing the physical right-of-way of link a; that is, if a is the passenger-car copy of a particular highway link, $\mathbf{X}_{\bar{a}}^n$ will include at least the flow on the truck copy of that same physical highway link. In addition, $\mathbf{X}_{\bar{a}}^n$ may also contain flows on the links defined especially to study the impact of truck lanes and other truck-related improvement options, as described in the previous section.

Note that the effect of X_a on t_a is referred to as a "main" effect, whereas the effect of the components of $\mathbf{X}_{\bar{a}}$ on t_a are called "cross-link" effects. The diagonalization algorithm requires, at the nth iteration, that all cross-link effects be fixed at their current levels, with only the main effect allowed to vary in the solution of the equilibrium problem at any given iteration. In other words, at the nth iteration of the diagonalization algorithm, t_a^n and X_a^n are solved for jointly such that $t_a^n = f_a(X_a^n, \mathbf{X}_{\bar{a}}^{n-1})$ and user equilibrium conditions are satisfied assuming that the cross-link effects are fixed at their values from the $(n-1)$th iteration. The next iteration, if convergence is not yet achieved, will then fix the cross-link effects at their values from the nth iteration in solving for t_a^{n+1} and X_a^{n+1}.

The steps of the algorithm can be summarized as follows:

Step 0: Initialization; find a feasible link-flow vector.

Step 1: Diagonalization; at the nth iteration, solve a user equilibrium subproblem assuming that cross-link effects are fixed. This yields a link flow pattern X_a^n and associated link travel times t_a^n for all links a.

Step 2: Convergence test; if $X_a^n \equiv X_a^{n-1}$, for all a, then convergence is reached. X_a^n is the desired solution. Otherwise, set $n = n + 1$ and go back to Step 1.

As can be seen, the diagonalization algorithm involves a number of iterations, referred to as outer iterations to distinguish them from the inner iterations that must be performed in solving the subproblem in Step 1. This subproblem is solved by using the F-W or convex combinations algorithm, the steps of which can be summarized as follows:

Step 0: Initialization; perform all-or-nothing assignment (by assigning all flows from a given origin to a particular destination to the shortest path between the two points) based on the free-flow or uncongested link travel times.

Step 1: Update travel times by using the link performance functions; only the main effects due to each link's own flow are allowed to affect that link's travel time. Cross-link effects remain constant throughout the solution of this subproblem, fixed at their values from the previous outer iteration.

Step 2: Direction finding; after solving a shortest-path problem based on the updated travel times from Step 1, perform all-or-nothing assignment from each origin to all destinations.

Step 3: Line search; find optimal move size parameter to update the current link flows by combining them with the flows generated in the direction-finding step (Step 2).

Step 4: Update the link flows by using the move size parameter calculated in Step 3.

Step 5: Convergence test; if the link flows generated in Step 4 are about equal to the flows from the previous iteration, convergence is reached and the subproblem is solved; otherwise, increment the iteration counter and go back to Step 1.

A more detailed mathematical presentation of the algorithm has been presented by Sheffi *(3)* and by Mouskos et al. *(2)* for the problem context of interest here. The main point for this discussion is that at each (outer) iteration of the diagonalization algorithm, a number of inner iterations need to be performed. Each of these inner iterations requires the solution of a shortest-path problem, from each origin to all destinations, which can be quite demanding computationally. However, by noting that the flow pattern of only the last outer iteration needs to be determined accurately and that the accuracy of the solution to that subproblem improves only marginally with each additional inner iteration, a streamlining of the algorithm has been proposed by Sheffi *(3)* whereby only one inner iteration is performed per outer iteration. Although more outer iterations will be required to reach convergence, the total number of (inner) iterations involving shortest-path calculations will in most cases be less than what it would have been in the algorithm's original form (i.e., if the diagonalized subproblem is solved to convergence in every outer iteration). This streamlined version has not, however, been sufficiently tested on realistic networks to ascertain its relative computational efficiency compared with other possible streamlining strategies.

Sheffi's streamlining strategy is generalized here by considering a family of shortcut strategies, each characterized by a different value of the (maximum) number of inner iterations per outer iteration. Numerical tests are conducted for values ranging from 1 to 10 inner iterations, in order to determine whether there is a "best" value that appears to apply in most cases for the type of network application of interest. The test networks and experimental conditions considered in the numerical work are described in the next section.

In addition to testing the foregoing shortcut strategies, another objective is to establish that the algorithm does indeed converge for this type of application, because, as noted earlier, convergence is not guaranteed a priori. It can be noted in this regard that the known sufficient conditions for convergence require that the cross-link effects resulting from the interaction between the various user classes be relatively weak compared with the main effect. For the problem of interest, this condition would require that the marginal effect of trucks on the travel time experience by cars sharing the same facility be much smaller than the corresponding marginal effect of cars. Such a

requirement cannot be expected to hold for car–truck interactions and will therefore violate the foregoing sufficient conditions. For this reason, experience with the algorithm in this problem context is quite valuable in assessing its applicability.

COMPUTATIONAL TEST RESULTS

The Experiments

The performance of the diagonalization algorithm and the shortcut strategies was examined in several numerical experiments conducted on three different test network configurations. Network 1 is an arbitrary hypothetical network, whereas Networks 2 and 3 are developed from the Texas highway network. Network 2 is a condensed abstraction of the Texas highway system, developed as the principal network for methodological development and testing. It reflects the essential features of the Texas system in terms of the location of the major origins and destinations, as well as the principal transport arteries, with a minimum of unessential detail. The motivation for this condensed version is of course to limit cost and effort associated with methodological development yet to obtain sufficiently meaningful insights for the large-scale implementation of the algorithm. However, less intensive testing was also performed on Network 3, which is a full-scale representation of the Texas network. As seen from the following summary of characteristics of each network, Network 3 contains an order of magnitude more nodes and links than does Network 2:

	No. of OD Pairs	*No. of Nodes*	*No. of Links*
Network 1	220	78	202
Network 2	364	128	336
Network 3 (Texas network)	364	1,400	3,912

As explained in the previous section, the basic experiment performed is to apply the diagonalization algorithm with different constraints on the maximum number of inner iterations of the F-W algorithm in solving the diagonalized subproblems. The values tested included all integer values from 1 to 10 inner iterations (per outer iteration) for Networks 1 and 2, and from 1 to 5 for Network 3 (values greater than 5 were not run for this large network, primarily because of cost considerations).

These experiments were performed both for cases of "closed" improvement options and for "open" improvement options for Networks 2 and 3, and only for "closed" options for Network 1. Under the "closed" improvement options, and referring to the three-link representation of improvement options presented earlier, the additional-lane links were not available for either cars or trucks (i.e., a high-cost M is placed on the dummy access links for both car and truck copies). Under the "open"improvement options, it is assumed that exclusive lanes are available to both cars and trucks, respectively, in addition to the common existing lanes.

In addition, and for Network 2 only, the experiments were performed under four different congestion levels. The mechanism for controlling this factor is the parameter C_a in the link performance functions. With the origin-destination (OD) matrices unchanged, four levels of this parameter were tested on the configuration of Network 2: the reference, or actual, value C; $0.5C$; $0.8C$; and $4.0C$.

The results of these experiments are summarized next.

Summary of Results

For each test case, two principal figures of merit were considered: the total number of internal iterations until convergence and the corresponding CPU time, which is virtually perfectly correlated with the total number of iterations (for a given network configuration and capacity level).

Table 1 is a summary of the results for each network configuration under the experimental conditions described. It presents the ranking of each shortcut strategy based on its relative performance (with the best-ranked first) for the various test cases. In addition, in that same table, the difference between the total number of (inner) iterations required for convergence and the corresponding number for the optimal (minimum) strategy is given in parentheses next to the ranking. Table 2 presents the total number of inner iterations until overall convergence for Network 2 under the open and closed improvement options. Table 3 presents similar information for Network 3 (the full-scale Texas network).

The principal conclusion from these results is that considerable reduction in computational effort can be achieved by constraining the maximum number of inner iterations in solving each diagonalized subproblem. Table 1 indicates that no single strategy (i.e., value of the maximum number of inner iterations) consistently outperforms all others. However, a strong case can be made for not going beyond three inner iterations under any circumstance, because using one, two, or three inner iterations ranked in the top three in most cases. Of the nine tests conducted with each strategy on Networks 1 and 2, the two-iteration strategy performed best five times, the three-iteration strategy performed best twice, and using one and six iterations performed best once each. For the two tests conducted on the large Texas network (Network 3), the one-iteration strategy performed best when the improvement options were open, whereas the two-iteration strategy performed best when the improvement options were closed (Table 3). The latter is more representative of actual traffic conditions because the assumption of both car and truck exclusive lanes (associated with all physical highway segments) in the "options open" scenario is intended as an extreme case. It can also be noted from the results in Table 2 (for Network 2) that the best performance of the one-iteration strategy occurred exclusively under the "options open" scenario. Table 1 further indicates that the one-iteration strategy was ranked second six times, with a corresponding deviation from the "best" strategy ranging from one to eight total iterations.

When not ranked best, the deviation (in terms of total number of iterations) of the two-iteration strategy from the corresponding best strategy ranged from 1 to 26 iterations, whereas that of the three-iteration strategy ranged from 5 to 53. However, the upper bounds of these ranges of deviation (26 and 53, respectively) occurred in the same test case. The latter used Network 2 under extremely low congestion levels (options open and capacity C_a equal to 4 times the reference

TABLE 1 SUMMARY OF RESULTS; RANK-ORDER POSITION OF SHORTCUT STRATEGIES

Shortcut Strategy[a]	Network 1	Network 2								Network 3	
		Options Open				Options Closed				Options Open	Options Closed
		1.0C	0.8C	0.5C	4.0C	1.0C	0.8C	0.5C	4.0C		
1	1(0)	2(1)	2(2)	2(1)	7(56)	2(8)	2(8)	2(7)	3(6)	1(0)	2(7)
2	2(12)	1(0)	1(0)	4(8)	3(26)	1(0)	1(0)	1(0)	2(1)	2(7)	1(0)
3	3(21)	5(9)	3(5)	1(0)	5(53)	3(15)	3(9)	3(17)	1(0)	4(10)	3(17)
4	5(30)	8(13)	7(14)	3(6)	2(2)	4(25)	4(13)	4(18)	6(21)	3(8)	5(22)
5	4(29)	3(6)	6(13)	4(8)	6(53)	6(31)	5(25)	7(34)	4(8)	5(12)	4(20)
6	6(46)	5(9)	4(9)	5(9)	1(0)	8(36)	5(25)	10(51)	5(12)		
7	7(55)	7(11)	8(15)	7(13)	10(72)	7(34)	6(29)	6(25)	4(8)		
8	8(66)	6(10)	5(11)	7(13)	4(42)	5(26)	8(46)	5(24)	4(8)		
9	9(78)	9(15)	9(16)	6(11)	9(62)	7(34)	7(39)	8(42)	4(8)		
10	10(80)	4(8)	7(13)	7(13)	8(58)	9(46)	9(48)	9(45)	4(8)		

NOTE: Difference between total number of iterations required for convergence and corresponding number for the optimal strategy is given in parentheses.
[a]Maximum number of internal iterations per outer iteration.

TABLE 2 TOTAL NUMBER OF INTERNAL ITERATIONS: NETWORK 2

Shortcut Strategy[a]	Capacity Value			
	1.0C	0.8C	0.5C	4.0C
Options Open				
1	16	13	14	79
2	15	11	21	49
3	24	16	13	76
4	28	25	19	25
5	21	24	21	76
6	24	20	22	23
7	26	26	26	95
8	25	22	26	65
9	30	27	24	85
10	23	25	26	81
Options Closed				
1	21	25	24	10
2	13	17	17	5
3	28	26	34	4
4	38	30	35	25
5	44	42	51	12
6	49	42	68	16
7	47	46	42	12
8	39	63	41	12
9	47	52	59	12
10	59	65	62	12

[a]Maximum number of internal iterations per outer iteration.

TABLE 3 SUMMARY OF RESULTS FOR NETWORK 3

Shortcut Strategy[a]	Total No. of Internal Iterations Until Convergence	
	Options Open	Options Closed
1	16	32
2	23	25
3	26	42
4	24	47
5	28	45

[a]Maximum number of internal iterations per outer iteration.

value). Excluding that test case, the upper bounds of the deviation ranges become 17 and 25 iterations for the two- and three-iteration strategies, respectively.

As has been seen, Sheffi's one-iteration streamlining strategy is not necessarily nor most frequently the most efficient strategy. It is, however, ranked second best in many cases and definitely provides a considerable improvement over the complete version of the algorithm. In the problem context used here, it seems that the two-iteration strategy is the most frequent best performer, though one cannot guarantee a priori that it will outperform the others in any given application.

The analysis of the CPU time required for each test case indicated that, as expected, the cost per iteration will increase with the number of OD pairs and with the size (number of nodes and links) of the network. For instance, the average cost per iteration for Network 3 is 7.51 sec compared with 0.29 and 0.53 sec for Networks 1 and 2, respectively. Interestingly, although the cost per iteration is obviously higher for the large Texas network, the total number of iterations required for this network was not much different from that of its reduced abstracted version (Network 2), thereby validating the premise of using the smaller network as a laboratory for computational testing before addressing the full-scale network. This approach will be pursued for evaluating heuristics for a special version of the network design problem (to determine optimal truck-related improvements in the network), for which extensive testing on the large network is prohibitive.

Another result that is quite important to this study's objective is the fact that convergence was reached in all tests conducted for all three network configurations and experimental conditions considered. It is notable that convergence was achieved despite the fact that the known sufficiency condition was violated in this type of application, thereby confirming the applicability of the algorithm to problems involving asymmetric interactions between cars and trucks on highway links. However, it must be noted that uniqueness of the solutions cannot be guaranteed.

Further details on the results of these experiments, as well as on the application of the diagonalization algorithm for the system optimal assignment problem, can be found in reports by Mouskos et al. *(2)* and by Mahmassani and Mouskos *(10)*.

CONCLUSION

This paper has addressed the applicability of the diagonalization algorithm to the evaluation of truck-related link improvements in a highway network. This problem involves asymmetric interaction between cars and trucks sharing the right-of-way, posing potentially serious methodological questions that need to be resolved in the problem context of interest. The results were generally positive; they indicated that the algorithm converged in most situations even though it did not conform to the only known sufficient conditions for convergence, which appear to be too strict and far from necessary.

In addition, it was demonstrated that considerable savings can be achieved by using shortcut strategies in implementing the diagonalization algorithm. Sheffi's *(3)* streamlined version of the algorithm, which uses only one inner F-W iteration in solving the diagonalized subproblems, was shown to perform considerably better than the unmodified algorithm and some of the other strategies tested here. However, it was not uniformly the best streamlining strategy. In these tests, the two-iteration strategy (using two internal F-W iterations per outer iteration) performed best in the majority of cases; however, the one- and three-iteration strategies were often acceptable second- or third-best strategies. It is nevertheless difficult to determine a priori which strategy will perform best in a particular situation. Additional testing may be necessary if more specific practical guidelines are to be developed in this regard.

A general mechanism was introduced for the network representation of an array of truck-related highway improvements, consisting not only of physical capacity expansion, but also of operating strategies in the form of lane access restrictions for either cars or trucks. This representation allows the use of the assignment methodology to support the analysis, identification, and selection of specific sections of the highway network as candidates for the construction and implementation of particular types of truck-related improvements such as exclusive truck lanes or facilities, which are of great interest to transportation and highway agencies.

Naturally, it is highly desirable to develop link performance functions that explicitly capture the interactions between cars and trucks, based on actual observations of such traffic. Limited attempts along these lines have been described by Kim *(11)*. Finally, it should be noted that because the primary interest of the present work is to examine the operationa implications on the highway system, the OD patterns of both cars and trucks were assumed known and given. However, for studies with a strategic orientation, where the broader implications for the economy, carrier operations, or shipper logistics are of concern, more general formulations along the lines discussed by Friesz et al. *(12)* should be considered.

ACKNOWLEDGMENTS

The diagonalization code used in this study was modified from an F-W User Equilibrium program listing provided by Fred Mannering of Pennsylvania State University, who in turn modified a code supplied by Stella Defermos of Brown University. The collaboration of John J. Massimi, formerly a graduate assistant working with the study team, in defining the improvement options is acknowledged. The work presented in this paper was performed in conjunction with Research Project 356, Study of Truck Lane Needs, funded by the Texas State Department of Highways and Public Transportation. The authors, of course, remain solely responsible for the contents of this paper.

REFERENCES

1. H. S. Mahmassani, C. M. Walton, K. Mouskos, J. J. Massimi, and I. L. Levinton. *A Methodology for the Assessment of Truck Lane Needs in the Texas Highway Network.* CTR Report 356-3F. University of Texas at Austin, 1986.
2. K. Mouskos, H. S. Mahmassani, and C. M. Walton. *Network Assignment Methods for the Analysis of Truck-Related Highway Improvements.* CTR Report 356-2. University of Texas at Austin, 1986.
3. Y. Sheffi. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods.* Prentice-Hall, Englewood Cliffs, N.J., 1984.
4. S. C. Dafermos. Relaxation Algorithm for the General Asymmetric Traffic Equilibrium Problem. *Transportation Science*, Vol. 16, No. 2, 1982, pp. 231–240.
5. T. Friesz. Transportation Network Equilibrium, Design and Aggregation: Key Developments and Research Opportunities. *Transportation Research A*, Vol. 19A, No. 516, 1985, pp. 413–427.
6. S. C. Dafermos. An Iterative Scheme for Variational Inequalities. *Mathematical Programming*, Vol. 26, No. 1, 1983, pp. 40–47.
7. S. Lawphongpanich and D. W. Hearn. Simplicial Decomposition of the Asymmetric Traffic Assignment Problem. *Transportation Research B*, Vol. 18B, No. 2, 1984, pp. 123–133.
8. *Traffic Assignment Manual.* Bureau of Public Roads, U.S. Department of Commerce, 1984.
9. M. Abdulaal and L. J. Leblanc. Methods for Combining Modal Split and Equilibrium Assignment Models. *Transportation Science*, Vol. 13, No. 4, 1979, pp. 292–314.
10. H. S. Mahmassani and K. Mouskos. Some Numerical Results on the Diagonalization Algorithm for Network Assignment with Asymmetric Interactions Between Cars and Trucks. *Transportation Research B* (in press).
11. Y.-G. Kim. *Performance Functions for Highway Links with Asymmetric Interactions Between Cars and Trucks.* M.S. thesis. Department of Civil Engineering, University of Texas at Austin, 1986.
12. T. L. Friesz, R. L. Tobin, and P. T. Harker. Predictive Intercity Freight Network Models: The State of the Art. *Transportation Research A*, Vol. 17A, No. 6, 1983, pp. 409–417.

Link Performance Functions for Urban Freeways with Asymmetric Car–Truck Interactions

YOUNG GEOL KIM AND HANI S. MAHMASSANI

Link performance functions, which capture the relationship between travel time per unit distance and traffic volume per unit time on the links of a network, constitute an essential element in the equilibrium assignment of traffic flows to congested transportation networks. Results are presented of the empirical development and calibration of performance functions that capture the dependence of travel time on the respective volumes of passenger cars and trucks sharing the physical right-of-way on urban freeway sections. The data used for model calibration, individual vehicle trajectories on urban freeway sections, originally collected for FHWA, are developed from a secondary data base. Despite the data limitations, useful relations applicable to a broad range of freeway traffic conditions are developed that yield insights into the effect of trucks on freeway performance. Two types of functions are presented: (a) linear functions over the range of operating volumes extending up to 1,300 vehicles per hour per lane and (b) nonlinear functions, based on the widely used Bureau of Public Roads form, over the full range of flow values. A secondary analysis of the relation between truck and car average travel times is also discussed. These functions are intended for use in network equilibrium studies requiring the assignment of explicit car and truck flows and therefore involving asymmetric interactions between these vehicle classes (or equivalently between links). Such problems arise in the context of the evaluation of truck-related highway improvements, which is a problem of current interest to highway agencies.

Link performance functions constitute an essential element in the equilibrium assignment of traffic flows in congested transportation networks. These functions capture the relationship between travel time per unit distance and traffic volume per unit time on the links of a network. There has been considerable development in theoretical and algorithmic aspects of the network equilibrium problem over the past decade [state-of-the-art reviews have been published by Friesz *(1)* and Sheffi *(2)*]. Recent advances have addressed very general formulations that recognize the presence of multiple user classes interacting in their shared use of the physical right-of-way of the links and, more generally, where asymmetric interactions exist between the network's links. A special case of this problem is that in which separate passenger-car and truck flows must be assigned to a highway network.

The advances have not been accompanied by any significant research into the form and parameter values of the link performance functions, which are essential for the application and further theoretical development of network equilibrium models. The most comprehensive study published after 1975 was conducted by the research group of the University of Montreal, in conjunction with the application of EMME (a multimodal network equilibrium model), using data from the city of Winnipeg *(3)*. Prior to that, Branston's *(4)* work is probably the latest serious investigation of this subject. In neither of these studies, however, was the issue of interaction among vehicle classes addressed.

Judging by the effort invested over the past decade in the network equilibrium problem, it seems surprising that the question of properly specified and calibrated link performance functions has received so little attention in the community of transportation researchers and practitioners. To a large extent, the issue is an empirical one that can only be addressed by using observations of actual traffic behavior on highway links. The sheer size of the data and corresponding cost requirements for a systematic investigation of this problem have probably been serious hindrances.

One particular area in which network equilibrium applications lack any significant observationally calibrated link performance functions is that involving the assignment of cars and trucks, with the interaction resulting from the joint use of roadways by both classes explicitly captured in the link performance functions. Such interactions are asymmetric because the effect of an additional truck in the traffic stream on the average car travel time is different from that of a car on truck travel time. This problem arises, for example, in studies of truck-related improvements in highway networks, as described by Mahmassani et al. earlier *(5)* and in another paper in this Record. This paper presents a first step toward addressing this gap through the calibration of such functions with car and truck flows on freeway sections by using secondary data in the form of individual vehicle trajectories initially collected for an entirely different purpose *(6)*.

This paper is organized as follows. First, the pertinent conceptual background related to link performance functions and truck effects is given. Next, the data are described so that the inherent limitations for this purpose will be understood. In the fourth section, linear functions are presented for the range of operating volumes extending up to 1,300 vehicles per hour per lane to gain insight into the underlying traffic behavior. This is followed by the calibration results for a nonlinear function defined over the full range of flow values. A useful relation between truck and car travel times is then presented, followed by concluding comments.

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

BACKGROUND ON LINK PERFORMANCE FUNCTIONS

It is generally recognized that as the volume of traffic on a link increases, so does the average travel time along that link. One difficulty with performance functions for urban freeways is the complex traffic flow dynamics under heavily congested conditions, in which traffic operates in a highly unstable regime. In particular, the presence or absence of queues at freeway bottlenecks affects the travel time associated with an observed volume on a particular section and raises methodological issues in properly defining the observations for studying the relation of interest, as discussed recently by Hurdle and Solomon *(7)* and alluded to by Branston *(4)*. However, it is not the purpose of link performance functions, intended for use in traffic assignment applications, to capture the dynamic aspects of traffic flow on the facility. Traffic assignment models are essentially planning tools, concerned primarily with presumed steady-state conditions. This is an important consideration in defining the observations used for the calibration of these functions.

The dependent variable in the link performance functions of interest here is the average travel time incurred by vehicles traversing a particular link. The average travel time is actually the reciprocal of the space mean speed of the vehicles traveling over the highway section under consideration. This average travel time is related to the prevailing traffic volume per time unit [or "rate of flow" in *Highway Capacity Manual* (HCM) terminology *(8)*]. Of particular concern in this study are the differential effects of the respective car and heavy-truck (and heavy-vehicle) volumes using the link. In current practice, trucks and heavy vehicles are converted into passenger-car equivalents (pce's) using multipliers reported in the traffic engineering literature, particularly in the HCM *(8)*. Values reported in the HCM are intended to capture the amount of "capacity" taken up by a truck relative to a car, and thus do not necessarily ensure that a truck's impact on travel time (or speed) is correctly reflected. Furthermore, considerable debate exists in the traffic engineering community regarding the appropriateness of the 1985 HCM pce values, which apparently tend to underestimate effects of trucks and heavy vehicles *(9, 10)*. In the transportation planning and traffic assignment literature, virtually no effort has specifically calibrated link performance functions with explicit car and truck volumes.

Network equilibrium models require that the link performance functions be monotonically increasing functions of flow, and commonly used solution algorithms require these functions to be continuously differentiable. When the function has more than one argument, such as flows of multiple vehicle classes, it is required that the Jacobian matrix (of first-order partial derivatives with respect to the flow variables) have a positive diagonal. The most commonly used functional form for link performance functions is that of the Bureau of Public Roads (BPR) *(11)*:

$$T = T_0 [1 + \alpha (V/\kappa)^{\beta}] \quad (1)$$

where

T = average travel time per unit distance (at prevailing volume V),
T_0 = travel time per unit distance under free-flowing conditions,
α, β = link-specific parameters to be calibrated, and
κ = "capacity" of the link (pce's per time unit).

In the original BPR form, κ was intended as the so-called "practical capacity" of the link *(11)*. Other researchers have defined it as the "steady-state capacity" [e.g., Steenbrink *(12)*], which is effectively equivalent to the maximum service flow (MSF) corresponding to level-of-service E in 1985 HCM terminology *(8)*. Essentially, κ is a link-specific parameter that could be estimated like the other parameters; this is often inconvenient given the above functional form. The main concern is to use a particular definition consistently for calibration and subsequent application.

When multiple classes of vehicles are present in the traffic stream, say $V_1, V_2, \ldots, V_K$, then the standard approach is to replace the volume V in the foregoing equation by $(\eta_1 V_1 + \eta_2 V_2 + \ldots + \eta_M V_M)$, where η_i is the pce factor for vehicle class i, $i = 1, \ldots, M$. As noted earlier, the source for these pce factors is the HCM, which suffers from the limitations mentioned earlier from the standpoint of link performance modeling.

In this paper, functions of the same basic BPR form but with separate car and truck volume components and no prior restrictions on the pce multipliers have been calibrated and found to provide relatively good agreement with the data. In addition, linear functions, applicable only over a limited range of traffic volumes corresponding to stable, mildly congested conditions, are reported for those facilities for which data were available. Next the data used in this study are described and some of the key features are highlighted.

DATA CHARACTERISTICS

The data used in this study were developed from a large data base intended as a source of information on urban freeway truck characteristics *(6)*, collected for the Federal Highway Administration in 1981-1982 on 11 different freeway facilities in four major metropolitan areas: Houston, Dallas, Atlanta, and Detroit. Because the original intent of the data was to examine the effect of different geometric features on freeway performance, the facilities selected exhibited four different basic geometries: merge, diverge, weave, and basic (pipe) freeway sections. The sections selected are generally characterized by level terrain, to avoid grade-induced complications. All facilities included in the analysis are six-lane facilities (three in each direction), with 12-ft lane width and adequate shoulders and medians.

The data contain more than 0.5 million individual vehicle trajectories, constructed from records of the activation of detectors consisting of low-profile tapeswitches affixed to the road surface and configured in standard traps within the travel lanes. The passage time of each vehicle at each trap is thus available for the duration of the observation periods, which range from 1 to more than 16 hr at the various locations, resulting in grand totals of 561,227 individual vehicle traces observed over 240 hr. With that information, the travel time per unit distance of a vehicle could be obtained by subtracting the time at which the

entry trap was activated from the time at the exit trap and dividing the resulting value by the known distance between the entry and exit traps.

As noted in the previous section, the data needed for the calibration of the link performance functions must necessarily be in aggregate form, because the intent is not to explain the considerable variation in individual vehicle performance nor to predict the minute-by-minute dynamics of traffic flow in the facility, but to characterize the effect of a prevailing average volume level on the average travel time experienced by users of the facility. The average volume on the link corresponding to a particular aggregation period is defined as the number of vehicles passing a certain point on the link (typically the entry or exit points) during that period divided by the length of that period. The average travel time per unit distance for that period is then the reciprocal of the space mean speed, as noted in the previous section. The selection of the length of the aggregation period, over which the individual vehicle data are aggregated to form valid observations for performance function calibration, is not a straightforward matter. Ideally, as noted by Branston *(4)*, "long time intervals" must be used in order to approximate steady-state conditions and avoid dealing with the accompanying dynamic phenomena. On the other hand, if the sampling interval is too long, averages might include distinctly different operating conditions. Furthermore, longer intervals might result in fewer observations (in the calibration data set, given a fixed total number of individual vehicle traces), covering a spectrum of operating conditions that is too limited to properly identify the underlying relation. Judgment needs to be applied in this regard given the particular conditions under consideration. In this study, different sampling interval lengths were tested, and an aggregation period of 60 min was ultimately selected. However, 10-min data were also used in some instances where observations would have been too limited or where conditions were sufficiently stable to yield good estimates of the pertinent average quantities.

CALIBRATION RESULTS

It is accepted in traffic engineering practice that average travel time (or speed) on freeways is only mildly sensitive to volume over a relatively wide range of volume levels, beyond which it increases rapidly and nonlinearly. This phenomenon is shown in Figure 1, a scatterplot of the average travel time (per unit distance) versus the corresponding prevailing volume (in vehicles per hour per lane), for the data points obtained at the test locations included in the data base. Similar patterns could be observed for each individual test section, though the extent of data availability across the volume spectrum varied greatly across locations *(13)*. Plots similar to Figure 1 for each section revealed that the first portion of the curve, referred to hereafter as the "linear" portion, extended up to volumes of about 1,300 vehicles per hour per lane. Unfortunately, data points in the "nonlinear" range were very sparse for many of the test locations and could not support the reliable estimation of site-specific link performance functions that apply across the full spectrum of traffic volume levels. This is because high volume conditions were either not attained or not sustained for a sufficiently meaningful period of time at many of the locations

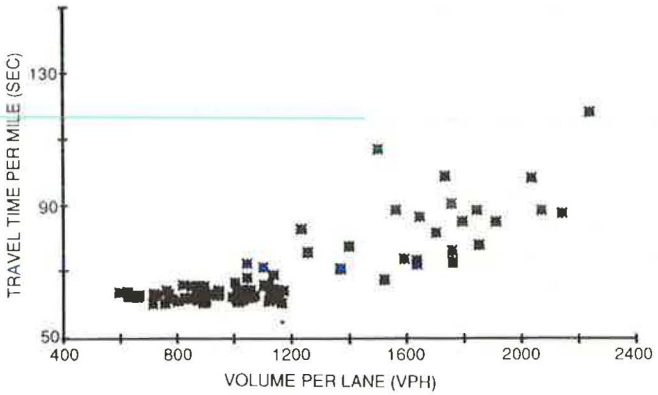

FIGURE 1 Travel-time–volume relationship, all data.

under consideration. Despite the fact that the data set was far from ideal in this regard, it is believed that useful insights and relations could still be obtained from the analysis presented here.

By pooling observations from sections with similar geometries, it was possible to calibrate the desired nonlinear performance functions, applicable over all volume levels, for pipe and merge sections. In addition, more detailed analyses were performed on the "linear" portion only for total vehicular volume levels below the 1,300-vph-per-lane threshold, to gain insight into the effect of trucks on freeway performance. Furthermore, the resulting equations may be of direct use in network assignment applications if the average operating conditions for certain facilities remain in the "linear" range. Relevant results from the analysis of the linear range are presented next, followed by the calibrated nonlinear functions.

Analysis of Linear Portion

The estimation results for three different specifications are reported for the range of vehicular volumes below 1,300 vph per lane:

1. A linear (in parameters as well as in variables) model with both car and truck volumes as the independent variables, as follows:

$$T_i = T_0 + C_1 V_{1i} + C_2 V_{2i} + \varepsilon_i \quad (2)$$

where

T_i = average vehicular travel time per unit distance (sec/mi) for the ith observation,
V_{1i}, V_{2i} = respective volumes of cars and trucks (vph per lane),
C_1, C_2 = parameters to be estimated, and
ε_i = random disturbance term, assumed, as usual, to be normally distributed with zero mean.

2. A linear model with the total volume in pce's (V_i) as the only explanatory variable; it is intended as a byproduct of the analysis to provide a useful model for the "linear" range that

would be very closely compatible with current practice and therefore uses prespecified pce multipliers from the 1985 HCM to convert trucks to pce's. The model therefore has the following form:

$$T_i = T_0 + C.V_i + \varepsilon_i \quad (3)$$

In this case, a pce factor of 1.7 was found to be applicable according to Table 3.3 in the 1985 HCM *(8)*.

3. A linear-in-parameters specification where the square of truck volume enters the model instead of V_2 in Equation 2. Actually, a number of such intrinsically linear specifications were explored, but only the results from this particular one are worthy of reporting.

The first specification (Equation 2) is the principal one for the purposes of this discussion. Results for the other models are only summarized in this paper. Because operating conditions were relatively stable at these volume levels, 10-min data were used in estimating the foregoing models.

The least-squares estimates for T_0, C_1, and C_2 in Equation 2 for pipe and diverge sections are presented in Table 1. All parameters are statistically significant (different from zero) at any reasonable level of significance, as is the overall regression. In Table 1, T_0, which corresponds to the free mean travel time per unit distance (i.e., reciprocal of the free mean speed), is expressed in seconds per mile; taking the inverse of the estimated values and converting to miles per hour yields respective values of about 63 and 61 mph, which is what one would expect for U.S. urban freeways.

Note that the volume effect coefficients C_1 and C_2 are expressed in seconds per mile per 100 cars or trucks; that is, they capture the expected changes in travel time with a change of link volume by 100 cars or trucks per lane. To formally establish what the numerical estimates for these coefficients strongly suggest, namely, that cars and trucks have different effects on average travel time, the hypothesis that $C_1 = C_2$ was tested for both facility types by using the general F-test for linear models *(14)*. To perform this test, the parameters of the "restricted" model (i.e., with $C_1 = C_2$) are estimated, yielding the sum of squared errors Q^R; similarly, the sum of squared errors for the "unrestricted" model, already estimated, is denoted Q^U. The test statistic is then calculated as $F^* = [(Q^R - Q^U)/r]/[Q^U/(n - k)]$, where n is the number of observations, k is the number of parameters to be estimated in the "unrestricted" model (in this case, $k = 3$), and r is the number of restrictions (in this case, $r = 1$). Under the null hypothesis that the restriction is true, this statistic is F-distributed with $(r, n - k)$ degrees of freedom. The results of this test for both types of sections are summarized in Table 2, revealing that, as expected, the null hypothesis should clearly be rejected, and implying that the truck effect on average travel time is in general different from that of passenger cars. Of course, C_1 is considerably smaller in magnitude than C_2. The ratio C_2/C_1 may be interpreted as a "volume effect" pce of trucks in the traffic stream in the volume range under consideration. Note, however, that this pce definition, which is the relevant one from the standpoint of link performance functions, is not altogether consistent with that used to come up with the HCM values. In particular, values of 3.6 and 13.9 are obtained for this ratio for the pipe and diverge sections, respectively, a far cry from the 1.7 suggested by the 1985 HCM for these types of facilities and typically used in current traffic assignment practice.

It should also be noted that the differentiation on the basis of geometric features, as is done in this study between pipe and diverge sections, constitutes a level of detail that is not usually associated with link performance functions in the context of network assignment problems. It was possible here because the data were available in that form. However, in practice it is unlikely that freeway links will be defined in that manner. In that case, the equation calibrated for pipe sections will be the more appropriate one to use.

As noted earlier, the foregoing linear specification was also estimated with the a priori restriction that the ratio C_2/C_1 was equal to the HCM value of 1.7 (i.e., the specification of Equation 3). The results are shown in Table 3. Naturally, because this model is a restricted version of the previous one, it cannot provide a better fit to the data. Furthermore, the volume effect pce values found earlier are quite different from the HCM value of 1.7 that would be used in conventional capacity analysis. The principal reason for including these results here is their potential usefulness in applications where the only infor-

TABLE 1 RESULTS OF PARAMETER ESTIMATION FOR LINEAR MODEL

Section Type	T_0 [a]	C_1	C_2 [b]	R^2	No. of Observations
Pipe	56.86	0.322	1.16	0.471	132
Diverge	58.93	0.205	2.84	0.574	146

NOTE: Range = $V \leq 1{,}300$ vph/lane.
[a] T_0 is expressed in seconds per mile.
[b] The coefficients C_1 and C_2 are expressed in seconds per mile per 100 vehicles.

TABLE 2 RESULTS OF SIGNIFICANCE TESTS OF DIFFERENCE BETWEEN CAR AND TRUCK VOLUME EFFECTS ON TRAVEL TIME ($C_1 = C_2$)

Site Type	$n - k$	r	Q^R [a]	Q^U [b]	F^* [c]	$F_{(0.05,r,n-k)}$ [d]
Pipe	129	1	241.8	233.8	4.41	3.90
Diverge	143	1	235.0	205.0	20.90	3.90

[a] Q^R is the sum of squared errors for restricted model.
[b] Q^U is the sum of squared errors for unrestricted model.
[c] F^* is the calculated value for F-test statistic.
[d] $F_{(0.05,r,n-k)}$ is the theoretical value for F-distributed statistic with r df for numerator and $n-k$ df for denominator at the 5 percent significance level.

TABLE 3 RESULTS OF PARAMETER ESTIMATION FOR RESTRICTED LINEAR MODEL WITH HCM pce VALUES

Section Type	T_0	C_1 [a]	R^2	No. of Observations
Pipe	56.36	0.515	0.419	132
Diverge	59.63	0.348	0.523	146

NOTE: Range = $V \leq 1{,}300$ vph/lane.
[a] The coefficient C_1 is expressed in seconds per mile per 100 vehicles.

mation available is given in pce flows; this could arise when the agency providing the data used in network assignment has already applied HCM pce factors, or when future-year forecasts do not break down the projected traffic into its constituent elements.

Given the empirical basis of these (and most other link performance) functions, several alternative specifications have been considered and estimated in the course of this analysis. In particular, specifications including power terms of the two principal independent variables, as well as multiplicative interaction terms, were tested. In general, these specifications were inferior to the simple linear model presented earlier. It is worthwhile to comment on the results of one specification, where the squared value of the truck volume is used, as follows:

$$T_i = T_0 + C_1.V_{1i} + C_2.(V_{2i})^2 + \varepsilon_i \quad (4)$$

where all terms are as defined previously. Table 4 summarizes the parameter estimation results. The model did not exhibit a discernible improvement in terms of statistical performance relative to the earlier linear version (slight improvement for diverge sections, but inferior performance for pipe data). However, its implications appear intuitively plausible and worthy of further examination.

TABLE 4 RESULTS OF PARAMETER ESTIMATION FOR MODEL 3

Section Type	T_0	C_1[a]	C_2[b]	R^2	No. of Observations
Pipe	58.38	0.315	0.212	0.425	132
Diverge	59.77	0.184	2.14	0.597	146

NOTE: Range = $V \leq 1{,}300$ vph/lane.
[a] C_1 is expressed in seconds per mile per 100 vehicles.
[b] C2 is expressed in seconds per mile per (100 vehicles)^2.

Although the values of T_0 and C_1 are directly comparable to those obtained with the previous model (Equation 2), the interpretation of C_2 is not as straightforward. The assumption here is that the marginal effect of truck volume is proportional to the prevailing volume of trucks, with $\partial T/\partial V_2 = 2C_2V_2$. Other similar assumptions, but with truck effect proportional to car or total volumes, were also considered but did not perform satisfactorily. The values reported in Table 4 indicate that the volume effect pce of trucks (now given by $2C_2V_2/C_1$) can vary over a rather wide range. For example, when truck volume is 100 vph (per lane), the ratio of the (marginal) truck effect to that of cars is 23.26 for diverge sections and 1.34 for pipe sections. Unfortunately, the limited nature of the data precludes more definitive conclusions, but suggests that the truck effect on freeway performance may not be captured very well by the HCM values for certain geometric features.

Nonlinear Performance Function

A nonlinear function of the BPR type, applicable over the full volume spectrum, is presented here. As noted earlier, adequate data for this analysis were available only for pipe and merge type locations. Furthermore, the number of corresponding observations was severely limited, especially for the higher-volume portions of these curves. Nevertheless, the resulting calibrated functions are useful, because they satisfy the desired properties for network equilibrium applications, in addition to providing insights on the functional form and the relative magnitude of the coefficients.

Before the BPR-type specification was estimated, an extensive exploratory analysis was conducted with linear-in-parameters specifications using polynomial regressions in which power terms of independent variables (ranging from first to fifth power), as well as multiplicative interaction terms, were included. The principal general results of this analysis can be summarized as follows: (a) higher-power terms appear to contribute more to explaining the variation in the dependent variable (travel time) than lower-power terms do; (b) different powers of the same independent variable exhibit high correlation with each other; (c) the multiplicative interaction terms are also highly correlated with the other independent variables; and (d) the foregoing leads to at least one negative value among the estimated coefficients whenever more than two of these terms are used; such negative values are not plausible and cannot be accepted in a well-specified model.

Consistent with the foregoing results and with the findings of the earlier analysis for the lower volume range, a BPR-type model was specified as follows:

$$T_i = T_0 \{1 + \alpha[(V_{1i} + \eta.V_{2i})/\kappa]^\beta\} + \varepsilon_i \quad (5)$$

where α, β, and η are parameters to be estimated, and all other terms are as previously defined. Note that the capacity κ, as discussed in the second section, is also a parameter describing the link. In the present analysis, it could not be identified separately; rather, the term $(\alpha T_0/\kappa^\beta)$ was treated as a single parameter value in the least-squares estimation. The value of α was then recovered by setting the value of κ at its HCM value of 2,000 pce's/hr, as discussed hereafter.

Nonlinear least-squares estimators for these parameters were obtained by performing a numerical search for the global optimum over a grid of η- and β-values and using linear least squares to estimate the resulting linear specification for each combination of η- and β-values. The parameter estimates are given in Table 5 for pipe and merge sections separately, and are also pooled for both types of sections. The values obtained in this table are plausible and consistent with prior engineering knowledge. For instance, the values of β (4.7, 4.5, and 4.8) are in the range of 4 to 6 reported by other researchers.

To formally establish that the truck effect on travel time is significantly different from that of cars, an F-test of the hypothesis that the corresponding restriction is true (i.e., that $\eta = 1$) was performed. The test procedure is similar to that used earlier for the linear models, because it is still applicable, with minor adjustments, to the nonlinear case [details on the test procedure in conjunction with nonlinear models have been discussed by Amemiya *(15)*]. The results, presented in Table 6, indicate that the volume effect of trucks is significantly different from that of cars for the pipe sections and for the pooled data, but not for merge sections taken separately. Therefore trucks appear to be more disruptive relative to cars on pipe sections than on merge

TABLE 5 RESULTS OF PARAMETER ESTIMATION FOR NONLINEAR MODEL

Section Type	T_0	C_1[a]	α[b]	η	β	R^2	No. of Observations
Pipe	60.62	0.81×10^{-14}	0.438	2.2	4.7	0.908	43
Merge	61.77	0.41×10^{-13}	0.477	1.5	4.5	0.655	65
Pooled Data	61.51	0.38×10^{-14}	0.431	2.1	4.8	0.861	108

NOTE: Applicable over full volume range.
[a] $C = \alpha \cdot T_0/\kappa^{\beta}$.
[b] The value of α reported here is recovered from C assuming that $\kappa = 2{,}000$ pce per hour per lane.

TABLE 6 RESULTS OF SIGNIFICANCE TESTS OF DIFFERENCE BETWEEN CAR AND TRUCK VOLUME EFFECTS ON TRAVEL TIME (NONLINEAR MODEL)

Site Type	$n-k$	r	Q^R [a]	Q^U [b]	F^* [c]	$F_{(0.05,r,n-k)}$[d]
Pipe	39	1	2181.7	1562.9	15.44	4.08
Merge	61	1	1854.0	1849.0	0.17	4.0
Pooled data	104	1	4150	3459	20.78	3.92

[a] Q^R is the sum of squared errors for restricted model.
[b] Q^U is the sum of squared errors for unrestricted model.
[c] F^* is the calculated value for F-test statistic.
[d] $F_{(0.05,r,n-k)}$ is the theoretical value for F-distributed statistic with r df for numerator and $n-k$ df for denominator at the 5 percent significance level.

sections. The pce values obtained here also appear to be smaller than the values obtained in the linear model calibrated for the lower volume range. Both conclusions are consistent with the view held by traffic researchers *(9, 16)* that the constraining effect of trucks on travel time is greater at higher speeds than at lower speeds (associated with higher volumes), where vehicles are already operating in a constrained mode.

As noted in conjunction with the models developed for the linear portion, the results calibrated for merge sections are not likely to be useful in the context of traffic assignment applications, because links are rarely defined at this level of detail. The functions calibrated for either pipe sections or pooled data would be more appropriate for such applications.

The functions presented so far yield the travel time per unit distance for an average vehicle. In applications involving the explicit differentiation between cars and trucks as separate user classes, there is concern that the average travel time experienced by cars may be different from that experienced by trucks. This problem is addressed in the next section, in which the results are given of an investigation of the relation between the respective averages for both vehicle classes. This provides the basis for obtaining separate estimates of these quantities given the average vehicular travel time determined from the foregoing performance functions.

RELATION BETWEEN CAR AND TRUCK TRAVEL TIMES

To examine whether the same value for the average travel time applies for both cars and trucks using a given link, the average travel time was calculated for cars and trucks separately for each observation period. In a plot of these averages for all the 60-min data points, which includes the best-fitting straight line (Figure 2), the linearity of the resulting relation between these two quantities is striking. This suggests a simple relation that allows the calculation of the average travel time experienced by either class of vehicles given that of the other or, as in this case, given that of an average vehicle. The following equation was thus calibrated:

$$T_{Ti} = B_0 + B_1 T_{Ai} + \varepsilon_i \tag{6}$$

where T_{Ti} and T_{Ai} are the respective average travel times per unit distance for trucks and cars for the ith observation period and B_0 and B_1 are parameters to be estimated.

The least-squares estimates, based on observations from all section types, are -4.78 for B_0, and 1.075 for B_1 (the corresponding R^2 is 0.80). Both parameters are statistically significant (different from zero) at any reasonable level of confidence. However, the hypothesis that $B_1 = 1.0$ (against the alternative that $B_1 \neq 1.0$), tested using the standard quasi-t-test, can be rejected at the 11 percent significance level but not at the 10 percent level. A slope of 1.0 would of course imply that average travel time for trucks increases at the same rate as car travel time; the results obtained here seem to suggest that overall, truck travel time increases at a slightly faster rate than that of cars. However, some differences in this pattern were

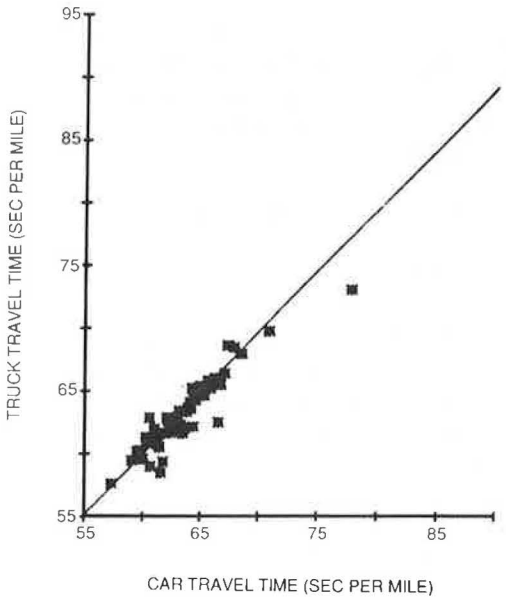

FIGURE 2 Truck versus car average travel time, all data.

observed across different geometric features, though this level of detail is not central to the focus of this paper and is presented elsewhere *(13)*.

The negative value of the estimated intercept B_0 is also worthy of note. Of course, its literal interpretation (i.e., average truck travel time when car travel time is equal to zero) is meaningless in this context, given the range of operating conditions encountered on freeways and in the estimation data base. Essentially, the meaningful range of average travel times to which this analysis is applicable has a lower bound of about 50 sec/mi (corresponding to an average speed of about 72 mph). The reason for the negative intercept is that truck drivers tend to go faster than passenger car drivers when traffic conditions are essentially free flowing, therefore allowing higher speeds (lower travel times per unit distance) to be reached. However, this trend is reversed at lower speeds, when the positive contribution from $(B_1.T_A - T_A)$ offsets the negative intercept (in Equation 6). In other words, when unimpeded, truck drivers go faster, on average, than passenger car drivers; however, their speed deteriorates more rapidly than that of cars with increasing congestion in the facility, given their lower acceleration capability and lack of maneuverability, to where average truck travel time per unit distance exceeds the corresponding value experienced by cars. This effect should be even more noticeable on steep grades, which were not available in this data base.

However, the net travel time differentials between cars and trucks predicted by Equation 6 are minute, and can, for all practical purposes, be ignored in the context of traffic assignment applications. Nevertheless, the equation can be used to calculate the respective average travel times for cars and trucks given the average vehicular travel time obtained by the link performance functions presented earlier. This is accomplished by noting that the average vehicular travel time T is the weighted average of the respective car and truck travel times; that is,

$$T = (V_1/V)T_A + (V_2/V)T_T \quad (7)$$

where $V = V_1 + V_2$.

If T_A and T_T are related by Equation 6, then substituting this relation into Equation 7 and some algebraic manipulation yield expressions for these two quantities in terms of V_1, V_2, and T (itself obtained from a function such as Equation 5, given V_1 and V_2), as follows:

$$T_T = (B_1 V T + B_0 V_1)/(V_1 + B_1 + B_1 V_2)$$

$$T_A = (V T - B_0 V_2)/(V_1 + B_1 V_2)$$

However, it would be simpler, and justified in light of these results, to use T for both vehicle classes unless a particular application involves links with unusual geometric features that might more severely bring out limitations in truck performance characteristics.

CONCLUDING COMMENTS

Results have been presented of the empirical development and calibration of link performance functions that capture the dependence of travel time on the respective volumes of passenger cars and trucks sharing the physical right-of-way. These functions are intended for use in network equilibrium studies requiring the assignment of explicit car and truck flows, and therefore involving asymmetric interactions between these vehicle classes (or equivalently between links). This problem arises, for example, in the context of the evaluation of truck-related highway improvements, which is a problem of current interest to highway agencies.

This work is primarily exploratory in nature, given its reliance on less than ideal secondary data initially developed for the microscopic analysis of certain aspects of truck traffic on freeways. Nevertheless, useful relations applicable to a broad range of freeway traffic conditions were developed. These performance functions, based on the widely used BPR form, can be used directly in current traffic assignment models. Useful insights were also obtained regarding the effect of trucks on freeway performance; for example, the estimated parameter values appeared to suggest that there may be differences in the marginal effect of trucks (relative to that of cars) at high versus low speeds.

The work presented here must be viewed as only a first step toward better understanding of the interaction of various vehicle classes in determining the performance of transportation facilities. In the context of equilibrium studies in urban networks, link performance functions that capture interactions among vehicle classes in signalized arterials and urban streets have not received adequate attention. It is for the latter type of facilities that the models used in current practice may be severely underestimating the effect of slow or heavy vehicles, especially in light of the continuing trend toward using larger trucks for urban goods movement.

ACKNOWLEDGMENTS

The authors are grateful to Paul Ross and Stephen Cohen of FHWA for making the data base available for this study. The assistance of Diane Gierish and Kyriacos Mouskos in handling and processing the large data files is appreciated. Partial funding for this work came from several sources, including the Department of Civil Engineering and the Bureau of Engineering Research at the University of Texas at Austin and the Texas State Department of Highways and Public Transportation through their support of truck-related research projects in the Center for Transportation Research at the university. The authors, of course, remain solely responsible for the contents of this paper.

REFERENCES

1. T. L. Friesz. Transportation Network Equilibrium, Design and Aggregation: Key Developments and Research Opportunities. *Transportation Research*, Vol. 19A, No. 5/6, 1985, pp. 413–427.
2. Y. Sheffi. *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Prentice-Hall, Englewood Cliffs, N.J., 1984.
3. M. Florian et al. *Validation and Application of EMME: An Equilibrium Based Two-Mode Urban Transportation Planning Method*. Centre de Recherche Sur les Transports, University of Montreal, Canada, 1979.

4. D. Branston. Link Capacity Functions: Review. *Transportation Research,* Vol. 10, No. 4, 1976, pp. 223–236.
5. H. S. Mahmassani, C. M. Walton, K. Mouskos, J. J. Massimi, and I. L. Levinton. *A Methodology for the Assessment of Truck Lane Needs in the Texas Highway Network.* Center for Transportation Research Report 356-3F. University of Texas at Austin, 1985.
6. *Urban Freeway Truck Characteristics.* FHWA, U.S. Department of Transportation, 1982.
7. V. F. Hurdle and D. Y. Solomon. Service Functions for Urban Freeways—An Empirical Study. *Transportation Science,* Vol. 20, No. 3, 1986, pp. 153–163.
8. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
9. M. Van Aerde and S. Yagar. Capacity, Speed, and Platooning Vehicle Equivalents for Two-Lane Rural Highways. In *Transportation Research Record 971,* TRB, National Research Council, Washington, D.C., 1984, pp. 58–65.
10. R. P. Roess and C. J. Messer. Passenger Car Equivalents for Uninterrupted Flow: Revision of Circular 212 Values. In *Transportation Research Record 971,* TRB, National Research Council, Washington, D.C., 1984, pp. 7–14.
11. *Traffic Assignment Manual.* Bureau of Public Roads, U.S. Department of Commerce, 1964.
12. P. A. Steenbrink. *Optimization of Transport Networks.* Wiley, New York, 1974.
13. Y. G. Kim. *Performance Functions for Highway Links with Asymmetric Interactions Between Cars and Trucks.* M.S. thesis. Department of Civil Engineering, University of Texas at Austin, 1986.
14. J. Neter and W. Wasserman. *Applied Linear Statistical Models.* Irwin, New York, 1974.
15. T. Amemiya. Nonlinear Regression Models. In *Handbook of Econometrics* (Z. Griliches and M.D. Intriligator, eds.), Vol. 1, North-Holland, Amsterdam, 1983.
16. J. Craus. Revised Method for the Determination of Passenger Car Equivalencies. *Transportation Research A,* Vol. 4A, 1980, pp. 241–246.

A Methodology for Feeder-Bus Network Design

GEOK KOON KUAH AND JOSSEF PERL

The U.S. transit industry faces financial difficulties. Among the strategies suggested for improving transit financial conditions, the development of better-integrated intermodal systems has the advantage of potentially achieving both cost reduction and improved service. A network optimization methodology for the design of an integrated feeder-bus–rail rapid transit system is presented. The Feeder-Bus Network Design Problem (FBNDP) is defined as that of designing a set of feeder-bus routes and determining the frequency on each route so as to minimize operator and user costs. The FBNDP is first considered under many-to-one demand and its formulation is discussed as a mathematical programming problem. Then the generalization of the formulation to the many-to-many demand pattern is reviewed. The FBNDP is a larger and complex routing-type problem that can be solved only heuristically. A heuristic method that generalizes the savings approach to consider operating frequency is presented. The analysis presented illustrates the capabilities of the proposed model as a strategic planning tool for feeder-bus network design. It indicates that changes that increase the relative weight of operator cost often result in feeder-bus networks with less circuitous routes operated at lower frequencies, whereas changes that increase the relative weight of user cost result in feeder routes operated at higher frequencies. The solutions provided by the proposed model have been tested and found superior to manually designed networks, particularly under variable demand.

The U.S. transit industry faces financial difficulties *(1)*. Rising costs and shrinking resources are the main causes for these difficult financial conditions. From 1960 to 1983, the annual urban transit operating costs rose by more than $6.7 billion. Only a small declining fraction of these operating expenditures were covered by farebox revenues *(2, 3)*. In recent years, the uncertain financial conditions of the transit industry have been made more acute by the reductions in federal funding for mass transit. Since FY 1981, total federal funding available for mass transit has declined by 28 percent *(2)*. In view of declining federal assistance, transit agencies can expect to face growing pressure to reduce deficits.

Several strategies have been suggested for improving the financial conditions of transit agencies: (a) new funding schemes to raise subsidies from local or state governments, or both; (b) new fare structures to increase revenues; (c) improvements in service quality to attract ridership; (d) reduction in the commitment to peak-period services ("shedding the peak"); (e) reduction or elimination of services to low-density areas; (f) use of more cost-effective technologies; (g) privatization of transit services; and (h) development of better-integrated intermodal systems.

Each of these strategies has its shortcomings. New funding schemes may be infeasible under the current political climate, because they would require governments to increase taxes for mass transit. New fare structures based on distance or time, or both, may increase revenue (depending on demand elasticities) with a loss of patronage to the automobile mode. Such a shift would increase the need for investments in highway facilities and would have negative environmental impacts. Service improvements would result in higher transit operating costs, which most likely would not be fully compensated for by increased revenues. The logic behind the strategy of shedding the peak is that a substantial proportion of transit investment is needed only for peak-period service. If private operators could share the burden of these services, it would reduce the capital requirements of transit agencies. The success of the strategy of shedding the peak as well as that of privatization depends on the willingness of the private sector to enter the transit industry, which, so far, has been limited.

The strategy of reducing (or eliminating) services to low-density areas may cause some captive riders to lose their mobility. Public transit operators may be reluctant to consider this strategy *(4)*. The employment of more cost-effective transportation technologies, such as vanpools, carpools, paratransit, and taxi, has in some cases reduced the cost of transit services. However, these technologies are suitable primarily for low-density areas and are usable only as components of an integrated intermodal transit system.

The development of better-integrated intermodal transit systems can achieve both reduction in cost and improvement in service quality. Better integration can reduce transit cost by eliminating duplications and employing the most cost-effective mode in each segment of the system. An improvement in service quality would result from better coverage and reduced access cost, fewer transfers, and shorter travel times. In turn, improved service would increase revenues to provide further deficit reduction. One type of integration with specific additional advantages involves the employment of bus transportation as access mode to a rail rapid transit system. First, a well-integrated feeder-bus–rail rapid transit system reduces parking requirements at the rail stations, thereby reducing the rail system's capital cost. Second, a well-integrated feeder-bus–rail system may attract automobile trips (primarily work trips), thereby increasing the economic viability of the rail system.

The potential for improving the financial condition of transit by designing an integrated feeder-bus–rail rapid transit system has been recognized for some time *(5–8)*. The integration of

G. K. Kuah, Gorove/Slade Associates, Inc., 1140 Connecticut Avenue, N.W., Suite 1200, Washington, D.C. 20036. J. Perl, Department of Civil Engineering, University of Maryland, College Park, Md. 20742.

feeder bus and rail rapid transit has been suggested as one of the most promising future directions for public transit in large U.S. cities. In the last 15 years, several new rail rapid transit systems have been constructed in large U.S. cities (San Francisco; Atlanta; Washington, D.C.; Buffalo; Miami). The bus and rail rapid transit systems in most of these cities are not well integrated *(9)*. A common practice in designing feeder-bus networks has been to turn the buses back at the nearest rail stations. The duplication of bus service along rail lines is also a common phenomenon *(9)*.

There is currently no methodology for designing an integrated feeder-bus–rail transit network. Most of the existing work on transit network design has focused on single-mode networks *(10–12)*. Often the focus has been on individual components of the network-design problem such as route structure *(13–15)*, service frequency *(16, 17)*, and station spacing *(18, 19)*. Results are presented of a recent study in which a network optimization methodology was developed for designing an integrated feeder-bus–rail transit system. The proposed methodology focuses on the network design elements of the integration problem while including related operational elements such as service frequency.

THE FEEDER-BUS NETWORK DESIGN PROBLEM

Integrated feeder-bus–rail transit systems have a variety of design components, which may include the network structure of the rail and bus systems and the levels of service on each component of the system. However, the basic decisions regarding the structure of the rail network, such as the locations of rail lines and rail stations, are based on projected land use and are not greatly affected by decisions regarding the feeder-bus system. The integration problem can therefore be viewed as one of designing a feeder-bus system that can access an existing rail network, which is defined as the Feeder-Bus Network Design Problem (FBNDP). Specifically, the FBNDP is the problem of designing a set of feeder-bus routes and determining the service frequency on each route so as to minimize the sum of operator and user costs.

The proposed methodology represents the FBNDP as a network optimization problem. The network includes two types of nodes—rail nodes and bus nodes—that represent rail stations and bus stops, respectively. Similarly, rail links represent rail line segments, whereas bus links represent feeder-bus route segments. The demand is assumed to be concentrated at nodes and the temporal distribution of demand is not represented. This representation of demand is common to network models, primarily those dealing with strategic problems such as network design and location. In the FBNDP, the demand can be viewed as hourly averages for a given time period, for example, peak period or off-peak period.

The FBNDP can be viewed as the problem of achieving the optimal balance between operator cost and user cost. Operator cost includes the capital cost associated with the fleet and variable costs, which are related to vehicle hours of travel. In the context of the FBNDP, user cost includes the time costs of access, wait, and riding. Because the design of the feeder-bus system determines the riding time in the rail system, riding time includes in-vehicle time on both bus and rail.

Two different demand patterns are considered—many-to-one and many-to-many. Many-to-one (M-to-1) refers to a demand pattern with multiple origins and a single destination. Peak-period work trips to and from the central business district (CBD) may exhibit this pattern. Many-to-many (M-to-M) demand refers to a pattern with multiple origins and destinations. Clearly, nonwork trips would likely follow this pattern.

M-to-1 FBNDP

In the development of a network optimization model for the M-to-1 FBNDP, the following assumptions are made:

1. Each bus stop is served by one feeder-bus route.
2. Each bus route is linked to exactly one rail station.
3. Buses have standard capacity and operating speed.

The first assumption may appear restrictive. However, it is valid for a system serving M-to-1 demand. When all the passengers have a common destination, it is unnecessary to have multiple bus routes serving the same bus stop. This assumption is relaxed in the case of the M-to-M demand pattern. The second assumption implies that buses are not allowed to travel along rail lines. This assumption is consistent with one of the basic purposes of integration discussed earlier—the elimination of duplicate services. The third assumption is consistent with the common practice of operating fixed-route transit service with the same type of vehicle. It is also a widely accepted assumption in mathematical models for routing-type problems, which significantly reduces the complexity of the models.

Figure 1 presents a feeder-bus–rail transit system with five rail lines serving a single destination. Also shown are the feeder-bus routes for two of the rail lines. The feeder-bus–rail system shown in Figure 1 can be represented as a spanning-tree network (Figure 2) in which the destination is the root, the rail stations are first-level nodes, and the bus stops are higher-level nodes. Figure 2 constitutes a conceptual representation of an integrated feeder-bus–rail system, to be used in the formulation of a network optimization model for the FBNDP. The costs associated with the first-level links (rail links) represent riding-time costs in the rail system. Those associated with the higher-level links (bus links) represent bus operating costs and passenger riding-time costs. Those associated with the bus nodes represent passenger wait-time costs.

Given the association between a feeder-bus–rail system as shown in Figure 1 and the corresponding spanning-tree network of Figure 2, the optimal solution to the FBNDP under an M-to-1 demand pattern would be obtained by finding the spanning-tree network that minimizes the sum of operator and user costs. The problem of finding the minimum-cost spanning-tree network can be formulated as a mathematical programming model *(20)*. The structure of the model is as follows:

Minimize (rail riding cost + bus operator cost + bus user cost)
subject to logical route constraints, route capacity constraints, fleet size constraint, and route length constraints.

There are five types of logical route constraints. The first places each bus node on a single feeder-bus route. The second

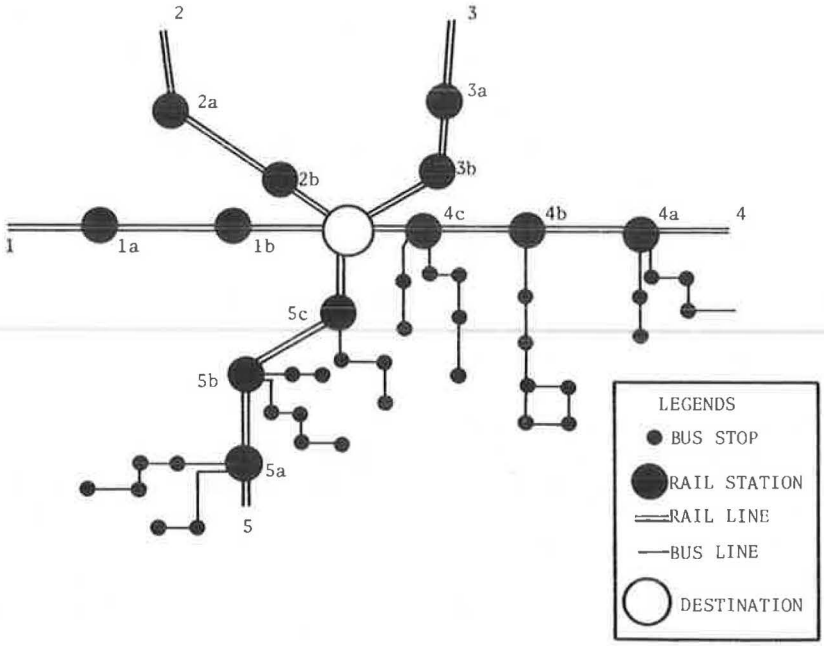

FIGURE 1 Feeder-bus–rail transit system for the case of many-to-one demand.

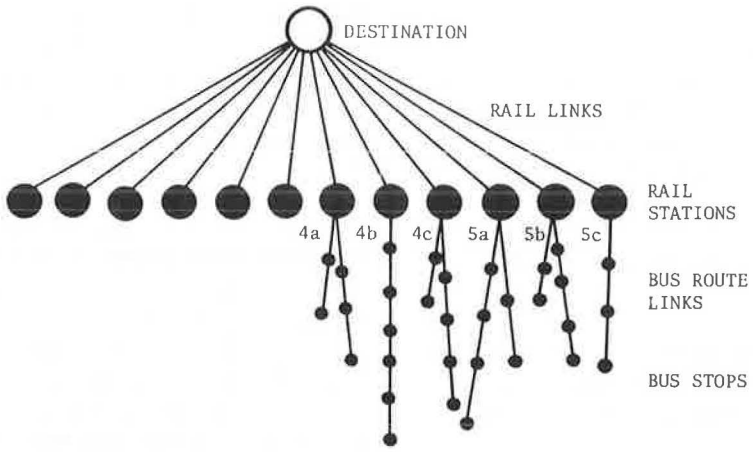

FIGURE 2 Spanning-tree network representation of the feeder-bus–rail transit system for the case of many-to-one demand.

ensures that each bus route is linked to a single rail node. The third is a route continuity constraint, which states that a route that enters a bus node must leave that node. The fourth ensures that every route is linked to a rail node. This is a "subtour elimination" constraint, which appears in mathematical models for the well-known Traveling Salesman Problem (TSP). The route capacity constraints ensure that the demand on any route does not exceed the capacity. It should be noted that unlike existing vehicle routing problems, in the FBNDP route capacity is not an input parameter, because route frequency is a decision variable. The fleet size constraint ensures that the total seat-hours offered on the feeder-bus system does not exceed the available seat-hours. Finally, the route length constraints ensure that the length of any route does not exceed the prespecified maximum route length.

The mathematical programming model for the M-to-1 FBNDP is a large and difficult vehicle-routing type model. There are two elements that contribute to the added complexity of this model relative to mathematical models for existing vehicle-routing problems. First, the objective function in the FBNDP model is nonlinear. Second, the FBNDP includes an additional decision variable—service frequency. The proposed solution method for the FBNDP is discussed next.

M-to-M FBNDP

The proposed mathematical model for the M-to-1 FBNDP can be generalized to the M-to-M demand pattern. The M-to-M FBNDP differs from the M-to-1 FBNDP in that the set of

destinations includes the entire set of rail stations. In the M-to-M FBNDP, the demand at each bus stop is a multidimensional quantity. Clearly, the M-to-M FBNDP is not simply a sum of M-to-1 FBNDPs, because under an M-to-M demand pattern a single feeder-bus route usually serves demands to multiple destinations. The problem under the M-to-M demand pattern appears significantly more difficult. First, the design of the feeder-bus network should take into account not only the linkings to alternative rail stations, but also alternative connections to rail lines. Depending on which rail line is chosen for connection, passengers may or may not have to transfer between rail lines. Second, the optimal feeder-bus network may include some bus stops on more than a single feeder-bus route. This results in a significantly more complex feeder-bus network.

Interestingly, with relatively minor modifications the conceptual representation of Figure 2 is applicable to the M-to-M FBNDP. Figure 3 shows a feeder-bus–rail transit system with five destinations, and Figure 4 shows the spanning-tree network representation of the system. Unlike the spanning tree of Figure 2, that in Figure 4 includes multiple rail links between each rail node and the dummy node. These links represent the travel from the associated rail station to all other destinations. In the M-to-1 FBNDP there is only a single destination and therefore only a single link is needed to represent the travel from any given rail station to the destination. Unlike Figure 2, the tree network representation in Figure 4 does not provide a suitable representation of the M-to-1 FBNDP for a network optimization model, because a bus node in Figure 4 does not uniquely identify the destination of demand.

To formulate a mathematical model for the M-to-M FBNDP, the network representation of Figure 4 needs to be modified so that a single destination is associated with each bus node. This is done by splitting each bus node into multiple subnodes (one for each destination). The locations of the subnodes are the same as that of the original bus node. However, each subnode

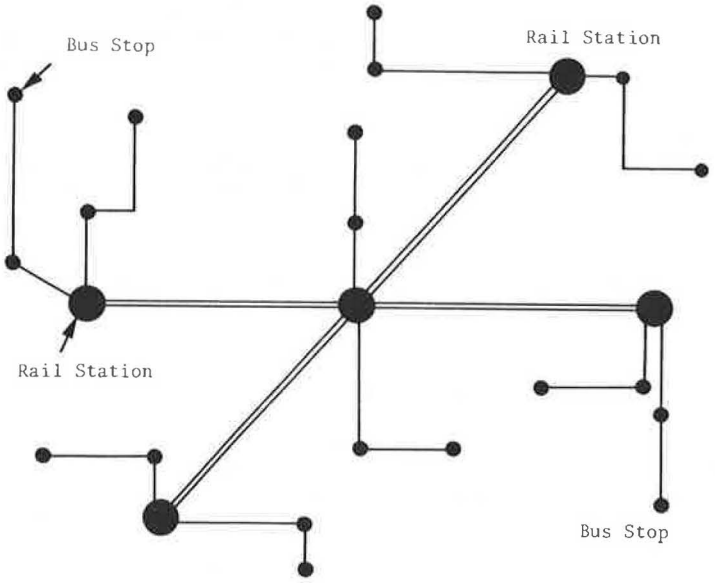

FIGURE 3 Feeder-bus–rail transit system for the case of many-to-many demand.

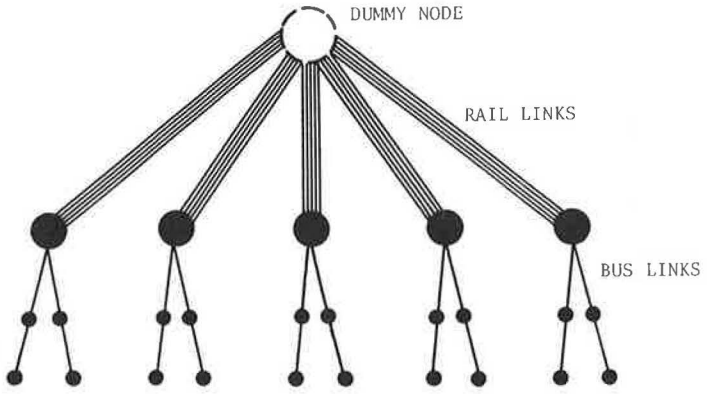

FIGURE 4 Spanning-tree network representation of the feeder-bus–rail transit system for the case of many-to-many demand.

represents the demand to a single destination. With this transformation, the spanning-tree representation of Figure 4 becomes similar to that of Figure 2. The only difference is the multiple rail links of Figure 4. However, this does not represent any conceptual difficulty, because the flow on the rail link connected to any given rail node is completely defined by the allocation of bus subnodes to that rail node. The spanning-tree representation of Figure 4 (with the foregoing transformation) represents the M-to-M FBNDP as an M-to-1 FBNDP. The mathematical model for M-to-1 FBNDP can therefore be used to represent the M-to-M case.

As stated, the proposed mathematical model includes a fleet size constraint. As such it represents the FBNDP in the context of short-term planning. In the long term, the fleet size can be adjusted to changes in the system. With a relatively minor modification, the proposed model can be adapted to represent the long-term FBNDP. In the long-term model, the fleet size constraint is removed and a fleet size cost component is added to the objective function. This additional cost component represents the depreciation cost associated with the fleet, based on an assumed unit depreciation cost per bus hours used.

SOLUTION ALGORITHM

As stated earlier, the FBNDP is a large and complex routing-type problem. Routing models with linear objective functions can be solved optimally only for very small test networks *(21)*. Under the M-to-1 demand pattern, for a system with 5 rail stations, 50 bus stops, and 6 feeder-bus routes, the proposed mathematical model would include 1,756 variables and more than 10^{15} constraints. Clearly such a model can be solved only heuristically.

The spanning-tree networks of Figures 2 and 4 are similar to the network representation of a single-echelon physical distribution system. On the basis of this similarity, one can notice the similarities between the FBNDP and the Multi-Depot Vehicle-Routing Problem (MDVRP), which is the problem of designing a set of delivery routes from several depots to a large number of demand points so as to minimize the total route distance. The similarity between the two problems is useful to the development of a heuristic method for the FBNDP, because it may allow the use of certain concepts of existing heuristics for the MDVRP. However, the FBNDP differs from the MDVRP in several basic elements: (a) although the objective in the MDVRP is to minimize operator cost, the FBNDP objective includes both operator and user costs; and (b) in the MDVRP the operating frequency is predetermined, whereas in the FBNDP operating frequency is a decision variable.

A detailed description of the proposed heuristic method for the FBNDP is beyond the scope of this paper and can be found elsewhere *(20)*. Only a summary of the basic elements is provided here. The variables, parameters, and measures used in the discussion of the heuristic method are defined in Table 1. The method consists of an initial algorithm and two improvement procedures. The initial algorithm generates an initial feasible feeder-bus network and determines the frequency on each feeder-bus route. Subsequently, the initial network is improved by these procedures.

The initial algorithm uses the sequential savings approach, which was used in previous algorithms for the vehicle-routing problem *(22)*. It starts by computing the cost of a direct route

TABLE 1 VARIABLES, PARAMETERS, AND MEASURES FOR HEURISTIC METHOD

Symbol	Units	Description
q_i	Pass./Hour	Average demand per hour at bus stop i
f_{ij}	Veh./Hour	Bus operating frequency of direct route from stop i to station j
l_{ij}	Miles	Distance from stop i to station j
C_{sj}	$/Passenger	Unit rail cost from station j to destination s
λ_0	$/Veh.-Mile	Unit bus operating cost
λ_r	$/Pass.-Hour	Value of passenger riding time
λ_w	$/Pass.-Hour	Value of passenger wait time
U	Miles/Hour	Average bus operating speed
L_k	Miles	Length of route segment k
Q_k	Pass./Hour	Demand on route segment k
TC_i^j	$/Hour	Total cost of direct route from stop i to station j
$SAVTC_i^k$	$/Hour	Saving from including stop i in route segment k

from each bus node to its nearest rail node as given by Equation 1. The four terms of Equation 1 represent rail riding cost, bus operating cost, bus passenger riding cost, and bus passenger waiting cost, respectively. The cost of wait time at rail stations is assumed to be relatively small and is not represented in Equation 1.

$$TC_i^j = C_{sj}q_i + 2 \cdot \lambda_0 f_{ij} \cdot l_{ij} + (\lambda_r/U) \cdot l_{ij} \cdot q_i + \lambda_w q_i/2f_{ij} \quad (1)$$

Equation 1 provides the direct-route cost as a function of operating frequency. To obtain the minimum direct-route cost, the first-order conditions are used to obtain the optimal frequency, as given by Equation 2.

$$f_{ij}^* = 0.5(\lambda_w q_i/\lambda_0 l_{ij})^{1/2} \quad (2)$$

From Equation 2, it can be seen that the optimal frequency is that which results in equality between operator cost and passenger wait-time cost. Substituting Equation 2 in Equation 1, the minimum cost of a direct route is obtained as follows:

$$TC_i^j = C_{sj}q_i + 2(\lambda_0\lambda_w l_{ij}q_i)^{1/2} + (\lambda_r/U) \cdot l_{ij}q_i \quad (3)$$

The second and third components of Equation 3 represent the direct-route cost associated with the feeder-bus system. The algorithm initiates a route by selecting the unassigned bus node with the largest feeder-bus component of direct-route cost. The route is then expanded by including unassigned bus nodes based on the criterion of maximum savings. The savings from including unassigned bus node i in route k is estimated by Equation 4:

$$SAVTC_i^k = TC_i^j - \{C_{sj}q_i + 2(\lambda_0\lambda_w)^{1/2}[(L_k^a Q_k^a)^{1/2} - (L_k Q_k)^{1/2}] + (\lambda_r/2U)(L_k^a Q_k^a - L_k Q_k)\} \quad (4)$$

where L_k^a *and* Q_k^a are the route length and demand after the inclusion of bus node i.

Equation 4 provides a correct measure of the savings in transportation cost from including bus node i in route k only if the rail station that minimizes the direct cost for i is the same as that to which route k is linked. If this is not the case, Equation 4 needs to be modified. This is done by using the "modified distance" concept proposed by Tillman and Cain *(23)*. The initial algorithm continues to expand the emerging route until there is no cost savings from any further expansion or until any savings would result in a violation of the constraint on route length. If the emerging route cannot be expanded, a new route is initiated.

The improvement procedures attempt to correct two limitations of the initial algorithm. First, since the initial algorithm builds routes sequentially, the order of bus nodes on any given route may be suboptimal. Any reduction in route length would reduce operator and user riding costs. Therefore, the first improvement procedure attempts to reduce the length of each route by solving the TSP. This is done using the 2-optimal procedures proposed by Lin *(24)*. The second limitation of the initial algorithm is that a bus node assigned in an early stage is not reassigned at a later stage as new routes are developed. The second improvement procedure attempts to correct this problem by displacing a bus node from its current route to another route if it results in a reduction of total cost.

ANALYSIS

The purpose of the following analysis is twofold. First, it is shown that the proposed heuristic method provides reasonable feeder-bus networks. At this point the solutions provided by the method relative to optimality cannot be evaluated. To do so would require the optimal solution for one or more test problems or a "good" lower bound on the optimal solution. Given the complexity of the FBNDP and the computational requirement of the mathematical programming model, neither of these can be accomplished at this stage. There is currently no adequate method for deriving useful lower bounds for vehicle-routing problems. The validation of the proposed heuristic method for the FBNDP was done by first comparing various measures of the network structure with those observed in real-life transit systems. Then the results provided by the proposed model were evaluated by comparing them with those developed manually by five transportation planners. The overall results of this comparison show that the proposed model designs bus networks that are superior to manually developed networks, even for small problems. The advantage of the proposed model is found to be most significant when there are differences in the demand generated at various bus stops. It is important to state that the heuristic model can be used to provide an initial feeder-bus network that may be improved incrementally by an experienced transit planner.

The second and perhaps more important purpose of this analysis is to show the capabilities of the proposed model as a strategic planning tool for the design of a feeder-bus network. In this function, the model is used for answering "what if" questions, that is, for providing responses to changes in the system. The changes to be considered represent new transit design and operating policies as well as changes in the environment in which the transit agency operates, that is, changes that are beyond the control of the agency. These changes are represented by several test cases. The responses of the model are evaluated by comparing the network structure under each test case with that obtained for the base case.

A base case is defined in which the demand is constant over the entire service area. This avoids the irregularities in the solution that would be caused by variable demand and enables the inspection and identification of the changes in network structure that take place under the new conditions. The assessment of the model's responses is done based on systemwide measures such as vehicle miles of travel, passenger miles of travel, operator cost, user cost, and others. In each test case the changes in systemwide measures to be expected under that scenario can be specified. The model's responses are then evaluated relative to the expected changes.

The test problem for this analysis includes 55 bus stops and 4 rail stations, covering a service area of 2×2.5 mi^2. The test problem was designed so that the bus stop density is 11 stops per square mile with demand density of 2,200 passengers per square mile. Such a demand density is typical for urban areas *(25)*. Table 2 provides the bus stop locations and demands for the base case. The locations of the rail stations are given in

TABLE 2 BUS STOP LOCATIONS AND DEMANDS

Bus Stop No.	X-Coordinate[a]	Y-Coordinate[a]	Demand (trips/hour)
1	30	234	200
2	62	235	200
3	119	250	200
4	182	249	200
5	134	228	200
6	163	230	200
7	115	222	200
8	87	215	200
9	24	203	200
10	60	193	200
11	125	197	200
12	150	210	200
13	183	196	200
14	108	186	200
15	85	177	200
16	37	169	200
17	130	173	200
18	185	164	200
19	12	163	200
20	67	153	200
21	105	157	200
22	123	152	200
23	32	133	200
24	55	135	200
25	73	135	200
26	89	144	200
27	142	137	200
28	161	143	200
29	18	107	200
30	46	107	200
31	107	115	200
32	147	117	200
33	172	124	200
34	31	95	200
35	91	103	200
36	113	99	200
37	13	80	200
38	66	87	200
39	83	83	200
40	141	92	200
41	167	97	200
42	67	65	200
43	122	75	200
44	150	67	200
45	177	68	200
46	95	59	200
47	17	47	200
48	47	43	200
49	130	48	200
50	71	35	200
51	108	33	200
52	169	35	200
53	13	25	200
54	35	17	200
55	63	7	200

[a] The distances are specified in hundreds of miles.

TABLE 3 RAIL STATION LOCATIONS

Rail Station No.	X-Coordinate[a]	Y-Coordinate[a]
56 (Destination)	42	72
57	78	116
58	123	137
59	160	178

[a] The distances are specified in hundreds of miles.

Table 3. Rail station 56 represents the destination under the M-to-1 demand pattern. The parameters for the base case are given in Table 4. The bus operating cost is based on a cost per seat mile of $0.06 and bus capacity of 50 seats *(26)*. Typical values from the literature are used for riding-time and wait-time costs as well as for bus capacity and operating speed.

Table 5 provides the base-case solution, which is illustrated in Figure 5. The feeder-bus network includes 16 routes with service frequency ranging from 16 to almost 26 trips per hour (headways in the range of 2.3 to 3.75 min). The average headway is 2.8 min and the total route distance per square mile is 3.0. These values are within the range observed in real-life transit systems. The base-case solution includes 703 vehicle-mi of travel and 7,926 passenger-mi of travel and provides a system cost of $6,000/hr, consisting of 10 percent rail cost, 35 percent bus operating cost, and 55 percent bus user cost.

The analysis includes the following test cases:

1. Change in design objective,
2. Introduction of demand variability,
3. Change in vehicle capacity,
4. Change in labor and fuel cost,
5. Opening of a new rail station, and
6. Closing of a rail station.

In the first test case the objective function is defined as that of minimizing bus user cost. The purpose is not to suggest the minimization of user cost as an appropriate objective but to evaluate the model's response to changes in the design objective. Under the new objective, a reduction in bus user cost and an increase in bus operator cost should be expected relative to the base-case network. The second test case introduces variable demand. Using Monte Carlo simulation, the demand at bus stops is generated from a uniform distribution with a mean of 200 passengers per hour. Under variable demand a feeder-bus network with larger differences in route length and route frequency should be expected. In the third test case, a change is considered in the type of vehicle used; a standard bus with seating capacity of 50 is changed to an articulated bus with a seating capacity of 95. This change in vehicle type reduces operating cost per seat mile because of higher labor productivity. Assuming that labor cost constitutes 60 percent of total operating cost and fuel and maintenance cost represents 40 percent, the operating cost per vehicle mile for an articulated bus is estimated at $3.30. As vehicle capacity increases, a feeder-bus network with fewer but longer routes, operated at lower frequencies, should be expected. The combined effect of longer routes and lower frequencies should result in higher user costs. The change in total vehicle miles of travel and consequently in total operator cost depends on the spatial distribution of demand and cannot be determined a priori.

The operating and maintenance (O&M) cost of a typical transit service consists of 60 percent labor cost, 25 percent fuel cost, and 15 percent administrative cost *(27)*. In the fourth test an increase of 40 percent in both labor and fuel costs is considered. This translates into an increase of $1.02 per vehicle mile. With an increase in operating cost per vehicle mile, operators can be expected to restructure the bus network to reduce vehicle miles of travel. This can be achieved by a

TABLE 4 BASE-CASE PARAMETERS

Descriptions	Units	Value
Bus Operating Cost	$/veh.-mile	3.0
Rail User Cost	$/pass.-mile	0.15
Riding Time Cost	$/pass.-hr.	4.0
Waiting Time Cost	$/pass.-hr.	8.0
Max. Allowable Route Length	mile	2.5
Bus Capacity	seat	50
Bus Operating Speed	mile/hr.	20
Maximum Available Seat-Hours	seat-hrs.	5500

TABLE 5 MODEL SOLUTION FOR THE BASE CASE

Route No.	Route Structure	Route Demand (passengers)	Route Length (miles)	Route Frequency (trips/hr.)
1	1 2 10 24 57	800	1.62	18.14
2	51 46 42 38 56	800	1.08	22.22
3	12 13 18 59	600	0.97	20.31
4	9 16 57	400	1.03	16.09
5	4 6 59	400	0.79	18.37
6	41 32 58	400	0.59	21.26
7	21 26 58	400	0.56	21.82
8	3 5 7 11 17 59	1000	1.29	22.73
9	40 39 35 57	600	0.99	20.10
10	19 23 30 29 34 56	1000	1.37	22.06
11	8 14 15 20 25 57	1000	1.30	22.65
12	52 45 44 49 50 56	1000	1.96	20.00
13	43 36 31 58	600	0.70	23.90
14	55 54 48 56	600	0.88	21.32
15	33 28 27 22 58	800	0.81	25.66
16	53 47 37 56	600	0.85	21.69

Systemwide Performance Measures

NR	=	16	FS = 54	TPM = 7926	BWC	= 2075
AF	=	21.1	TRL = 16.8	TSC = 6033	BRC	= 1255
RSH	=	2706	TVM = 703	RC = 592	BUC	= 3330
					BOC	= 2110

NR = no. of routes
RSH = required seat-hrs. (seat-hrs.)
TRL = total route length (miles)
TPM = total pass.-miles (pass.-miles)
RC = rail cost ($/hr.)
BRC = riding cost on bus ($/hr.)
BOC = bus operating cost ($/hr.)
AF = average frequency (trips/hr.)
FS = fleet size (no. of vehicles)
TVM = total veh.-miles (veh.-miles)
BWC = waiting cost for bus ($/hr.)
TSC = total system cost ($/hr.)
BUC = bus user cost ($/hr.)

reduction in total route length, reduction in operating frequencies, or both. The last two test cases consider changes in the rail network. First, adding a new rail station would improve the accessibility to the rail system and should be expected to reduce both operator and user costs for the feeder-bus system. These cost reductions would result from a reduction in total route length. The closing of a rail station should be expected to have the opposite effect, that is, an increase in vehicle miles of travel and passenger miles of travel. It would increase travel time in the rail system for passengers who previously boarded at the closed station.

The results of the analysis are summarized in Tables 6 and 7. Table 6 shows the responses of the model in terms of systemwide operation measures, whereas Table 7 provides the re-

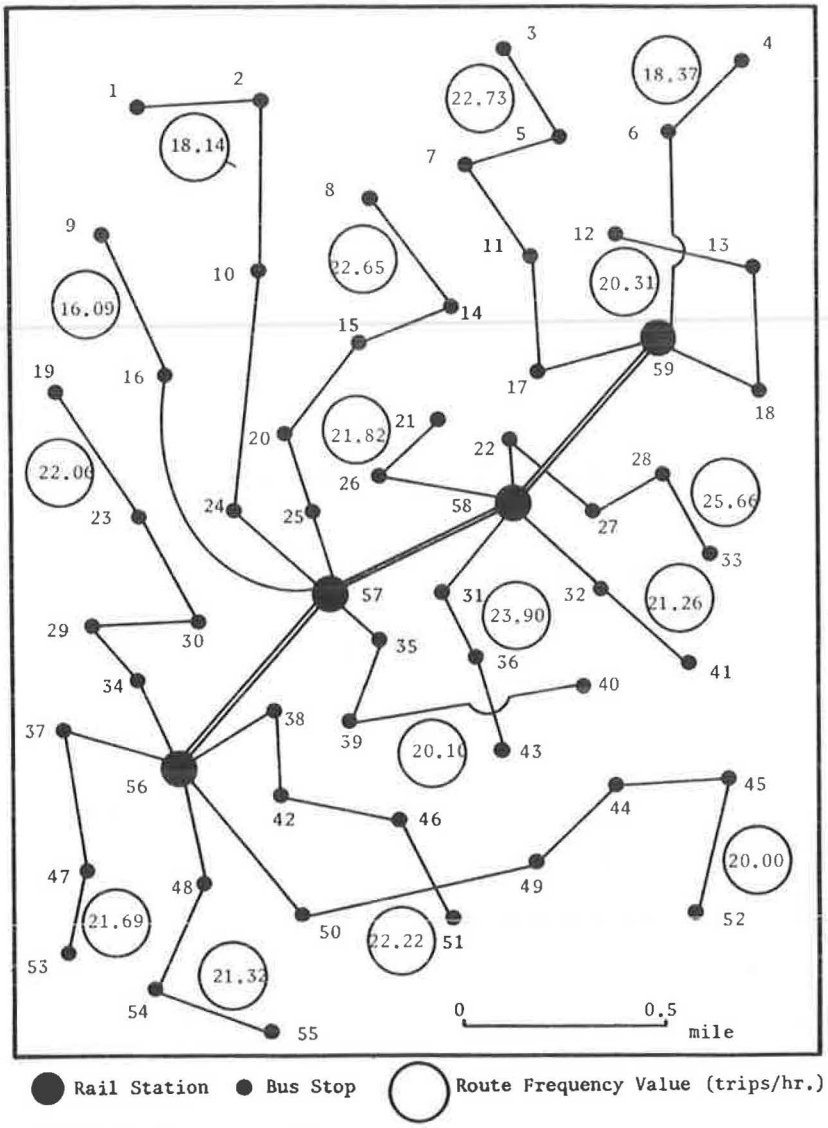

FIGURE 5 Base-case solution.

TABLE 6 SYSTEMWIDE OPERATION MEASURES

Case	Number of Routes	Fleet Size	Average Frequency	% change	Total Vehicle -Miles	% change	Total Passenger Miles	% change
Base	16	54	21.1	--	703	--	7926	--
Min. Bus User Cost	15	59	22.6	+7.1	761	+8.3	7654	-3.4
Variable Demand	15	53	20.4	-3.3	687	-2.3	7840	-1.1
Inc. in Veh. Capacity	15	50	20.3	-3.8	652	-7.3	7832	-1.2
Inc. in Optr. Cost	20	54	21.0	-0.5	700	-0.4	6952	-12.3
Add a Rail Station	15	52	21.7	+2.8	682	-3.0	7548	-4.8
Close a Station	14	54	21.4	+1.4	707	+0.6	8482	+7.0
M-to-M Demand	15	54	20.8	-1.4	704	+0.2	7950	+0.3

TABLE 7 SYSTEMWIDE COST COMPONENTS

Case	Rail Cost	% change	Bus User Cost	% change	Bus Optr. Cost	% change	Total System Cost	% change
Base	592	--	3330	--	2110	--	6033	--
Min. Bus User Cost	822	+38.9	3186	-4.3	2283	+8.2	6291	+4.3
Variable Demand	777	+31.3	3314	-0.5	2060	-2.4	6151	+2.0
Inc. in Veh.Capacity	727	+22.8	3387	+1.7	2152	+2.0	6266	+3.9
Inc. in Optr. Cost	1110	+87.5	3129	-6.0	2104	-0.3	6343	+5.1
Add a Rail Station	755	+27.5	3283	-1.4	2046	-3.0	6084	+0.9
Close a Station	673	+13.7	3440	+3.3	2120	+0.5	6234	+3.3
M-to-M Analysis	973	+64.4	3396	+2.0	2112	+0.1	6481	+7.4

TABLE 8 SOLUTION FOR M-TO-M DEMAND PATTERN

Route No.	Route Structure	Route Demand (passengers)	Route Length (miles)	Route Frequency (trips/hr.)
1	54 53 47 37 34 56	1000	1.28	22.83
2	55 50 48 56	600	0.84	21.84
3	19 23 29 30 57	800	1.27	20.51
4	9 16 24 57	600	1.05	19.55
5	1 10 20 25 57	700	1.30	18.94
6	43 39 35 57	500	0.80	20.45
7	51 46 42 38 57	800	1.11	21.91
8	2 10 15 26 58	700	1.40	18.28
9	7 8 14 21 58	800	1.21	21.02
10	3 5 12 11 17 22 58	1100	1.41	22.84
11	4 6 12 59	500	0.84	19.89
12	13 18 28 27 59	800	1.29	20.36
13	41 33 32 58	600	0.85	21.74
14	52 45 44 40 58	800	1.36	19.80
15	49 43 36 31 58	700	0.98	21.81

Systemwide Performance Measures

NR	=	15	FS = 54	TPM = 7850	BWC	=	2112
AF	=	20.8	TRL = 17.0	TSC = 6481	BRC	=	1285
RSH	=	2707	TVM = 704	RC = 973	BUC	=	3396
					BOC	=	2112

NR = no. of routes
RSH = required seat-hrs. (seat-hrs.)
TRL = total route length (miles)
TPM = total pass.-miles (pass.-miles)
RC = rail cost ($/hr.)
BRC = riding cost on bus ($/hr.)
BOC = bus operating cost ($/hr.)
AF = average frequency (trips/hr.)
FS = fleet size (no. of vehicles)
TVM = total veh.-miles (veh.-miles)
BWC = waiting cost for bus ($/hr.)
TSC = total system cost ($/hr.)
BUC = bus user cost ($/hr.)

sponses in terms of cost measures. In all the cases, the proposed model provides consistent and reasonable responses to the various "what if" questions. Under the objective of minimizing bus user cost, user cost has decreased by 4.3 percent. However, this required an increase of 8.2 percent in bus operator cost and 38.9 percent in rail user cost, with a net effect of 4.3 percent increase in total system cost. The variability of demand increased the variations in route length and operating frequencies. The operating frequencies under variable demand range between 11.6 and 28.4 buses per hour compared with a range of 16.1 to 25.7 in the base-case network. The increase in bus capacity resulted in a reduction in the number of routes and

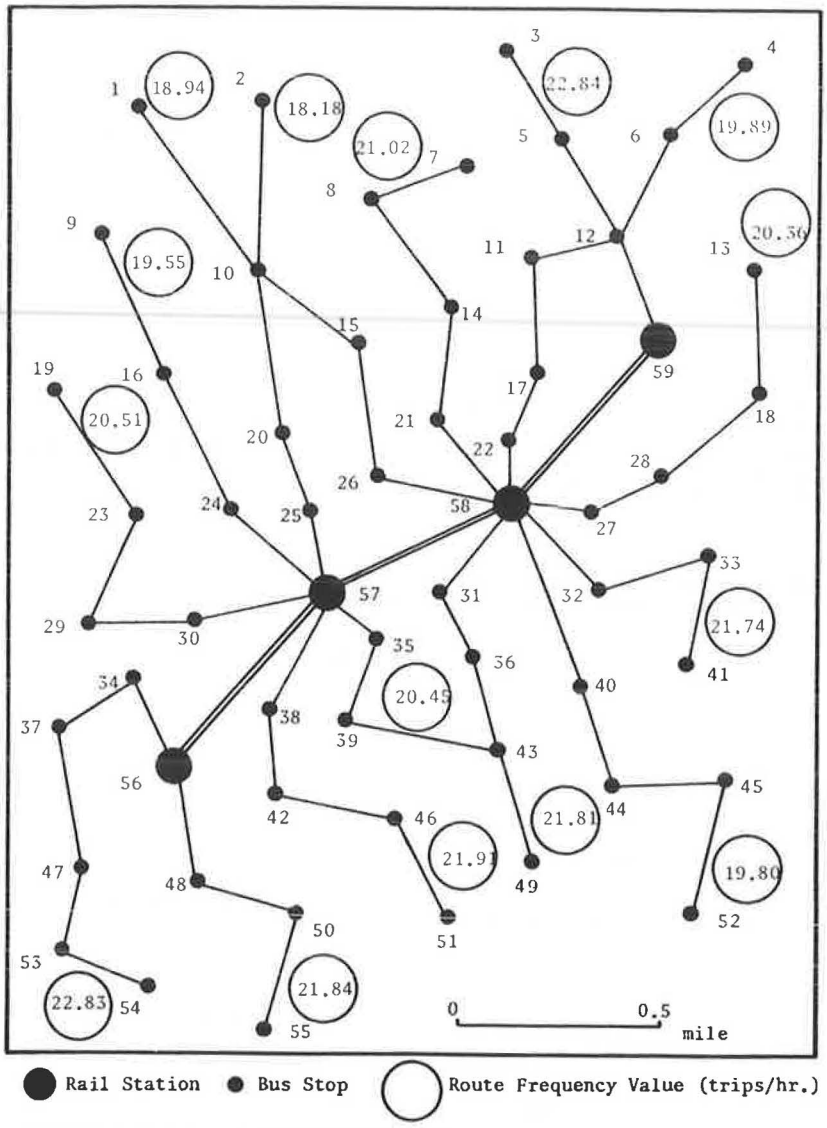

FIGURE 6 Solution for M-to-M demand.

increase in average route length, with an associated decrease in average operating frequency from 21.1 trips per hour to 20.3 trips per hour. The decrease in operating frequencies increased total user cost and resulted in a 3.9 percent increase in total system cost. The increase in labor and fuel costs results in a marginal reduction in bus vehicle miles of travel and in a significant increase in rail passenger miles. The opening of a new rail station decreases both vehicle miles and passenger miles of travel in the bus system. Consequently, it results in the expected reduction in both bus operator and user costs. The closing of a rail station has the opposite effect.

In the last part of the analysis the FBNDP is considered under the M-to-M demand pattern. This part of the analysis is used to evaluate the model's responses to changes in demand pattern. The demand at each bus stop is divided equally among the four destinations (rail stations). As discussed earlier, under the M-to-M demand pattern a bus stop may be served by multiple feeder-bus routes. Consequently, a bus network with more circuitous routes, greater total route length, higher vehicle miles of travel (therefore higher operator cost), higher total passenger miles of travel, and lower operating frequencies should be expected. The combined effect of lower frequencies and higher passenger miles of travel would increase bus user cost.

Table 8 shows the solution for the M-to-M base case, which is illustrated in Figure 6. It can be noted that some of the bus stops are served by multiple routes. The results of Table 6 show the expected reduction in average operating frequency as well as the increases in total vehicle miles and passenger miles of travel relative to the base-case network under the M-to-1 demand pattern. Table 7 shows the increases expected in bus operator cost and bus user cost.

SUMMARY AND CONCLUSION

The U.S. transit industry faces financial difficulties. Among the strategies suggested for improving the financial conditions of transit agencies, the development of better-integrated intermodal systems has the advantage of potentially achieving both

cost reduction and improvement in service quality. A network-optimization based methodology for the design of an integrated feeder-bus–rail transit system has been presented. The FBNDP is defined as that of designing a set of feeder-bus routes and determining the service frequency on each route so as to minimize the sum of operator and user costs. The FBNDP is considered under two demand patterns—M-to-1 and M-to-M. A conceptual representation of the M-to-1 FBNDP as a spanning-tree network is presented. Based on this representation, the M-to-1 FBNDP is formulated as a mathematical programming problem. It is shown that the spanning-tree network representation can be generalized to the M-to-M FBNDP. Consequently, the FBNDP under the M-to-M demand pattern can be solved using a model for the M-to-1 case.

The FBNDP is a large and complex routing problem that can be solved only heuristically. The proposed heuristic method for the FBNDP is based on the savings approach while generalizing it to represent operating frequency. A comparison shows that the solutions provided by the proposed model were generally superior to manually designed networks. The advantage of the model's solution is found to increase under variable demand. The proposed model may be viewed as a way of deriving an initial feeder-bus network that may be improved incrementally by an experienced transit planner. The proposed heuristic method is not limited to small problems and can efficiently construct networks of reasonable size. It is shown to be capable of providing consistent and reasonable answers to "what if" questions.

As stated earlier, the FBNDP can be viewed as the problem of achieving the optimal balance between operator and user costs. This analysis indicates (see Table 6) that changes that increase the relative weight of operator cost result in a feeder-bus network with lower total vehicle miles of travel, achieved by operating more but less circuitous routes at lower frequencies. Changes that increase the relative weight of user cost result in feeder-bus networks with higher vehicle miles of travel. The reduction in user cost is often achieved by operating fewer but more circuitous routes at higher frequencies, thereby reducing wait time at bus stops. This operating strategy reduces user cost because unit wait-time cost is higher than the unit cost of riding time.

REFERENCES

1. J. R. Bonnell. Transit's Growing Financial Crisis. *Traffic Quarterly,* Vol. 35, 1981, p. 541.
2. *Transit Facts Book.* American Public Transit Association, Washington, D.C., 1985.
3. *National Urban Mass Transportation Statistics.* U.S. Department of Transportation, 1982.
4. M. Pikarsky. Trends and Options for Increasing the Role of the Private Sector in Urban Transportation. Presented at the Conference on Mass Transit in Chicago Metropolitan Region, Northwestern University, 1983.
5. L. J. Pignataro. Future Directions for Public Transportation. In *Transportation Research Record 793,* TRB, National Research Council, Washington, D.C., 1981, pp. 5–11.
5. J. A. Dunn, Jr. Coordination of Urban Transit Services: The German Model. *Transportation,* Vol. 9, 1980, pp. 33–43.
7. R. M. Stanger and V. R. Vuchic. The Design of Bus-Rail Transit Facilities. *Transit Journal,* Vol. 5, No. 4, 1979, pp. 61–72.
8. B. E. Sullivan and C. Lovelock. Intermodal Integration. In *Special Report 182: Light Rail Transit: Planning and Technology,* TRB, National Research Council, Washington, D.C., 1978, pp. 170–171.
9. *Announcement, Office of Technical Assistance, University Research and Training Program.* UMTA, U.S. Department of Transportation, 1984.
10. L. A. Silman, Z. Brazily, and U. Passy. Planning the Route System for Urban Buses. *Computers and Operations Research,* Vol. 1, 1974, pp. 201–211.
11. 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. 49–59.
12. W. Lampkin and P. D. Saalman. The Design of Routes, Service Frequencies and Schedules for Municipal Bus Undertaking: A Case Study. *Operations Research Quarterly,* Vol. 18, 1967, pp. 375–397.
13. A. N. Bansal. Optimization of Bus Route Network for a Fixed Spatial Distribution. In *Scientific Management of Transportation Systems* (N. S. Jaiswal, ed.), North Holland Publishing Co., Amsterdam, 1981.
14. C. E. Mandl. Evaluation and Optimization of Urban Public Transportation Networks. *Europe Journal of Operations Research,* Vol. 5, 1980, pp. 396–404.
15. J. D. Hsu and V. H. Surti. Decomposition Approach to Bus Network Design. *Transportation Engineering Journal,* ASCE, Vol. 103, No. 4, 1977, pp. 447–460.
16. S. A. Scheele. Supply Model for Public Transit Service. *Transportation Research,* Vol. 14B, 1980, pp. 133–148.
17. B. P. Lingaraj et al. An Optimization Model for Determining Headways for Transit Routes. *Transportation Planning and Technologies,* Vol. 3, 1977, pp. 81–90.
18. L. J. S. Lesley. Optimum Bus Stop Spacing, Part I. *Traffic Engineering and Control,* Vol. 17, 1976, pp. 399–401.
19. V. R. Vuchic. Rapid Transit Interstation Spacings for Maximum Number of Passengers. *Transportation Science,* Vol. 3, No. 3, 1969, pp. 214–232.
20. G. K. Kuah. *The Feeder Bus Route Design Problem.* Ph.D. dissertation. Department of Civil Engineering, University of Maryland, College Park, 1986.
21. L. Bodin, B. Golden, A. Assad, and M. Ball. Routing and Scheduling of Vehicles and Crews. *Computers and Operations Research,* Vol. 2, No. 2, 1983, pp. 63–210.
22. J. Perl. Multidepot Routing Allocation Problem. *American Journal of Mathematical and Management Sciences* (forthcoming).
23. F. Tillman and T. Cain. An Upper Bound Algorithm for the Single and Multiple Terminal Delivery Problem. *Management Science,* Vol. 18, No. 11, 1972, pp. 669–682.
24. S. Lin. Computer Solutions of the Traveling Salesman Problem. *Bell System Technical Journal,* Vol. 44, 1965, pp. 2245–2269.
25. F. V. Webster and P. H. Bly. Public Transportation and the Planning of Residential Area. In *Urban Planning and Public Transport* (R. Cressell, ed.), The Construction Press, Harlow, Essex, England, 1984.
26. J. E. Anderson. Optimization of Transit System Characteristics. *Journal of Advanced Transportation,* Vol. 18, No. 1, 1984, pp. 77–111.
27. J. O. Jansson. A Simple Bus Line Model for Optimization of Service Frequency and Bus Size. *Journal of Transport Economics and Policy,* Vol. 14, No. 1, 1980, pp. 53–80.

Application of a Simultaneous Transportation Equilibrium Model to Intercity Passenger Travel in Egypt

K. NABIL A. SAFWAT

Safwat and Magnanti have developed a combined trip-generation, trip-distribution, modal-split, and trip-assignment model that can predict demand and performance levels on large-scale transportation networks simultaneously, that is, a Simultaneous Transportation Equilibrium Model (STEM). The major objective in this paper is to assess the behavioral applicability of the STEM methodology by using it to analyze intercity passenger travel in Egypt. Though results were greatly influenced by the misspecification of the trip-distribution model (particularly its attractiveness measure) and the existence of severe fleet capacity constraints (particularly along the Cairo-Banha-Tanta corridor in the middle Delta region), there were strong indications that the approach was able to predict rational behavioral responses of users to policy changes in the system. Appropriate modifications to current specifications, which can easily be incorporated within the STEM framework, were suggested. Computational issues of the application are addressed in detail in a companion paper by Safwat in this Record.

Modeling of transportation systems must invariably balance behavioral richness and computational tractability. Safwat and Magnanti *(1)* developed a combined trip-generation, trip-distribution, modal-split, and trip-assignment model that can predict demand and performance levels on large-scale transportation networks simultaneously, that is, a Simultaneous Transportation Equilibrium Model (STEM). Moavenzadeh et al. *(2)* developed a methodology for intercity transportation planning in Egypt with the STEM as one of its central components.

In this paper, the major objective is to assess the behavioral applicability of the STEM methodology. The main concern is assessing the ability to represent observed behavior and to predict behavioral changes. Computational issues of the application are addressed in detail in a companion paper by Safwat in this Record.

To achieve this objective, the STEM approach was applied to a real-world transportation network, the Egyptian intercity transportation system. In the next section, a STEM is briefly introduced. Next, the major issues related to intercity passenger travel in Egypt are presented. Then the focus is on the behavioral modeling of passenger transport on the system and the design of a case study. The results of the application are discussed last, followed by a summary and conclusions.

Department of Urban and Regional Planning, Texas A&M University System, College Station, Tex. 77843.

A STEM

In this section the behavioral assumptions and formulations of the components of a STEM are briefly introduced. For a detailed description, the reader is referred to the paper by Safwat and Magnanti *(1)*. It should be emphasized at this point that although the assumptions of individual components of the STEM are not necessarily new, it is their internally consistent combination and simultaneous prediction that distinguishes a STEM formulation from other approaches.

Trip distribution is given by a logit model whose measured utility functions include the average minimum origin-destination (O-D) travel costs as variables with linear parameters. That is,

$$T_{ij} = G_i \frac{\exp(-\theta U_{ij} + A_j)}{\sum_{k \varepsilon D_i} \exp(-U_{ik} + A_k)} \quad \text{for all } ij \varepsilon R$$

where

T_{ij} = number of trips distributed from origin i to destination j,
G_i = number of trips generated from origin i,
U_{ij} = average minimum travel cost from i to j,
A_j = attractiveness measure of destination j,
θ = parameter that measures the sensitivity of travelers to O-D travel costs,
D_i = set of destinations accessible from origin i, and
R = set of all O-D pairs ij.

Trip generation is given by any general function as long as it is linearly dependent on the system's performance through an accessibility measure based on the random utility theory of travel behavior (i.e., the expected maximum utility of travel). That is,

$$G_i = \alpha S_i + E_i \quad \text{for all } i \varepsilon I$$

$$S_i = \max [0, \ln \sum_{j \varepsilon D_i} \exp(-\theta U_{ij} + A_j)] \quad \text{for all } i \varepsilon I$$

where

S_i = accessibility variable that measures the expected maximum utility of travel for travelers at origin i,

E_i = given minimum trip generation from i due to socioeconomic and land use forces,
α = parameter that measures the sensitivity of travelers to accessibility of the system, and
I = set of origins in the network.

Modal split and traffic assignment are simultaneously user optimized [the STEM framework allows for modal split to be given by a logit model or to be system optimized together with traffic assignment *(3)*].

In the following sections, the issues involved in the application of this STEM to intercity passenger travel in Egypt are addressed.

INTERCITY PASSENGER TRAVEL IN EGYPT

For more details the reader is referred to *Egypt National Transport Study,* Phase 1 *(4)* and Phase 2 *(5)*; papers by Safwat *(3, 6)*; and *Egypt Intercity Transport Project (7).*

Existing modes for intercity passenger travel in Egypt include private car, taxi, and bus on a highway network with a total length of about 28,500 km, of which 15,000 km is paved. There are also different types of service on the railway network, which is owned by the Egyptian Railway Authority and has a total route length of about 3,260 km, excluding the Sinai lines.

Historically, rail was the dominant mode. During the last decade, however, rail has been quickly losing its position in favor of an increasingly competitive mode, the taxi. Table 1 shows the passenger kilometers produced in 1974 and 1979 by each mode in the system. In 1974 the system produced 23.5 billion passenger-km, of which 55 percent was produced by rail, 22.5 percent by taxi, 14.5 percent by bus, and 8 percent by private car. In 1979 the system produced 34.5 billion passenger-km (i.e., an increase of 12 billion passenger-km compared with 1974); only 0.9 billion passenger-km of this increase was absorbed by rail, whereas 6 billion passenger-km (i.e., one-half of the increase) was attracted to the taxi. As a result, the rail share dropped from 55 to only 40 percent and the biggest increase was for the taxi, from 22.5 to 33 percent. In terms of the number of passenger trips in 1979 (Table 2), the taxi had the highest share, 37 percent, followed by rail at 30 percent only. Netherlands Engineering Consultants (NEDECO) predicted that this trend was expected to continue for at least five more years *(5)*.

TABLE 1 PASSENGER TRANSPORT, 1974 AND 1979

	Passenger Kilometers (billions)			
	1974		1979	
Mode	No.	Percent	No.	Percent
Private car	1.9	8.0	3.0	9.0
Taxi	5.3	22.5	11.3	33.0
Public bus	3.4	14.5	6.4	18.0
Railway	12.9	55.0	13.8	40.0
Total	23.5		34.5	

This trend is strikingly counterintuitive because rail has extremely low tariffs compared with other modes. Table 3 shows revenues generated and financial costs incurred in millieemes (MM) per passenger-km by mode. In 1979, the rail revenues were 2.37 MM per passenger-km compared with 12.40 MM for taxi. The severity of the problem facing rail is evident given that the financial costs incurred were more than twice the generated revenues, which indicates a large deficit (see Table 3). The problem is becoming even more serious over time because of the rapid increase of demand, estimated at about 9 percent annually during 1974–1979.

The major constraint on the Egyptian Railway is a severe limitation on fleet capacity. NEDECO *(5)* found that between 20 and 30 percent of all rolling stock registered as book stock in 1979 was beyond repair. Officials added that by 1982, about 82 percent was estimated to be beyond repair. The lack of sufficient tractive power was believed to be the main source of limitation. In July 1976 only 889 trains ran and 2,699 were cancelled, 2,614 of them for lack of locomotives *(4)*.

Therefore, in the modeling and analysis of intercity passenger travel in Egypt, fleet capacity constraints must be accounted for in a satisfactory fashion. This is addressed in the next section.

BEHAVIORAL MODELING OF THE EGYPTIAN INTERCITY PASSENGER TRANSPORT SYSTEM

Modeling the Egyptian system to address the major issues indicated in the preceding section involves four major tasks: (a) the definition of passenger type–choice set mapping, (b) the specification of modal split and traffic assignment behavior and network representation, (c) the development of performance functions, and (d) the calibration of demand functions.

Passenger Type–Choice Set Mapping

Passengers may be classified into three types according to their income level: high, middle, and low. Transportation services may also be classified according to their level-of-service attributes such as travel time, tariff, comfort, safety, and so on. In

TABLE 2 PASSENGER TRANSPORT, 1979 *(5)*

	Passenger Trips (millions)		Passengers Kilometers (billions)		Avg Distance
Mode	No.	Percent	No.	Percent	(km)
Private car	47.3	8.0	3.0	9.0	63
Taxi	215.8	37.0	11.3	33.0	52
Public bus	146.3	25.0	6.4	18.0	44
Railway	174.2	30.0	13.8	40.0	79
Total	583.6		34.5		59

TABLE 3 REVENUES AND COSTS OF INTERCITY PASSENGER TRANSPORT IN EGYPT, 1979

Mode	Revenues (milliemes/ passenger kilometer)	Financial Costs (milliemes/ passenger kilometer)
Railway	2.37	6.41
Air-conditioned	4.94	19.9
Non-air-conditioned	2.14	5.22
Bus (52 seats)	6.65	6.65
Taxi	12.40	12.40

fact, service types on the Egyptian system are designed so that each is most suitable for a particular income level. Available services in the system may be categorized as taxi, lux bus (an aggregation of several types of bus service that are considered "excellent," "good," or "sufficient"), normal bus (an aggregation of two types of bus service—"moderate" and "poor"), diesel train [includes first- and second-class air-conditioned (AC) service], express train (includes all types of service on rail ranging between first- and second-class AC to second- and third-class non-AC), and local train (exclusively third-class non-AC service).

In the analysis, it is assumed that a "mapping" exists between passenger income types and transportation service types (choice sets). The required mapping, based on a quality index associated with each service type [from NEDECO *(5)* and discussions with Egyptian transport experts], is defined in Figure 1.

In Figure 1, solid lines indicate main modes and dashed lines indicate modes that may be selected by the particular passenger type whenever capacity is limited on the main modes. The mapping shown is further refined within the train types. The complete mapping may be summarized as follows:

1. Low-income passengers may select among local train, express train (third class only), normal bus, and taxi;
2. Middle-income passengers may select among express train (second class AC and non-AC), taxi, lux bus, and diesel train (second class AC only); and
3. High-income passengers may select among automobile, diesel, lux bus, express train (first class AC only), and taxi.

In this paper, analysis is focused on low-income passengers only, because they represent the majority of users in the system (about 80 percent of train users, 75 percent of bus users, and more than 65 percent of systemwide users have low incomes). Excluding other passenger types is expected to influence the results. Proper definition of low-income mapping, however, should overcome some of these inaccuracies.

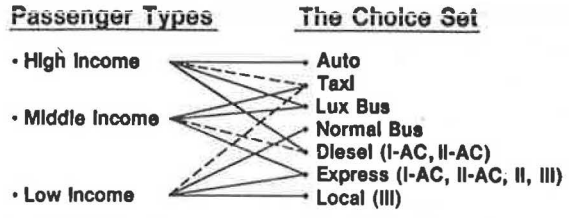

FIGURE 1 Passenger type–choice set mapping.

Multimodal Composed Networks

Because of the existence of transfer between transport modes in the middle of any given trip and because passengers travel as individuals or in small groups, modal split and traffic assignment on the Egyptian system are assumed to occur simultaneously in accordance with the user optimization principle of user behavior. With these behavioral characteristics (i.e., transfer between modes and user optimization behavior) the network may be best represented as a multimodal composed network. An example of such a composed network is shown in Figure 2.

The network in Figure 2 consists of two modal networks, express train and normal bus, connected through three zonal centroids with loading and unloading links that reflect the allowable types of transfers between these two modes at each of the three zones. For example, the leftmost zone is a destination only, whereas the rightmost one can be origin or destination for both modes.

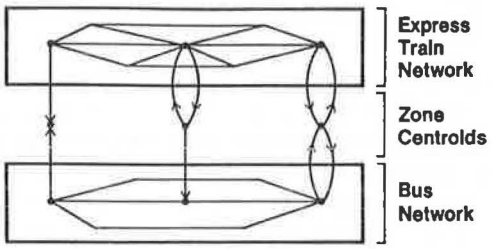

FIGURE 2 Example of multimodal composed network.

In the analysis, four major modal networks were created for express train, local train, taxi, and normal bus. Figure 3 shows one of these modal networks (that for the taxi). There were 24 zones, each represented by its centroid identified with a four-character name that corresponds to the name of its governorate whenever possible, as shown in Table 4. Points of transfer were assumed to be any zone centroid that is accessible to a given mode. For the purpose of analysis, three composed networks were specified. A brief description of each is as follows:

Network	*Modes*	*No. of Links*	*No. of Nodes*
NET1	Express, local trains	244	90
NET2	Express, local trains; normal bus	394	125
NET3	Express, local trains; normal bus and taxi	534	152

Link Performance Functions

System performance may be perceived by users through a set of generalized cost functions. For any given trip, average perceived cost may include travel-time cost, tariff cost, cost of

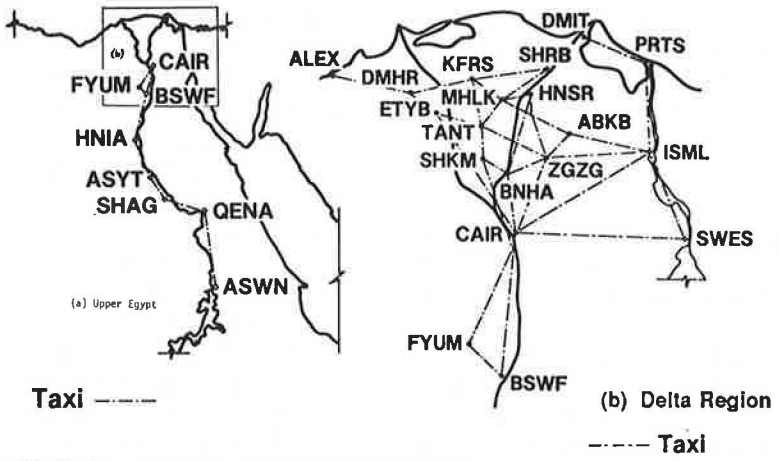

FIGURE 3 Egyptian intercity taxi network.

delay at intermediate nodes, and loading-unloading cost. The average perceived cost depends on several factors, such as system use, design, and operating policies, which may vary among different links and modes on the network. Therefore, performance functions are defined at the level of modal links on the composed networks.

A typical modal link User Perceived Cost (UPC) function, travel-time cost, is the value of time multiplied by the average travel time. The value of time is dependent on the average annual income and was estimated at 0.15078 L.E./hr for the low-income group [based on figures from NEDECO (*5*, Annex II, pp. 3.27 and 3.29)]. Average travel time depends on link free-flow speed (which in turn is a function of link design and classification) and mode type (as reflected through a modal speed factor). The highway classes, railway classes, and modal speed factors were obtained from the Intercity Project *(7)*. The speed factors for taxi and bus were assumed to include delay at intermediate nodes. For express and local trains, delays at intermediate nodes were assumed to depend on the importance of the node as determined by the Intercity Project *(7)*.

As far as traffic congestion is concerned, it is assumed that on the railway it is captured through the definition of "practical speed" on different track classes. On the highway, NEDECO *(5)* calculated the volume/capacity (V/C) ratio for 80 intercity roadway sections; 80 percent of these were found to have V/C ratios less than 0.5. Hence, it is assumed that traffic congestion on the Egyptian intercity highway network may be ignored.

Tariff cost is estimated by multiplying the tariff per unit distance by the link length. Tariff per unit distance was approximated by the Intercity Project *(7)* because the actual tariff structure is distance-dependent.

Loading and unloading costs are the delays multiplied by the value of time. The average loading and unloading delays associated with different modes at different zones in the system were estimated by the Intercity Project *(7)*. These delays were difficult to estimate and the available information may be considered "crude" estimates.

The fleet capacity constraint on the Egyptian system may have been dealt with accurately by introducing a set of mathematical constraints to the optimization formulation. Such constraints, however, are generally nonconcave or nonconvex functions of the vector of flows of all user types on a given mode, and hence the addition of such constraints to the optimization problem will create computational difficulties. In this application, an approximate solution to the problem is suggested by introducing a term in the link UPC function that drives the user cost to a very high value whenever that capacity is exceeded. Similar approximations to deal with the problem of hard link capacity constraints have been suggested by several researchers [Daganzo *(8)* and Hearn *(9)*]. In this application, however, "capacity" refers to the available fleet capacity

TABLE 4 ZONING SYSTEMS

NEDECO Zoning System			Safwat Zoning System[a]
No.	Zone (Governorate) Name	Zone Centroid	
1	Cairo central business district	Cairo-CBD	CAIR
2	Giza	Giza	–
3	Galyubia	Banha	BNHA
4	Sharkia (South)	Zagazig	ZGZG
5	Sharkia (North)	Abu-Kebir	ABKB
6	Dakahlia (East)	Mansoura	MNSR
7	Dakahlia (West)	Sherbin	SHRB
8	Domiat	Domiat	DMIT
9	Port Said	Port Said	PRTS
10	Ismailia	Ismailia	ISML
11	Swes	Swes	SWES
12	Minufia	Shebin El-Kom	SHKM
13	Gharbiya (South)	Tanta	TANT
14	Gharbiya (North)	Mahalla Kubra	MHLK
15	Kafr El-Shaikh	Kafr El-Shaikh	KFRS
16	Beheita (South)	Etay Baroud	ETYB
17	Beheira (North)	Damanhour	DMHR
18	Alexandria	Alexandria	ALEX
19	Western Desert	Marsa Matrah	–
20	Sinai	E-Swes Tunnel	–
21	El-Fayum	Fayum	FYUM
22	Bani-Swaif	Bani-Swaif	BSWF
23	El-Minia	Minia	MNIA
24	Asyut	Ayut	ASYT
25	New Valley	New Valley	–
26	Sohag	Sohag	SHAG
27	Qena	Qena	QENA
28	Aswan	Aswan	ASWN
29	Red Sea Coast	Port Safaga	–

[a]Abbreviations of zone centroid names.

for a given user type on a given modal link, which may be calculated given the passenger type–choice set mapping, train composition, load factors, and daily schedule *(3)*.

Demand Functions

As indicated earlier, trip generation is given by a general linear model and trip distribution is given by a logit model. The calibration of these demand functions was based on data from NEDECO *(5)*.

The main types of data required for calibration are O-D matrices of trips and costs and socioeconomic data of passengers and zones. Available data include O-D matrices for automobile, taxi, bus, and rail trips in 1979; O-D distance matrix; and zonal population divided into urban and rural *(5)*. The O-D matrices of trips include synthesized items for highway modes and estimated values for about 50 percent of railway passenger movements based on a sample survey. This weakens their reliability. In addition, it is clear that socioeconomic data are very limited. Furthermore, O-D cost matrices are not available.

In the Intercity Project *(7)* a trip-distribution model was specified as follows:

$$T_{ij} = G_i \frac{\exp(-\theta_0 d_{ij} + \theta_1 \ln D_j)}{\sum_k \exp(-\theta_0 d_{ik} + \theta_1 \ln D_k)}$$

where

d_{ij} = distance between zones i and j (km),
D_j = number of trips attracted to j (per day), and
θ_0, θ_1 = parameters to be estimated.

This trip-distribution model was calibrated within the Intercity Project *(7)* by Abdel-Nasser for each of the four major modes: automobile, taxi, bus, and rail, using the Gaussian least-squares method.

The corresponding calibration results for the trip-distribution model for low-income passengers are obtained by computing a weighted average of θ_0 and θ_1 where the weights correspond to the modal split of low-income groups between bus and rail as obtained from the modal-split survey conducted in 1979 by NEDECO *(5)*. That is,

$$(\theta_i)_{\text{low}} = \frac{0.75}{0.75 + 0.80} (\theta_i)_{\text{bus}} + \frac{0.80}{0.75 + 0.80} (\theta_i)_{\text{rail}}$$

where $i = 0$ and 1.

The results of this calculation imply the following values of θ_0 and θ_1 for low-income passengers: $\theta_0 = 0.01402$ and $\theta_1 = 1.1044$. To estimate the parameter θ so that $\theta_0 d_{ij} = \theta U_{ij}$ (recall that U_{ij} is minimum travel cost from i to j), reasonable estimates of U_{ij} for selected O-D pairs were obtained by assuming free-flow conditions (i.e., from the initial solution of the algorithm); a weighted average of those values was then computed, yielding $\theta = 1.50714$.

As far as the trip-generation model is concerned, the observed trip generation for the low-income group, G_i^o, was estimated as the weighted average of bus and rail trip generation based on the modal-split survey results by NEDECO *(5)*. The corresponding accessibility, S_i^o, was computed as the natural logarithm of the denominator of the trip-distribution model (by definition). The minimum trip generation, E_i, was assumed to represent about 90 percent of trips of the low-income group (i.e., there is a large portion of low-income passengers who must travel for socioeconomic reasons). That is, $E_i = 0.90\ G_i^o$ for all i. A parameter α_i was then estimated for each origin. Table 5 shows the values of G_i^o, S_i^o, E_i, and α_i for the 24 zones of this study. In the table, α_i is observed to be large for the major generators of traffic in the system (i.e., Cairo, Banha, Shebin Kom, and Alexandria). For the remaining zones, α_i is almost consistently less than 200. In this study, an $\alpha = 200$ was assumed; this would lead to underestimating total demands from the major generators.

BEHAVIORAL ASSESSMENT

In this section the applicability of the STEM methodology is assessed from the behavioral point of view. This includes assessment of the ability to represent observed behavior and to predict behavioral changes.

It should be mentioned at the outset that the behavioral capabilities of the STEM depend upon the state of the art of modeling travel behavior, the behavioral assumptions of the STEM itself, the ability of modeling behavior on the particular system, the availability and reliability of appropriate data, and the existence of special peculiar features on that system. The major limitation in relation to the state of the art is the lack of a well-defined theory of trip-generation and trip-distribution behavior. The STEM itself stands somewhere in the middle of the range of behaviorally acceptable transportation planning models. The behavioral modeling of the Egyptian intercity system

TABLE 5 TRIP GENERATION DATA

Zone	G_i^o	S_i^o	E_i	α_i
CAIR	125,540.	9.363	112,986.	1,340
BNHA	81,240.	9.9286	73,116.	818
ZGZG	20,178.	9.9134	18,160.	204
ABKB	6,803.	9.9134	6,123.	69
MNSR	21,004.	9.6107	18,904.	212
SHRB	9,700.	9.6107	8,730.	101
DMIT	6,494.	8.894	5,845.	73
PRTS	2,323.	8.3537	2,091.	28
ISML	7,846.	9.1587	7,061.	86
SWES	2,580.	8.5713	2,322.	30
SHKM	33,200.	10.0362	29,880.	331
TANT	38,677.	9.3537	34,809.	414
MHLK	9,781	9.3537	8,803.	105
KFRS	11,706.	9.538	10,535.	123
ETYB	6,000.	9.334	5,400.	64
DMHR	13,682.	9.334	12,314	147
ALEX	26,615.	8.5737	23,954	310
FYUM	11,226.	8.9846	10,103.	125
BSWF	13,751.	8.7727	12,376.	157
MNIA	6,237.	7.3347	5,613.	85
ASYT	9,867.	6.31	8,880.	156
SHAG	6,934.	5.948	6,241	117
QENA	6,062.	4.9553	5,456.	122
ASWN	2,834.	2.4167	2,550.	118

(see previous section) appears to be reasonable in some, but not all, of its components. The main areas of limitation appear to be the calibration of demand models (because of the lack of theory and supportive data), the estimation of loading and unloading delays (because of the considerable amount of research and data collection efforts required to obtain better estimates), and the modeling of fleet capacity constraints (because of the limitations of the state of the art). The major special features of the Egyptian intercity system are the nonexistence of the usual traffic congestion, the existence of fleet capacity constraints, and the topology of the network (i.e., a very dense network in Lower Egypt and one corridor in Upper Egypt).

The existence of these special features and limitations are expected to limit the generality of analysis. Nevertheless, results of analysis should be fruitful in terms of identifying areas of potential improvements in the particular application at hand as well as in the STEM approach itself.

Table 6 compares the "observed" trip-generation and trip-attraction data with those predicted on the network representing the existing situation (i.e., NET3). On the aggregate level, total demand predicted on NET3 is within 1.5 percent of the corresponding observed value (i.e., 473,045 versus 480,280 trips; see last row of Table 6); this is quite satisfactory. Percent difference between predicted and observed trip generation is less than 10 percent for all origins with number of trips exceeding 10,000 and is less than 20 percent for almost all origins with trips less than 10,000; again this is very reasonable. The greatest differences, percentagewise, are observed for PRTS and SWES (+63.8 and +57.8 percent, respectively). In absolute terms, however, these differences should not be overemphasized. The main reason for these discrepancies is related to the choice of an aggregate average value of the parameter α. This suggests that trip-generation predictions, in general, may be improved by defining a specific parameter α_i for each origin; note that the STEM methodology allows this modification.

Looking at trip-attraction results in Table 6, one finds that the differences are higher both percentagewise and in absolute terms. Recall that the STEM predicts trip attraction indirectly through trip distribution. Therefore, the reasons for these discrepancies may be revealed by analyzing trip-distribution results.

Table 7 shows the results of analysis of trip distribution between the single most important O-D pair in the system—Cairo and Banha. Table 7 shows the observed data and predicted trips distributed between Cairo and Banha. It is obvious that differences between predicted and observed values are very large percentagewise and in absolute terms. Predicted trips from Cairo to Banha represent an underestimation of 32 percent (i.e., 38,549 trips predicted on NET3 compared with 56,886 trips observed) and from Banha to Cairo represent an even greater underestimation of 54 percent (i.e., 32,080 trips predicted compared with 70,729 trips observed).

The main reason for these large biases appears to be the misspecification of the trip-distribution model; namely, the attractiveness measure $A_j = \theta_1 \ln D_j$. It appears that the total number of trips attracted, D_j, is "too aggregate" to capture the

TABLE 6 COMPARISON BETWEEN PREDICTED AND OBSERVED TRIP GENERATION AND ATTRACTION

	Trip Generation			Trip Attraction		
Zone	NET3	Observed	Percent Difference	NET3	Observed	Percent Difference
ALEX	25,688	26,615	−3.5	14,125	26,150	−46
DMHR	14,093	13,682	+3	13,898	13,824	+0.5
ETYB	7,187	6,000	+19.8	7,953	6,224	+27.8
KFRS	12,331	11,706	+5.3	9,820	11,196	−12.3
MHLK	10,644	9,781	+8.8	12,776	9,924	+28.7
TANT	36,644	38,677	−5.2	48,555	38,452	+26.3
SHKM	31,719	33,200	−4.5	44,208	33,628	+31.5
BNHA	74,949	81,240	−7.7	88,104	66,383	+32.7
CAIR	114,739	125,540	−8.6	110,823	139,565	−20.6
ZGZG	19,992	20,178	−0.9	29,205	21,768	+34.2
ABKB	7,939	6,803	+16.7	8,385	5,716	+46.7
MNSR	20,708	21,004	−1.4	18,178	19,692	−7.7
SHRB	10,519	9,700	+8.4	9,765	11,447	−14.7
DMIT	7,605	6,494	+17.1	4,024	6,209	−35.2
PRTS	3,805	2,323	+63.8	2,018	2,975	−32.2
ISML	8,827	7,846	+12.5	6,476	7,524	−13.9
SWES	4,070	2,580	+57.8	2,550	2,903	−12.1
FYUM	11,826	11,226	+5.3	8,156	12,156	−32.9
BSWF	14,102	13,751	+2.6	11,571	12,206	−5.2
MNIA	7,211	6,237	+15.6	4,607	6,777	−32
ASYT	10,368	9,867	+5.1	6,871	9,432	−2.7
SHAG	7,662	6,934	+10.5	5,702	7,294	−21.8
QENA	6,754	6,062	+11.4	4,115	6,025	−55.7
ASWN	3,664	2,834	+29.3	1,162	2,624	−55.7
Total	473,045	480,280	−1.5	473,045	480,280	−1.5

TABLE 7 ANALYSIS OF DEMAND BETWEEN CAIRO AND BANHA

From	To CAIR	BNHA	G_i
Observed Trips			
CAIR	–	56,886	125,540
BNHA	70,729	–	81,240
D_j	139,565	66,383	480,280
Predicted Trips (A2)			
CAIR	–	15,834	114,739
BNHA	21,735	–	74,949
D_j	–	–	473,045
Predicted Trips (A1)			
CAIR	–	51,977	114,739
BNHA	65,206	–	74,949
D_j	–	–	473,045
Predicted Trips (NET3)			
CAIR	–	38,549	114,739
BNHA	32,080	–	74,949
D_j	110,823	88,104	473,045

variability in destination choice behavior of users at different origins. To see this, compare the following two specifications in which the influence of perceived cost is neglected:

A1:

$$A_{ij} = \ln T_{ij} \quad T_{ij} = G_i \frac{T_{ij}}{\sum_k T_{ik}} \quad \text{for all } ij \epsilon R$$

A2:

$$A_j = \ln D_j \quad T_{ij} = G_i \frac{D_j}{\sum_k D_k} \quad \text{for all } ij \epsilon R$$

Table 7 also shows predictions based on A1 and A2 specifications. It is quite obvious that A1 represents a considerable improvement in the model specification, essentially by assuming that attractiveness of alternative destinations can be origin-specific. This modification can be incorporated easily into the STEM.

Another factor that would have introduced biases in trip-distribution behavior is the parameter θ. Again, this parameter is currently specified as a systemwide average. It can, however, be an origin-specific parameter θ_i for each origin. The STEM methodology also allows this modification. A third source of bias, particularly in the application at hand, is the existence of fleet capacity constraints.

The comparison between modal-split predictions on NET3 and observed values is shown in Table 8, which reveals that NET3 predictions imply more, longer trips than what was observed; that is, the average distance traveled is 56 percent greater on NET3 than it is for observed trips. This is one of the implications of the misspecification of attractiveness in the trip-distribution model. That is, the relative importance of destinations in Upper Egypt (Lower Egypt) to users originating from Lower Egypt (Upper Egypt) was overestimated, and the relative importance of destinations in the same region, particularly those in Upper Egypt, was underestimated *(3)*.

On the basis of that observation, the railway results in Table 8 are quite consistent and reasonable. As for the results of normal bus travel, predicted values are far below observed values. It seems that the assumption of the observed number of low-income passengers using the bus (i.e., 75 percent) may be relatively above the national average and hence the actual percentage may be far less than 75 percent. Unfortunately, available data could not provide a better estimate. In addition, it seems that the assumption that low-income passengers will not select lux bus under any circumstances is too restrictive, because lux bus is essentially an aggregation of five types of bus service—first class, Arrow, Flight Pullman, Lux Pullman, and Super Lux Pullman—and two of these types have a "sufficient" quality similar to the taxi *(5)*. Therefore, including these two types in the choice set of low-income passengers should improve modal-split predictions.

At the aggregate level, predictions appear to be reasonable except that they imply longer trips than those observed, as explained earlier.

As far as traffic assignment is concerned, no observed data were available for comparison. Nevertheless, results should essentially reflect the foregoing biases.

In order to assess the ability of the approach to predict behavioral changes, it is necessary to assume that NET3 predictions represent actual behavior. In view of the previous comparison between NET3 predicted and observed trips, this assumption is not valid. However, an attempt to address this issue was made by assuming that express train fleet capacity is doubled everywhere in the system (i.e., NET4). The implications of this change on user behavior were assessed by comparing predictions before (NET3) and after (NET4) the change. The results of this comparison, though influenced by the biases introduced through the misspecification of the trip-distribution model and the existence of fleet capacity constraints, reflected the potential capability of the approach to predict rational behavioral responses of users to policy changes in the system [details on the application have been given by Safwat *(3)*].

TABLE 8 COMPARISON OF MODAL-SPLIT PREDICTIONS ON NET3 AND OBSERVED VALUES

Mode	No. of Passengers			Passenger Kilometers (thousands)			Average Distance (km)		
	Observed	Predicted	Percent Difference	Observed	Predicted	Percent Difference	Observed	Predicted	Percent Difference
Rail	381,808	348,861	–8.6	30,247	43,268	+43	79	124	+57
Bus	300,616	75,707	–75	13,151	5,035	–62	44	66.5	+51
Taxi	–	59,097	–	–	3,903	–	52	66	+27
Total	682,424	519,615	–24	43,398	52,206	+20.3	64	100	+56

SUMMARY AND CONCLUSIONS

In this paper, the major objective was to assess the applicability from the behavioral point of view of the STEM methodology developed by Safwat and Magnanti *(1)*. The main concern was assessing the ability to represent observed behavior and to predict behavioral changes.

To achieve this objective, the STEM approach was applied to a real-world transportation network, namely, the Egyptian intercity transportation system.

The major conclusions may be summarized as follows:

- As to the ability to represent actual behavior, it was found that most of the biases between predicted and observed data were attributed to the misspecification of the trip-distribution model (particularly its attractiveness measure) and the existence of severe fleet capacity constraints (particularly along the Cairo-Banha-Tanta corridor in the Middle Delta region). Appropriate modifications to the current specification, which can easily be incorporated within the STEM framework, were suggested.
- As to the ability to predict behavioral changes, results were greatly influenced by the preceding conclusion. However, there were strong indications that the STEM would be capable of predicting rational behavioral responses of users to policy changes in the system.

As indicated at the introduction, the computational issues of the application are addressed in a companion paper by Safwat in this Record. In addition, a recent application to a large-scale urban transportation network (in Austin, Texas) has further demonstrated the computational tractability of the STEM methodology *(10)*. The value of the approach as compared with other existing methodologies was highlighted by Safwat and Magnanti *(1)*. Several case studies involving passengers and freight on the Egyptian intercity system have recently been completed *(11)*. An extended version of the STEM model was a central component of the methodology used in these case studies.

Further research in relation to the STEM methodology should include more applications, particularly in the urban context, as well as more refinement of the model assumptions and computational procedures.

ACKNOWLEDGMENTS

This work was funded by the U.S. Agency for International Development through the Technology Adaptation Program at Massachusetts Institute of Technology. Support for writing of the paper was provided by the Department of Civil Engineering, Michigan Technological University, Houghton, Michigan.

The author would like to thank three anonymous referees and Yosef Sheffi, Massachusetts Institute of Technology and chairman of the TRB Task Force on Transportation Supply Analysis, for their invaluable comments on earlier versions of the paper.

REFERENCES

1. K. N. A. Safwat and T. L. Magnanti. *A Combined Trip Generation, Trip Distribution, Modal Split and Trip Assignment Model.* Working Paper OR-112-82. Center for Operations Research, Massachusetts Institute of Technology, Cambridge, 1982.
2. F. Moavenzadeh, M. Markow, B. Brademeyer, and K. N. A. Safwat. A Methodology for Intercity Transportation Planning in Egypt. *Transportation Research A*, Vol. 17A, No. 6, 1983.
3. K. N. A. Safwat. *The Simultaneous Prediction of Equilibrium on Large-Scale Networks: A Unified Consistent Methodology for Transportation Planning.* Ph.D. dissertation. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, 1982.
4. Louis Berger, Inc. *Egypt National Transport Study, Phase I, Final Report.* Ministry of Transport, Cairo, Egypt, 1977.
5. Netherlands Engineering Consultants. *Egypt National Transport Study, Phase II, Final Report.* Ministry of Transport, Cairo, Egypt, 1981.
6. K. N. A. Safwat. *Egypt Intercity Transport Project—A Comprehensive Report.* Technology Adaptation Program, Massachusetts Institute of Technology, Cambridge, 1981.
7. *Egypt Intercity Transport Project.* Technology Adaptation Program, Massachusetts Institute of Technology, Cambridge, 1982.
8. C. F. Daganzo. On the Traffic Assignment Problem with Flow Dependent Costs, I and II. *Transportation Research,* Vol. 11, 1977, pp. 433–441.
9. D. W. Hearn. *Study of Introducing Flow Bounds to Traffic Assignment Models.* Final Report. Mathtech, Inc., Princeton, N.J., 1979.
10. K. N. A. Safwat and C. M. Walton. Application of a Simultaneous Transportation Equilibrium Model to Urban Travel in Austin, Texas. Presented at the Joint National Meeting of The Institute of Management Science and Operations Research Society of America, Los Angeles, Calif., April 1986.
11. *Update and Application of the Intercity Transport Model: Final Report.* CU/MIT Technology Adaptation Program, Development Research and Technological Planning Center, Cairo University, Egypt, 1986.

Computational Experience with a Convergent Algorithm for the Simultaneous Prediction of Transportation Equilibrium

K. NABIL A. SAFWAT

A report is given of the computational experience with a globally convergent algorithm [Shortest Path to the most Needy Destination (SPND)] that predicts trip generation, trip distribution, modal split, and traffic assignment simultaneously on a Simultaneous Transportation Equilibrium Model, developed by Safwat and Magnanti, when it is applied to analyze intercity passenger travel in Egypt. A good convergence criterion, known a priori to be zero at equilibrium, was found on the basis of the solution procedure itself. In order to achieve an accuracy of about 1 percent within the optimum value of the objective function, the CPU time on a VAX-11 VMS computer was 379 sec for a network with 24 origins, 552 origin-destination pairs, 152 nodes, and 224 links. The SPND algorithm is expected to perform better in applications involving the usual urban traffic congestion in contrast to the "fictitous severe congestion" caused by the existence of fleet capacity constraints on the Egyptian intercity system. A companion paper by Safwat in this Record addresses the behavioral aspects of the application.

Safwat and Magnanti *(1)* developed a combined trip-generation, trip-distribution, modal-split, and trip-assignment model that can predict demand and performance levels on large-scale transportation networks simultaneously, that is, a Simultaneous Transportation Equilibrium Model (STEM). The STEM is formulated as an equivalent convex optimization program (ECP) that is solved by a globally convergent algorithm [Shortest Path to the most Needy Destination (SPND)].

The STEM methodology is intended to achieve a practical compromise between behavioral and computational aspects of modeling transportation systems. The model was applied to analyze intercity passenger travel in Egypt.

In a companion paper in this Record, Safwat addresses the behavioral aspects of the application. In this paper, the major objective, however, is to assess the computational experience with the SPND algorithm when applied to the Egyptian intercity system.

In the next section, a brief description of the STEM methodology is provided. This is followed by a summary of the major aspects of modeling the Egyptian intercity passenger transport system. Computational issues are discussed next. This includes two major issues: the convergence criterion and computational efficiency. The last section contains a summary and conclusions.

Department of Urban and Regional Planning, Texas A&M University System, College Station, Tex. 77843.

STEM METHODOLOGY

A brief description is given of a STEM, an ECP, and an algorithm (SPND) for solving the ECP in order to predict equilibrium on the STEM. For a detailed description, the reader is referred to the paper by Safwat and Magnanti *(1)*.

A STEM

In this subsection, a STEM that describes user travel behavior in response to system performance on a transportation network is presented as follows:

$$G_i = \alpha_i S_i + E_i \quad \text{for all } i \in I \tag{1}$$

$$S_i = \max[0, \ln \sum_{J \in D_i} \exp(-\theta_i U_{ij} + A_j)] \quad \text{for all } i \in I \tag{2}$$

$$T_{ij} = G_i \frac{\exp(-\theta_i U_{ij} + A_j)}{\sum_{k \in D_i} \exp(-\theta_i U_{ik} + A_k)} \quad \text{for all } ij \in R \tag{3}$$

$$\left.\begin{aligned} C_p &= U_{ij} \quad \text{if } H_p > 0 \\ C_p &\geq U_{ij} \quad \text{if } H_p = 0 \end{aligned}\right\} \quad \text{for all } p \in P \tag{4}$$

$$C_p = \sum_{a \in A} \delta_{ap} C_a(F_a) \quad \text{for all } p \in P \tag{5}$$

In this model, the demand variables are as follows:

- G_i = number of trips generated from origin i,
- T_{ij} = number of trips distributed from origin i to destination j,
- H_p = number of trips via path p from any given origin i to any given destination j, and
- F_a = number of trips using link a.

The performance variables are as follows:

- S_i = accessibility variable that measures the expected maximum utility of travel on the transport system as perceived from origin i,
- U_{ij} = average minimum "perceived" cost of travel from i to j,
- C_p = average cost of travel via path p from any given i to any given j, and

C_a = average cost of travel on link a expressed as a function of the number of trips (F_a) on that link.

The remaining quantities are as follows:

E_i = composite measure of the effect that the socioeconomic variables, which are exogenous to the transport system, have on trip generation from origin i;

A_j = composite measure of the effect that the socioeconomic variables, which are exogenous to the transport system, have on trip attraction at destination j;

α_i = parameter that measures the additional number of trips that would be generated from any given origin i if the expected maximum utility of travel, as perceived by travelers at i, increased by unity;

θ_i = parameter that measures the sensitivity of travelers at any given origin i to changes in system performance between any given origin-destination (O-D) pair $ij : j \in D_i$;

$$\delta_{ap} = \begin{cases} 1 & \text{if link } a \text{ belongs to path } p \\ 0 & \text{otherwise} \end{cases}$$

and the defined sets are as follows:

I = set of origins,
R = set of O-D pairs ij,
P = set of simple paths in the network, and
D_i = set of destinations accessible from origin i.

The basic assumptions of this STEM may be summarized as follows:

1. Trip generation (G_i) is given by any general function as long as it is linearly dependent on the system's performance through an accessibility measure (S_i) based on the random utility theory of travel behavior (i.e., the expected maximum utility of travel).
2. Trip distribution (T_{ij}) is given by a logit model whose measured utility functions include the average minimum perceived travel costs (U_{ij} for all $j \in D_i$) as variables with a linear parameter θ_i.
3. Modal split and trip assignment are simultaneously user optimized. Note that the STEM framework allows for the modal split to be given by a logit model or to be (together with trip assignment) system optimized *(2)*.

An ECP

Consider the following optimization problem:

Minimize

$$Z(S, T, H) = J(S) + \psi(T) + \phi(H)$$

subject to

$$\sum_{j \in D_i} T_{ij} = \alpha S_i + E_i \quad \text{for all } i \in I \tag{6}$$

$$\sum_{p \in P_{ij}} H_p = T_{ij} \quad \text{for all } ij \in R \tag{7}$$

$$\left.\begin{array}{ll} S_i \geq 0 & \text{for all } i \in I \\ T_{ij} \geq 0 & \text{for all } ij \in R \\ H_p \geq 0 & \text{for all } p \in P \end{array}\right\} \tag{8}$$

where

$$J(S) = \sum_{i \in I} \frac{1}{\theta_i} [\alpha_i S_i^2 + \alpha_i S_i - (\alpha_i S_i + E_i) \times \ln(\alpha_i S_i + E_i)]$$

$$\psi(T) = \sum_{i \in I} \frac{1}{\theta_i} \sum_{j \in D_i} (T_{ij} \ln T_{ij} - A_j T_{ij} - T_{ij})$$

$$\phi(H) = \sum_{a \in A} \int_0^{F_a} C_a(w)\, dw$$

$$F_a = \sum_p \delta_{ap} H_p \quad \text{for all } a \in A \tag{9}$$

Constraints 6 and 7 are flow conservation equations on the transport network stating (a) that the number of trips distributed from a given origin to all possible destinations must equal the total number generated from that origin, and (b) that the number of trips on all paths joining a given O-D pair must equal the total number distributed from that origin to that destination. Constraints 8 state, as postulated earlier, that all the decision variables must be nonnegative. Expressions 9 define the link-path incidence relationships, stating that the flow on a given link equals the sum of flows on all paths sharing that link.

The objective function Z has three sets of terms. The last of these, $\phi(H)$, corresponds to the familiar transformation introduced by Beckman et al. *(3)*. The second set of terms, $\psi(T)$, is similar to those used by Evans *(4)* and by Florian and Nguyen *(5)*, as well as in other related models. The first set of terms, $J(S)$, was introduced by Safwat and Magnanti *(1)*, who proved that under mild monotonicity assumptions on performance functions and nonnegativity and inequality assumptions on demand parameters (i.e., $\theta_i > 0$, $E_i > \alpha_i > 0$ for all $i \in I$), the ECP has a unique solution that is equivalent to equilibrium on the STEM.

Algorithm for Predicting Equilibrium on the STEM (SPND)

In this subsection, an algorithm (SPND) for solving the ECP to predict equilibrium on the STEM is introduced. The (SPND) algorithm belongs essentially to the class of feasible-direction methods.

At any given iteration r, the method involves two main steps. The first step determines a direction for improvement, d^r. The second step determines an optimum step size, λ^*, along that direction. The current solution, X^r, is then updated, $X^{r+1} = X^r + \lambda^* . d^r$, and the process is repeated until a convergence criterion is met.

In accordance with the Frank-Wolf method *(6)*, the feasible direction d^r is determined as follows:

$$d^r = Y^r - X^r$$

where X^r is the given current solution (S^r, T^r, F^r) and Y^r is the solution of the following linearized subproblem (LP1):

Minimize $Z_L^r(Y) = \nabla Z(X^r)Y$ subject to Equations 6, 7, 8, and 9. The steps of the (SPND) algorithm to determine a feasible direction d^r at iteration r are as follows:

Step 1. Update link costs by calculating $C_a^r = C_a(F_a^r)$ for all $a \in A$. Set $i = 1$ in an ordered set of origins I.

Step 2. Find the shortest path tree from i to all $j \in D_i$. Let U_{ij}^r be the cost on the shortest path from i to j.

Step 3. Calculate $w_{ij}^r = 1/\theta_i\,(\ln T_{ij}^r - A_j) + U_{ij}^r$ for all $j \in D_i$.

Step 4. Determine j^* satisfying $w_{ij^*}^r = \min_{J \in D_i} \{w_{ij}^r\}$.

Step 5. Calculate $C_i^r = 1/\theta_i\,[S_i^r - \ln(\alpha_i S_i^r + E_i)]$.

Step 6. Set $i \leftarrow i + 1$. If $i \in I$, go to Step 2. Otherwise, continue.

Step 7. Find an optimum solution to LP1 and a feasible direction d^r as described in the following.

The optimum solution to LP1 is the vector $\mathbf{y}^r = (\hat{S}^r, \hat{T}^r, \hat{H}^r)$, given by

$$\hat{S}_i^r = \begin{cases} 0 \text{ if } C_i^r + w_{ij^*} \geq 0 \\ (M_i - E_i)/\alpha_i \text{ otherwise} \end{cases} \quad \text{for all } i \in I$$

Hence

$$\hat{G}_i^r = \alpha_i \hat{S}_i^r + E_i \quad \text{for all } i \in I$$

$$\hat{T}_{ij}^r = \begin{cases} \hat{G}_i^r \text{ if } j = j^* \in D_i \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } ij \in R$$

$$\hat{H}_p^r = \begin{cases} \hat{G}_i^r \text{ if } p = p_{ij}^* \in P_{ij^*} \\ 0 \text{ otherwise} \end{cases} \quad \text{for all } p \in p_{ij} \text{ and } ij \in R$$

where M_i is the maximum trip generation from i (assuming zero travel cost everywhere in the system).

The path flows can be decomposed into link flows using the link-path incidence relationship as follows:

$$\hat{F}_a^r = \begin{cases} \sum_i \delta_{ap^*} \hat{G}_i^r \text{ if link } a \text{ belongs to path } p^* \text{ between } i \text{ and } j^* \\ 0 \text{ otherwise} \end{cases}$$

Hence, the feasible direction at iteration r is the vector $\mathbf{d}^r$ with the following components:

$$\mathbf{d}_i^r = \hat{S}_i^r - S_i^r \quad \text{for all } i \in I$$

$$\mathbf{d}_{ij}^r = \hat{T}_j^r - T_{ij}^r \quad \text{for all } ij \in R$$

$$\mathbf{d}_a^r = \hat{F}_a^r - F_a^r \quad \text{for all } a \in A$$

The main computational effort in this direction-finding algorithm is finding the set of shortest paths from all origins to all destinations in Step 2, which is identical to that of the traffic assignment problem with fixed demand. The additional calculations in Steps 3–5 are insignificant compared with those in Step 2. Step 7 just loads the shortest paths with the total demand to the most "needy" destinations, which involves even less effort than the all-or-nothing loading procedure. This procedure is referred to as the shortest path to the most needy destination (SPND) algorithm as dictated by its direction-finding step.

In the following two sections, the computational issues involved in the application of STEM to intercity passenger travel in Egypt are addressed. In the next section, the major aspects related to modeling intercity passenger travel in Egypt are introduced. Then the computational results of the application are analyzed.

MODELING THE EGYPTIAN INTERCITY PASSENGER SYSTEM

Intercity passengers in Egypt utilize two major networks: highway and railway. On highways, passengers may select among private automobiles, taxis, and buses. On railways they may select among diesel, express, and local trains. Furthermore, passengers may be divided into three types according to whether their income is high, middle, or low. For more details on this topic, the reader is referred to papers by Safwat *(2, 7)*. This application focused on one passenger type (the low-income group) and considered taxis, buses, express trains, and local trains as the feasible modes for this passenger type.

The usual link congestion problems encountered in urban travel are insignificant on the Egyptian intercity system. Instead the system is congested because of its fleet capacities. Though these fleet capacity constraints are essentially "hard" constraints, the modeling approach was to treat them as a congestion term added to the link cost functions in a way that drives the user cost to a very high value whenever flows exceed fleet capacity. That is, for any given link, the fleet capacity cost (FCC) may be expressed as follows:

$$\text{FCC} = \delta(\text{flow/fleet capacity})^{\beta}$$

where δ and β are link congestion parameters, assumed equal to 0.1 and 20, respectively. These assumptions, particularly $\beta = 20$, give very steep cost functions. Consequently, more iterations would be required than are customary to achieve any given level of accuracy. Similar approximations to deal with hard link capacities have been suggested by several researchers [Daganzo *(8)* and Hearn *(9)*].

To model the demand functions, a logit trip-distribution model based on observed data was calibrated. Trip generation was assumed to have a minimum E_i of 90 percent of observed values. The maximum trip-generation M_i was calculated assuming that transportation costs are zero everywhere in the system. This choice certainly gives a sufficiently large value, which may be large enough to significantly reduce the step size between successive iterations and hence to adversely affect the performance of the algorithm. A better choice of M_i would be to assume that transportation costs are at their minimum values corresponding to zero flows on the system.

The application included four network sizes (see Table 1) ranging from 90 nodes and 224 links to 152 nodes and 534 links. All networks had 24 origins and 552 origin-destination pairs. Each of these four networks is essentially a composed

TABLE 1 MAJOR CHARACTERISTICS OF THE FOUR PROBLEMS IN THE ANALYSIS

Name	Description	Network Size				Fleet Capacity
		No. of Origins	No. of O-D Pairs	No. of Nodes	No. of Links	
NET1	Express and local	24	552	90	224	Severely constrained
NET2	Express, local, and normal bus	24	552	125	394	Less constrained
NET3	Express, local, and normal bus and taxi	24	555	152	534	Not constrained
NET4	Express (doubled), local, and normal bus and taxi	24	552	152	534	More relaxed than NET3

multimodal network consisting of individual modal networks connected through a set of loading and unloading links at different zonal centroids. This network representation allows modal split and trip assignment to be performed simultaneously.

COMPUTATIONAL RESULTS

In this section, the application is assessed from the computational point of view. The major issues considered are related to the convergence criterion and computational efficiency.

Convergence Criterion

Several convergence measures were tested to find the best criterion for stopping the algorithm when the current solution is sufficiently close to the exact equilibrium. Many of the measures tested were essentially comparisons between the results of the last two iterations. It was obvious that there is a strong correlation between these criteria and the optimum step size, λ^r, at any given iteration r. Figure 1 shows the step size λ at each iteration of the SPND algorithm for the first 200 iterations. It is clear that using any criterion related to λ may cause the procedure to stop prematurely (e.g., at iteration 13).

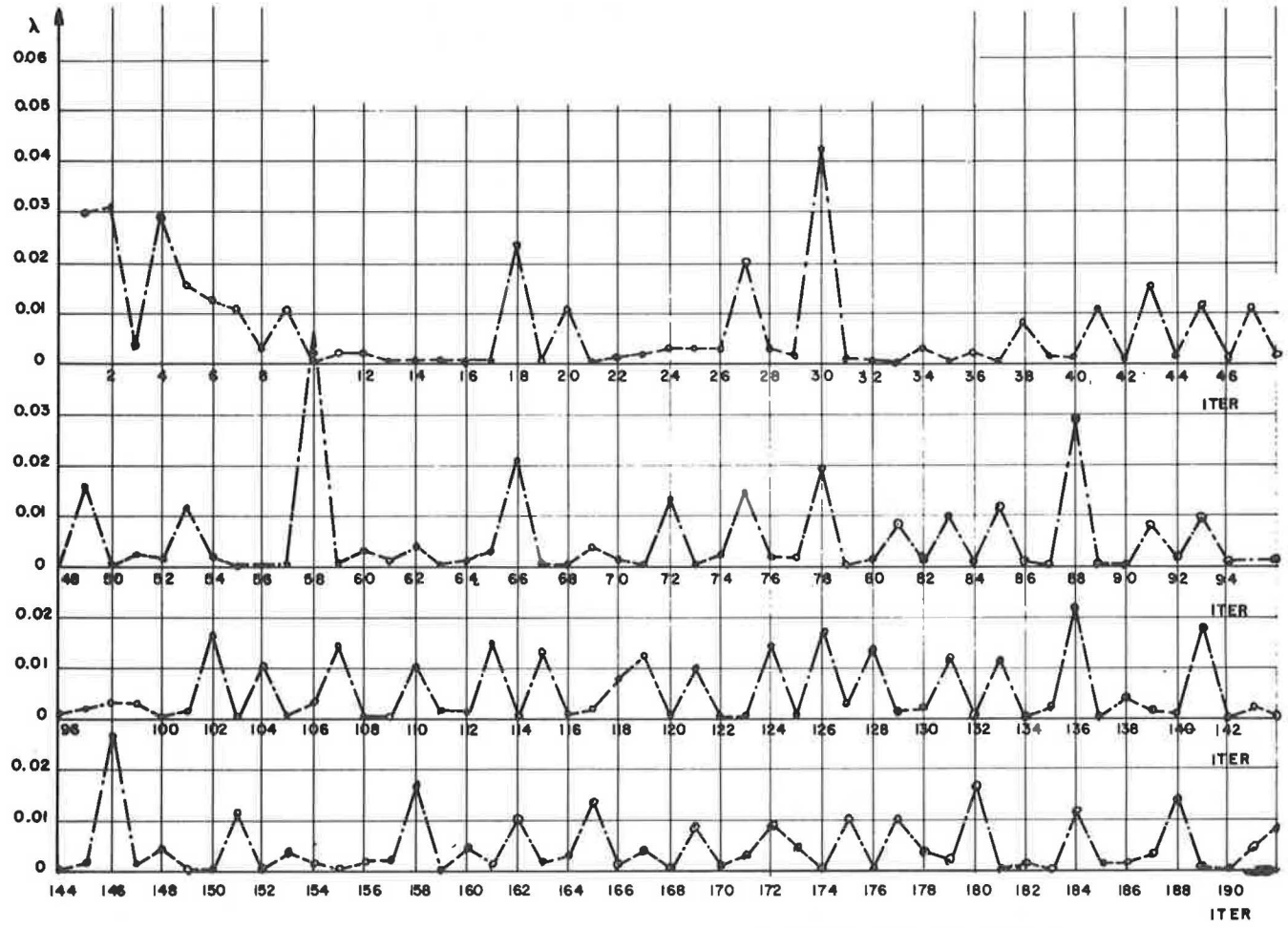

FIGURE 1 Step size (λ) versus number of iterations (*ITER*) for express and local trains, bus, and taxi.

The value of the objective function Z, itself, is monotonically decreasing in the SPND algorithm and could have been a good criterion except that its optimum value is not known a priori.

An apparent criterion (which was found to be the best one) was one based on the procedure itself. As discussed earlier, the direction for trip generation at each iteration is determined on the basis of the following calculation:

$$U_i^r = \frac{1}{\theta_i}[S_i^r - \ln(\alpha_i S_i^r + E_i)] + \frac{1}{\theta_i}[\ln T_{ij^*}^r - A_{j^*}] + U_{ij^*}^r \quad \text{for all } i$$

where j^* is the most needy destination in the set D_i at iteration r. The value of U_i^r may be interpreted as the marginal cost of generating an additional trip from origin i going through the shortest path to the most needy destination j^*. If U_i^r is negative, the current level of demand generated at i may be increased, and vice versa. Hence at equilibrium, U_i^* should satisfy the following conditions:

1. $U_i^* = 0$ if $E_i < G_i^* < M_i$,
2. $U_i^* \geq 0$ if $E_i = G_i^*$,
3. $U_i^* \leq 0$ If $G_i^* = M_i$.
3. $U_i^* \leq 0$ if $G_i^* = M_i$.

Let

$$\text{ERMSE} = \left(\sum_i U_i^r\right)^{1/2}$$

where ERMSE is the equilibrium root-mean-square (rms) error and the summation is over all i such that $E_i < G_i^r < M_i$. (This assumption may not represent any limitation, because M_i can always be selected to ensure the strict inequality. Also, if the problem is defined such that there is always at least one "attractive" destination for any given origin i, then $S_i^r > 0$ and hence $G_i^r > E_i$ for all i and r.)

Then at equilibrium ERMSE is 0 and hence can be used as a "good" convergence criterion for the SPND algorithm. Figure 2 shows the performance of the ERMSE measure for the four problems included in the analysis (see Table 1). It is obvious that the ERMSE has the desirable properties of a good convergence criterion. Why it has a slow convergence toward its optimum value (i.e., zero) is explained next.

Computational Efficiency

Computational efficiency depends on several factors, such as network size, fleet capacity constraints, initialization, steepness of cost functions, parameters of demand functions, and the nature of the algorithm itself.

It turns out that the existence of fleet capacity constraints (FCC), which is essentially a special feature of the Egyptian system, has had an adverse influence on the computational performance of the SPND algorithm.

First, in order to obtain a reasonable initial feasible solution, the existence of FCC required performance of an incremental traffic assignment process in the initialization step of the SPND algorithm. (A description of the modified initialization step is given in the next section.) Second, in order to maintain reasonable feasibility as the algorithm proceeds, the step size in the one-dimensional search at any given iteration had to be restricted in a similar manner to that suggested by Daganzo *(8)*. Results by Hearn and Ribera *(10)* ensure convergence of the modified procedure. Third, the FCC term in the user cost function produces steep cost functions. These three major implications of the FCC have resulted in additional computational efforts for the SPND algorithm, in the sense of increasing the number of iterations and the CPU time for initialization. At any given iteration, however, the method is extremely efficient, as indicated earlier.

Different components of the computer CPU time (in seconds) on VAX-11/VMS for all four problems considered in analysis are shown in Table 2. The CPU time for a typical iteration varied between 1.57 and 3.32 sec (excluding input-output time). Total CPU time for 100 iterations varied between 216 and 379 sec (including input-output time and initialization).

At the 100th iteration the objective function was approximately within 1 percent of its optimum value. Figure 3 shows the value of the objective function at different iterations of the algorithm. The tailing-off phenomenon of the SPND algorithm (a well-known characteristic of the Frank-Wolf method) is evident from Figure 3. The decrease in the objective value during the first 100 iterations was 5 to 6 times that of the following 100 iterations. Also, for problems that are more relaxed in terms of fleet capacities, convergence is relatively faster. This confirms earlier comments on the influence of fleet capacity on computational efficiency. Nevertheless, in view of the foregoing computational results, the SPND algorithm appears to be reasonably efficient for analyzing large-scale systems.

Initialization Process and One-Dimensional Modified Search Procedures

Step 0—Initialization

Step 0.1: Assume that the network is empty and calculate minimum link perceived costs; that is, set $F^o = 0$ and calculate $C^o = C_a(0)$, for all $a \in A$. Set $i = 1$ in an ordered set of origins I.

Step 0.2: Find the shortest tree from i to all other destinations. That is, U_{ij}^o for all $j \in D_i$. Set $j = 1$ in an ordered set of O-D pairs D_i.

Step 0.3: Calculate initial trip generation and trip distribution as follows:

$$G_i^o = E_i$$

$$T_{ij}^o = G_i^o \frac{\exp(-\theta u_{ij}^o + A_j)}{\sum_{k \in D_i} \exp(-\theta u_{ik}^o + A_k)} \quad \text{for all } j \in D_i$$

Then set $i \leftarrow i + 1$; if $i \in I$, go to Step 0.2; otherwise, set i and $j = 1$, and continue.

Step 0.4: Determine the increment ΔT_{ij}^o to be assigned to the shortest path, p^o, from i to j such that the fleet capacity on any

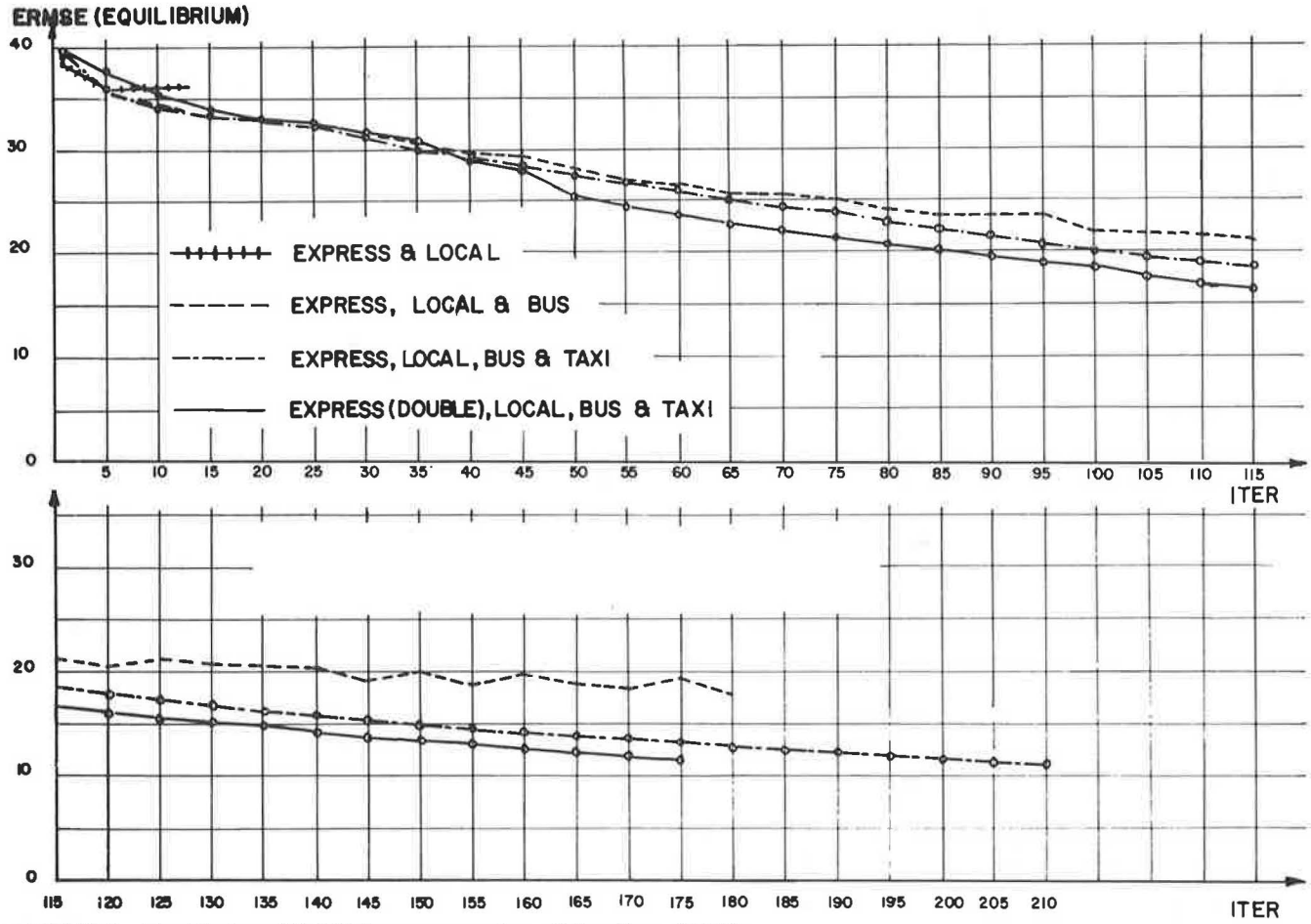

FIGURE 2 Equilibrium (*ERMSE*) versus number of iterations (*ITER*).

TABLE 2 COMPUTER CPU TIME ON VAX-11/VMS

	Time (sec) by Problem Name			
CPU Time Components	NET1	NET2	NET3	NET4
Initialization	11.5	68	20.2	19.4
Typical iteration				
Direction finding	1.37	1.95	2.57	2.57
One-dimensional search	0.2 ~ 0.5	0.3 ~ 0.5	0.38 ~ 0.75	0.38 ~ 0.75
Convergence test, intermediate output	0.26	0.35	0.38	0.38
Total per iteration	1.98	2.7	3.5	3.5
Final output	6.13	8.	8.85	8.85
Total CPU time (for 100 iterations)	215.63	346.	379.	378.2

link on p^o may not be exceeded by more than 20 percent. That is,

$$(\text{CAPACITY})_a \leftarrow 1.2 * (\text{CAPACITY})_a \quad \text{for all } a \in p^o$$

$$(\text{CAPACITY})_p^o = \min_{a \in p^o} (\text{CAPACITY})_a$$

$$\Delta T_{ij}^o = \min[T_{ij}^o, (\text{CAPACITY})_p^o]$$

Step 0.5: Assign the increment ΔT_{ij}^o and update link fleet capacities and flows. That is,

$$F_a^o = \begin{cases} f_a^o + \Delta T_{ij}^o & \text{if } a \in p^o \\ F_a^o & \text{otherwise} \end{cases}$$

$$(\text{CAPACITY})_a \leftarrow (\text{CAPACITY})_a - \Delta T_{ij}^o \quad \text{for all } a \in p^o$$

$$T_{ij}^o \leftarrow T_{ij}^o - \Delta T_{ij}^o$$

Then set $j \leftarrow j + 1$; if $j \in D_i$, go to Step 0.4; otherwise, continue.

Step 0.6: Update link costs and shortest trees. That is,

$$C_a^o = C_a(F_a^o) \quad \text{for all } a \in A$$

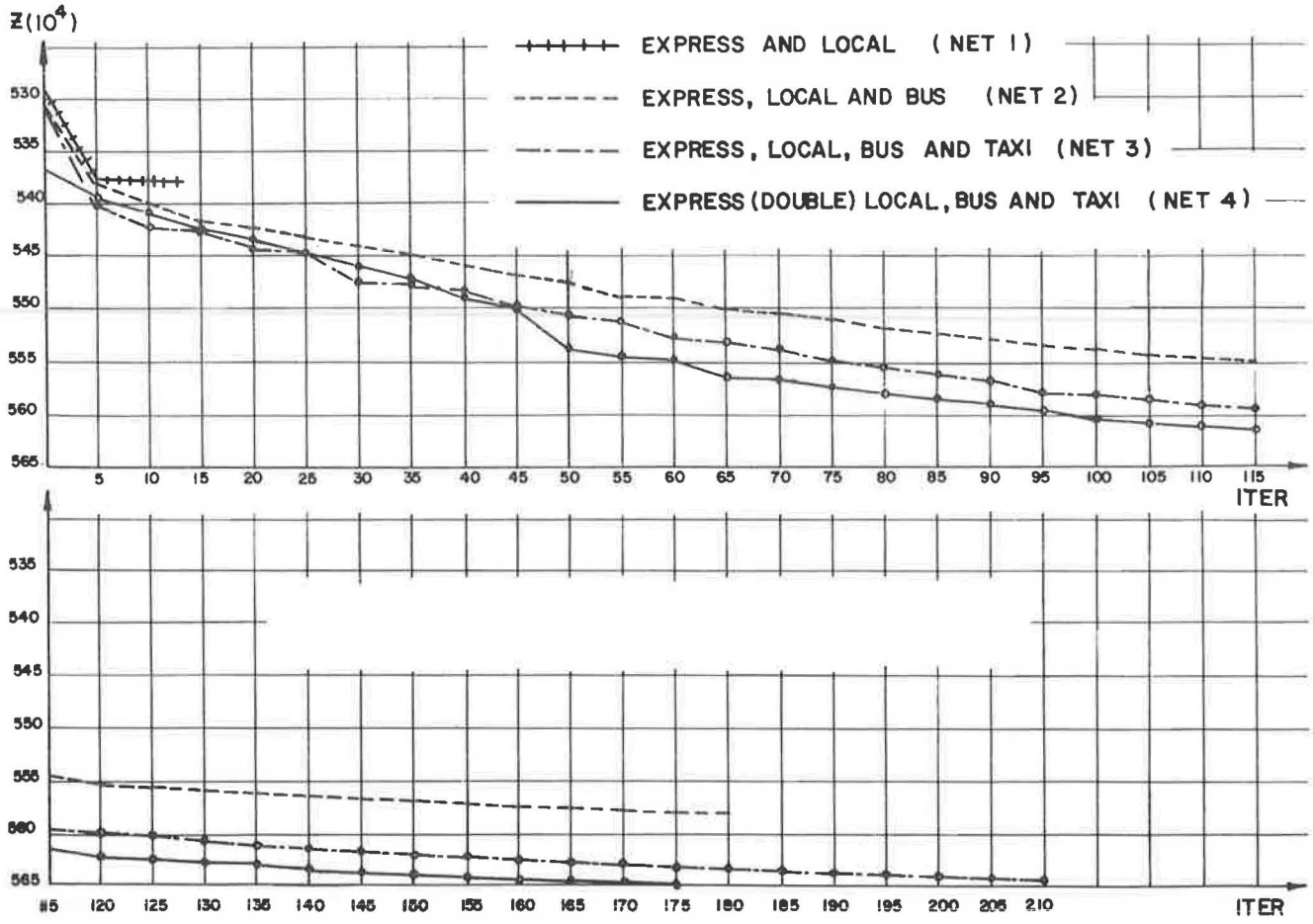

FIGURE 3 Objective function (Z) versus number of iterations (*ITER*).

Then set $i \leftarrow i + 1$; if $i \in I$, find the shortest tree, and go to Step 0.4; otherwise, continue.

Step 0.7: Check for termination. That is, if $T^o_{ij} = 0$ for all ij, then stop; an initial feasible solution is obtained.

If $T^o_{ij} \neq 0$ for some O-D pairs but has been constant for the last two iterations, then stop; an initial feasible solution cannot be obtained. Otherwise, set $i = 1$ and $j = 1$, and go to Step 0.4.

As far as the one-dimensional search is concerned, the step size is essentially restricted in such a way to maintain reasonable feasibility of the solution as the search proceeds. The idea has been suggested before by Daganzo *(8)*. The following modifications are formally introduced:

Step 2—One-Dimensional Search

Step 2.1: Calculate maximum step size, λ_{max}, as follows:

$$\lambda_{max} = \min \left[1, \min_{d_a>0} \frac{(\text{CAPACITY})_a - \bar{F}_a}{d_a} \right]$$

where d_a is the descent direction on link a.

$$(\text{CAPACITY})_a \leftarrow 1.3 * (\text{CAPACITY})_a$$

Step 2.2: Minimize $Z(\lambda)$ subject to $0 \leq \lambda \leq \lambda_{max}$.

In order for the modified procedure to converge, Daganzo *(8)* invokes a strong assumption that is not satisfied in this case, namely, the link cost is required to approach infinity as the link flow approaches its capacity. Hearn and Ribera *(10)* proved convergence of this modified procedure under a weaker and more natural assumption. They required that whenever the flow approaches capacity, the link cost be sufficiently large that the integral $\phi(H^o)$ in the objective function at the initial solution be strictly less than that $\phi(H^c)$ when the flows are at their capacities. This assumption is satisfied in the modified procedure because the flows in any initial solution cannot exceed more than 20 percent of capacities, whereas in subsequent iterations the flows can reach up to 30 percent more than capacities. The corresponding costs are magnified with a power of 20, implying that the value of $\phi(H^c)$ where the flows are at their relaxed capacities should always be strictly greater than that of $\phi(H^o)$ at the initial solution.

SUMMARY AND CONCLUSIONS

Modeling transportation systems must invariably balance behavioral richness and computational tractability. Safwat and Magnanti *(1)* developed a combined trip-generation, trip-distribution, modal-split, and trip-assignment model that can predict demand and performance levels on large-scale transporta-

tion networks simultaneously—a STEM methodology, which is intended to achieve a practical compromise between behavioral and computational aspects of modeling transportation systems.

In order to assess its applicability, the STEM was applied to analyze intercity passenger travel in Egypt. In a companion paper in this Record, Safwat addresses the behavioral aspects of the application. In this paper, the major objective was to report the computational experience with the SPND algorithm concerns were to suggest a convergence criterion for the algorithm and to assess its computational efficiency.

The major conclusions of this paper may be summarized as follows:

1. A good convergence criterion based on the solution procedure itself was found. As the algorithm proceeds, the convergence measure gets closer to its optimum value at equilibrium, which is known a priori to be zero.
2. As far as the computational efficiency of the SPND algorithm is concerned, the computer CPU time required to achieve an accuracy of about +1 percent within the optimum value of the objective function varied between 216 and 379 sec on a VAX-11/VMS minicomputer depending on the network size, which varied from 24 origins, 552 O-D pairs, 90 nodes, and 224 links to 24 origins, 552 O-D pairs, 152 nodes, and 534 links. The algorithm is expected to perform better in applications involving the usual urban traffic congestion in contrast to the fictitious severe congestion caused by the existence of fleet capacity constraints on the Egyptian system.

Safwat and Walton *(11)* applied the STEM to urban travel in Austin, Texas. The Austin network consisted of 520 zones, 19,214 O-D pairs, 7,096 links and 2,137 nodes. The computer CPU time on an IBM 4381 was 430 sec for a typical iteration. A modification of the SPND algorithm (though consistently converged to the unique equilibrium solution, it does not yet have a formal proof of global convergence) arrived at a reasonably accurate solution in only 10 iterations.

An extended version of the STEM was included in a comprehensive intercity transportation planning methodology in Egypt [see paper by Moavenzadeh et al. *(12)*]. Several case studies involving multimodal transportation of passengers and freight in Egypt have been completed *(13)*.

Further research in relation to the STEM methodology should include more applications, particularly in the urban context, as well as more refinement of the model assumptions and computational procedures.

ACKNOWLEDGMENTS

This work was funded by the U.S. Agency for International Development through the Technology Adaptation Program at Massachusetts Institute of Technology. Support for writing of the paper was provided by the Department of Civil Engineering, Michigan Technological University, Houghton, Michigan.

The author would like to thank three anonymous referees and Yosef Sheffi, Massachusetts Institute of Technology and chairman of the TRB Task Force on Transportation Supply Analysis, for their invaluable comments on earlier versions of the paper.

REFERENCES

1. K. N. A. Safwat and T. L. Magnanti. *A Combined Trip Generation, Trip Distribution, Modal Split and Trip Assignment Model.* Working Paper OR-112-82. Center for Operations Research, Massachusetts Institute of Technology, Cambridge, 1982.
2. K. N. A. Safwat. *The Simultaneous Prediction of Equilibrium on Large-Scale Networks: A Unified Consistent Methodology for Transportation Planning.* Ph.D. dissertation. Department of Civil Engineering, Massachusetts Institute of Technology, Cambridge, 1982.
3. M. Beckman, C. B. McGuire, and C. B. Winston. *Studies in the Economics of Transportation.* Yale University Press, New Haven, Conn., 1958.
4. S. P. Evans. Derivation and Analysis of Some Models for Combining Trip Distribution and Assignment. *Transportation Research,* Vol. 10, 1976, pp. 37–57.
5. M. Florian and S. Nguyen. A Combined Trip Distribution, Mode Split and Trip Assignment Model. *Transportation Research,* Vol. 12, 1978, pp. 241–246.
6. M. Frank and P. Wolf. An Algorithm for Quadratic Programming. *Naval Research Logistics Quarterly,* Vol. 3, 1956, pp. 95–110.
7. K. N. A. Safwat. *Egypt Intercity Transport Project—A Comprehensive Report.* Technology Adaptation Program, Massachusetts Institute of Technology, Cambridge, 1981.
8. C. F. Daganzo. On the Traffic Assignment Problem with Flow Dependent LOSTS, I and II. *Transportation Research,* Vol. 11, 1977, pp. 433–441.
9. D. W. Hearn. Study of Introducing Flow Bounds to Traffic Assignment Models. Final Report. Mathtech, Inc., Princeton, N.J., 1979.
10. D. W. Hearn and J. Ribera. Convergence of the Frank-Wolf Method for Certain Bounded Variable Traffic Assignment Problems. *Transportation Research B,* Vol. 15B, No. 6, 1981, pp. 437–442.
11. K. N. A. Safwat and C. M. Walton. Application of a Simultaneous Transportation Equilibrium Model to Urban Travel in Austin, Texas: Computational Experience. Presented at the Joint National Meeting of The Institute of Management Science and Operations Research Society of America, Los Angeles, Calif., April 1986.
12. F. Moavenzadeh, M. Markow, B. Brademeyer, and K. N. A. Safwat. A Methodology for Intercity Transportation Planning in Egypt. *Transportation Research A,* Vol. 17A, No. 6, 1983.
13. *Update and Application of Egypt Intercity Transport Model: Final Report.* CU/MIT Technology Adaptation Program, Development Research and Technological Planning Center, Cairo University, Egypt, 1986.