A Method for Direct Estimation of Origin/Destination Trip Matrices

MALACHY CAREY, CHRIS HENDRICKSON and KRISHNASWAMI SIDDHARTHAN

Carnegie-Mellon University, Pittsburgh, Pennsylvania

Estimates of the volume of travel between zones or locations is often a necessary step in transportation studies. This paper suggests a method of estimating origin/destination volumes, using a direct demand function and incomplete, aggregate data. Most attention is devoted to demand functions which are linear in their parameters. With such demand functions, volume estimates are obtained from a quadratic programming problem, which minimizes the sum of squared errors from a direct demand function, subject to constraints derived from observations of some travel volumes. A decomposition algorithm is suggested for solving the programming problem and is proven to converge. The method may also be used in the trip distribution phase of the conventional urban transportation model systems (UTMS).

An intermediate step in many travel analysis and planning studies is the estimation of origin/destination (OD) trip matrices. Each entry in such a matrix represents the volume of travel originating at a particular "zone" and destined for some other particular zone. For example, a trip matrix might represent peak-hour travel between districts in a metropolitan area, or daily air travel between major cities.

The most common means of estimating a matrix of interzonal trip volumes involves a survey of individual tripmakers. Trip volume forecasts are often obtained by sequential application of models which first estimate the total number of trips generated by or attracted to a zone (i.e., "trip generation") and then distribute the expected trips among potential destination zones (i.e., "trip distribution").

In the present paper we present a one-step method for estimating origin/destination matrices from aggregate data. The method is intended for use when the available information is limited; a general travel inter-
view survey is not required. The method is suggested for networks such as rapid transit systems, regional airlines, or major intercity roadways.

Three major assumptions are made in developing the model. First, we assume that the aggregate volume of travel between any two zones may be modeled as a function of socioeconomic data specific to that pair of zones and the impedance of travel between the pair; we shall call such a relationship a direct demand model. Second, we assume that interzonal volume estimates should be consistent with all available and relevant travel volume observations. Such consistency is ensured in the application of the method. Finally, we assume that interzonal volume estimates should be made so as to minimize the sum of squares of estimation errors with respect to the direct demand function. We shall discuss each of these assumptions at greater length below.

The plan of the paper is as follows. The trip distribution model is formulated in Section 1 and discussed in Section 2. The input data is discussed in Section 3. Sections 4 and 5 present solution techniques for large scale problems; for smaller scale problems any quadratic programming algorithm will suffice. Two appendices prove various properties of the model solutions and the convergence of a suggested decomposition solution technique.

In this paper the words “flow” and “volume” and the words “zone” and “node” will be used interchangeably. Also, it should be remembered that the OD travel volumes with which we are concerned are directly observable from traffic counts only if there are exclusive links between OD pairs; in most cases, however, the observed volume on a link (road, street, etc.) will consist of flows between several OD pairs.

1. MODEL FORMULATION

The precise formulation of the model depends upon the extent and the type of information which is available in particular applications. We assume that the available information may be of five types:

1. The travel impedance \( t_{ij} \) between each origin zone \( i \) and destination zone \( j \) (\( t_{ij} \) may be a single attribute—such as travel time—or a vector of travel attributes);
2. Socioeconomic characteristics of each origin zone and each destination zone, denoted by the vectors \( x_i \) and \( x_j \) respectively;
3. Observations of some interzonal flows \( q_{ij}^0 \);
4. Observations of the volume of travel between some mutually exclusive sets of nodes, i.e., cut volumes \( V_i \) (defined below);
5. Observations of the total flow \( M_i \) originating at zone \( i \) and the total flow \( A_j \) terminating at zone \( j \). (This is a special case of data of Type 4.)
Attributes of travel impedance (Category 1 above) and socioeconomic characteristics of each zone (Category 2) are required for each zonal pair and zone respectively. At least some observations of Types 3, 4, or 5 are required; minimum requirements are discussed below. We will assume that all the flow observations of Types 3, 4 and 5 are consistent with each other. The various observations of volumes and travel impedances are assumed to be representative of a particular time period (e.g., daily or peak period conditions) and may be restricted to a particular mode or type of travel.

The model is formulated as a mathematical programming model in which the objective is to minimize the sum of squared errors in prediction from a direct demand function, subject to the constraints derived from the various travel volume observations. A direct demand function expresses the aggregate travel flow, \( q_{ij} \), from zone \( i \) to zone \( j \) as a function of socioeconomic characteristics \( (x_i, x_j) \) of the two zones and of the travel impedance \( t_{ij} \) between the zones:

\[
q_{ij} = f(x_i, x_j, t_{ij}, \alpha) + \epsilon_{ij}
\]

(1)

where \( x_{ij} = [x_i, x_j, t_{ij}] \) is a data vector, \( \epsilon_{ij} \) is an error term, and \( f(\cdot) \) is some assumed functional form (see Section 2 below) with parameters \( \alpha \).

Other notation used below includes:

- \( S^0 \) = the set of all \( (i, j) \) pairs having observed interzonal flows, \( q_{ij}^0 \)
- \( \hat{q}_{ij} \) = the flow (to be estimated) from zone \( i \) to zone \( j \) where \( (i, j) \not\in S^0 \). Also, \( \hat{q} = [\ldots, \hat{q}_{ij}, \ldots] \)
- \( C(k) \) = a cut set, i.e. a set of direct \( (i, j) \) paths (see Section 3 below).
- \( C^0(k) \) = the subset of paths in \( C(k) \) for which there are flow observations \( q_{ij}^0 \), i.e., \( C^0(k) = C(k) \cap S^0 \). Also, let \( C''(k) \) represent the subset of paths in \( C(k) \) with unobserved flow, thus \( C(k) = C''(k) \cup C^0(k) \) and \( C''(k) \cap C^0(k) = \emptyset \).
- \( V_k \) = the observed total flow in the cut set \( C(k) \).

Mathematically, the model is:

\[
P0: \text{Min } z(\hat{q}, \alpha) = \sum_{(i,j) \not\in S^0} (\hat{q}_{ij} - f(x_{ij}, \alpha))^2 \\
+ \sum_{(i,j) \in S^0} (q_{ij}^0 - f(x_{ij}, \alpha))^2 \quad (2)
\]

subject to:

\[
\sum_{j \mid (i,j) \not\in S^0} \hat{q}_{ij} + \sum_{i \mid (i,j) \in S^0} q_{ij}^0 = M, \text{ for (some) origins } i
\]

(3)

\[
\sum_{i \mid (i,j) \not\in S^0} \hat{q}_{ij} + \sum_{j \mid (i,j) \in S^0} q_{ij}^0 = A, \text{ for (some) destinations } j
\]

(4)
\[ \sum_{(i,j) \in C^u(k)} \hat{q}_u + \sum_{(i,j) \in C^o(k)} q^0_{ij} = V_k \text{ for all cuts } k \quad (5) \]
\[ \hat{q}_u \geq 0 \text{ for all } (i, j). \quad (6) \]

We will refer to the problem, (2) to (6), as \( P_0 \).

If there are no volume observations of type \( M_i, A_j \), or \( V_k \), then \( P_0 \) reduces to an unconstrained least squares estimation of \( \hat{q}_u \). In this (unconstrained) case the first expression in the objective function of \( P_0 \) will go to zero, so that the predictions of unobserved interzonal flows \( \hat{q}_u \) will be exactly \( \hat{q}_u = f(\cdot) \), or \( \hat{q}_u = 0 \) in the unlikely case that \( f(\cdot) \leq 0 \). The parameters of \( \alpha \) will then be determined solely from minimizing the second expression in the objective of \( P_0 \), i.e., regressing \( q^0_{ij} \) on \( x_u \), \( (i, j) \in S^0 \).

Similarly, any flow variables \( \hat{q}_u \) which do not appear in the constraints (3), (4) and (5) in \( P_0 \) will equal \( f(x_u, \alpha) \), so that the term \( (\hat{q}_u - f(x_u, \alpha))^2 \) in the objective function will be zero. As a result, these unconstrained, unobserved flow variables (denoted \( q^o_{ij} \)) need not be included in the formulation of problem \( P_0 \). Once an estimate of \( \alpha \) is obtained from the solution of \( P_0 \), estimates of these unconstrained, unobserved flows \( q^o_{ij} \) can be calculated as \( q^o_{ij} = \max \{0, f(x_u, \alpha)\} \). In what follows, we shall therefore assume that all \( \hat{q}_u \) variables appear in both the objective function and the constraints of \( P_0 \).

The decision variables in \( P_0 \) are \( \hat{q} \) and \( \alpha \). A lemma in Appendix 1 shows that the problem \( P_0 \) will have a finite (bounded) solution for \( \hat{q} \) and \( \alpha \) as long as (a) the constraints of \( P_0 \) are consistent, so that a set of feasible flows exist, and (b) the least-squares value of \( \alpha \) and \( f(\cdot) \) obtained from (1) are finite when the vectors \( x = [x_u] \) and \( q = [q_{ij}] \) are finite.

For simplicity and clarity we have omitted adding more subscripts or superscripts, but these can be inserted for applications in which flows by mode of transport, purpose, time of day or other category are desired.

2. DISCUSSION OF THE MODEL

Direct demand functions for estimation of travel volumes have been used extensively for transportation planning, albeit with mixed success. Early examples include Domencich and Kraft\(^{[10]} \) and Quandt and Baumol\(^{[19]} \). A review appears in Stopher and Meyburg\(^{[21]} \). These models were suggested by consumer demand theory and have utilized a variety of functional forms. Two of the more common forms are a linear demand model:

\[ q_u = \alpha_0 + \alpha_1 x_i + \alpha_2 x_j + \alpha_3 t_u \quad (1b) \]

or a multiplicative form:

\[ q_u = \alpha_0 x_i^{\alpha_1} x_j^{\alpha_2} t_u^{\alpha_3}. \quad (1c) \]
If \( x_i, x_j, t_{ij} \) and \( \alpha_i \) represent vectors (i.e. \( x_i = [x_{ik}] \), \( t_{ij} = [t_{ijm}] \) and \( \alpha_i = [\alpha_{im}] \)) then (1b) retains the same form and (1c) is written as

\[
q_{ij} = \alpha_0 \prod_{m=1}^{p} x_{im}^{\alpha_{im}} \prod_{m=1}^{q} x_{jm}^{\alpha_{jm}} \prod_{m=1}^{r} t_{ijm}^{\alpha_{ijm}}. 
\quad (1d)
\]

In most applications of direct demand models, the only travel flow information used for estimation of parameters, \( \alpha \), is that of interzonal flow observations, \( q_{ij} \).

While the model \( P0 \) formulated above is simply an extension of the unconstrained direct demand modeling approach, it does have a number of advantages over the latter. First, volume observations of types other than interzonal flows provide additional information for estimation of flows and parameter values. In addition, the model ensures that all volume estimates are consistent with known information. The constraints may also serve to correct for deficiencies in the assumed functional form of the demand function, such as the omission of travel impedance to competing destinations. Finally, the interzonal flows may be estimated from the model even with very few or no direct observations of interzonal flows.

The modeling system \( P0 \) also has advantages with respect to more conventional sequential demand modeling techniques. First, sequential models often do not require consistency between modeling stages or with observed network flows. Second, travel impedances are used explicitly in the estimation of interzonal flows in our formulation. Finally, and most importantly, the modeling system \( P0 \) allows us to use various kinds of flow data, particularly traffic count data, which is much more easily available than OD flow data obtained from interview surveys. Consequently, the cost of obtaining data for the model is likely to be very much less than for more traditional methods.

Functional forms other than (1b), (1c) and (1d) can be used for the direct demand functions \( f(\cdot) \) in (2). Also, different functional forms or parameters (e.g. a piecewise linear form) can be used over various ranges of the \( q_{ij} \)'s. We should note however that functional forms which are nonlinear in their parameters can lead to computational difficulties; we shall discuss this point below.

As a parenthetical note, the problem \( P0 \) may be considered a generalization of the conventional trip distribution problem. In that problem, estimates of generating and terminating volumes (\( M_i \) and \( A_j \)) are developed in earlier analysis stages for all nodes. No observations of interzonal volumes, \( q_{ij} \), or other cut volumes, \( V_k \), are included. The solution algorithm developed below can also be applied to this specialized problem.

3. MEASUREMENT OF CUT VOLUMES AND TRAVEL IMPEDANCES

In cases in which only one physical route exists between two zones, the interzonal travel impedance is equal to the impedance on this unique
path. In a user equilibrium system, the impedance on each route between two zones must be equal on all paths with positive flow, so the interzonal travel impedance is equal to the travel impedance on any route with positive flow. In cases with multiple routes and different travel impedances on the routes, the analyst must develop some means of estimating a representative interzonal travel impedance.

Observations of specific interzonal flows ($q_{ij}^0$) may be obtained if an exclusive link connects two zones or if specialized interview surveys (such as cordon traveler surveys) are conducted. Observations of total trip volumes terminating or originating at a zone $A$, and $M_i$, respectively may be obtained from employment totals (for the case of work trips), special surveys, or counting devices, e.g., counters on freeway entry and exit ramps or turnstile counts at rail rapid transit stations.

Cut volume observations $V_k$ represent the total flow between two sets of nodes. Link volume, turnstile and ticket observations may be aggregated to obtain cut volume observations, as long as care is taken to include all relevant paths. Cut volumes may be obtained fairly easily for networks with sparse, radial lines such as the linear network in Figure 1, which might represent a rail rapid transit system line or a freeway through a metropolitan area. An observation of the total flow past point $P$ in Figure 1 toward the CBD represents a cut volume which includes all the flow from nodes 1 and 2 to all other nodes in the network, i.e., $V_1 = q_{13} + q_{14} + q_{23} + q_{24}$, where $V_1$ is the observed flow past $P$ toward the CBD. Similarly, an observation of travel flow past $P$ going away from the CBD is a cut volume which includes all flow from other network nodes to nodes 1 or 2.

In roadway or other networks with numerous connections and paths, it is possible to obtain cut volume observations, although the analyst must be particularly careful to ensure that all relevant flows, and only those flows, are included in the cut volume. Figure 2 illustrates a small roadway network with two unambiguous cut observations $C(1)$ and $C(2)$, which may be obtained by summing the volumes on links $(3, 1)$, $(3, 2)$ and $(3, 4)$ for $C(1)$ and on links $(1, 4)$, $(2, 4)$ and $(3, 4)$ for $C(2)$. The resulting constraint equations for $P0$ from $C(1)$ and $C(2)$ are shown on the figure. The sum of the link volumes $(1, 2)$, $(1, 4)$ and $(3, 4)$ can be taken as the cut volume from nodes 1 and 3 to nodes 2 and 4, assuming that no flow from node 2 to node 4 takes the path via node 1. In many cases, traffic cordon counts such as $C(3)$ may be taken as a true cut volume.

![Fig. 1. Example of a linear transportation network.](image-url)
Fig. 2. Examples of cut volumes calculated from link volume counts.

\[ q_{31} + q_{32} + q_{34} = V_1 \]
\[ q_{14} + q_{24} + q_{34} = V_2 \]
\[ q_{12} + q_{14} + q_{32} + q_{34} \leq V_3 \]

4. SOLVING THE ESTIMATION PROBLEM \( P0 \)

We will first consider solving the problem \( P0 \) (i.e. (2) to (6)) for the case in which the direct demand model \( f(\cdot) \) embodied in (2) is assumed to be linear with respect to the parameters \( \alpha \). In this case the demand function (1b) can be written as \( q_{ij} = x_{ij} \alpha + \epsilon_{ij} \), where \( \alpha \) is a column vector and \( x_{ij} = [1, x_i, x_j, t_{ij}] \) is a row vector. Then \( P0 \) becomes \( P0^1 \):

\[ P0^1: \ \text{Min} \ z(\hat{q}, \alpha) = \sum_{(i,j) \notin S^0} (\hat{q}_{ij} - x_{ij} \alpha)^2 + \sum_{(i,j) \in S^0} (q^0_{ij} - x_{ij} \alpha)^2 \quad (2b) \]

subject to (3) to (6). Problem \( P0^1 \) has a convex (see appendix) quadratic objective function and linear constraints. If the number of variables and constraints in \( P0^1 \) is not large, then any quadratic programming code may be used to solve \( P0^1 \) since computational efficiency will be unimportant.

However, the larger is \( P0^1 \), the more important is the choice of computational approach, and if \( P0^1 \) is very large then it can become prohibitively expensive to solve \( P0^1 \) using a general purpose quadratic program which does not take advantage of the special structure of the constraints of \( P0^1 \).

The model \( P0^1 \) can be decomposed into simpler subproblems. In particular, if \( \hat{q} \) is held constant, then \( P0^1 \) reduces to an unconstrained least squares regression problem for \( \alpha \). Accordingly, we suggest an iterative two stage decomposition of the problem:

(i) Let \( \bar{\alpha} \) be a given value of \( \alpha \), and compute the corresponding internodal flows \( \hat{q} \), i.e.,

\[ \text{Min} \ \{ z(\hat{q}, \bar{\alpha}) = \sum_{(i,j) \notin S^0} (\hat{q}_{ij} - x_{ij} \bar{\alpha})^2, \quad (7) \]

subject to (3) to (6)
(ii) Let \( \tilde{q} = [\tilde{q}_u] \) be a given value of \( \tilde{q} \), and compute the corresponding value of the parameter \( \alpha \), i.e.,

\[
\text{Min } z(\tilde{q}, \alpha) = \sum_{(i,j) \in S^0} (\tilde{q}_u - x_{ij}\alpha)^2 + \sum_{(u,j) \in S^0} (q_u^0 - x_{uj}\alpha)^2.
\] (8)

Problem (8) is equivalent to an unconstrained least-squares regression, and problem (7) is a quadratic programming problem which has a special structure, of which we will later take advantage. The above decomposition is formalized in the following algorithm. Appendix 2 shows that this algorithm will converge to a global optimum of \( P0' \).

**Algorithm A**

**Step 0. Initialize.** Let \( h = 1 \). Select an initial parameter estimate \( |\alpha^1| < +\infty \) and let \( z_i^0 = +\infty \). Select convergence tolerance parameters \( \epsilon_1 > 0 \) and \( \epsilon_2 > 0 \).

**Step 1. Flow Estimation.** Solve

\[
P1: \text{Min } \{z^h(\hat{q}, \alpha^h) = \sum_{(i,j) \in S^0} (\hat{q}_u - x_{ij}\alpha^h)^2, \} \tag{7'}
\]

subject to (3) to (6)) for \( \hat{q} \), using any suitable algorithm (see Section 5 below). Let the solution be \( \hat{q}^h \).

**Convergence test.** Many different tests are possible, e.g., for \( h \geq 2 \),

\[
\text{if } |\hat{q}_u^h - \hat{q}_u^{h-1}| \leq \epsilon_1 \text{ for all } i, j
\]

and/or

\[
|z^h - z^{h-1}| \leq \epsilon_2
\]

where \( z = z(\hat{q}, \alpha) \) is defined by (2b), so that \( z^h = z(\hat{q}^h, \alpha^h) \), then terminate. Otherwise, add 1 to \( h \).

**Step 2. Regression.** Solve the unconstrained problem

\[
P2: \text{Min } z^h(\hat{q}_u^h, \alpha) = \sum_{(i,j) \in S^0} (\hat{q}_u^h - x_{ij}\alpha)^2 + \sum_{(u,j) \in S^0} (q_u^0 - x_{uj}\alpha)^2 \tag{8'}
\]

for \( \alpha \). Problem P2 is equivalent to least-squares estimation of \( \alpha \) from a regression equation of the form \( q^h = X\alpha + [\epsilon_u] \), where \( q^h = [\hat{q}^h, q^0] \) is \( n^2 \times 1 \) and \( X = [X_u] \) is \( n^2 \times m \); hence the solution (\( \alpha^h \)) of P2 is given by the well known least-squares formula

\[
(X^TX)\alpha^h = X^Tq^h. \tag{9}
\]

(Usually the columns of \( X \) will be linearly independent, in which case (9) has a unique solution \( \alpha^h = (X^TX)^{-1}X^Tq^h \). In the very unlikely event that the columns of \( X \) happen to be linearly dependent (i.e., perfect multicollinearity), there will be a closed convex set of solutions for \( \alpha^h \), from (9). In this case we can choose any finite \( \alpha^h \) from this set.) Go to Step 1.

Step 0 above requires the assignment of preliminary values for the parameters of the direct demand function, \( \alpha \). These initial values might be obtained from an application of regression using the observations of
internodal flows, \( q^0_u \), as dependent variables. Alternatively, reasonable values for \( \alpha \) might be selected based on a priori knowledge.

Step 1 requires the solution of a convex quadratic programming problem \( P1 \) and, in large problems, can represent a relatively large computational burden. Solution techniques for \( P1 \) are discussed in Section 5 below.

The use of a direct demand function which is not linear in its parameters creates difficulties. For example, applications of direct demand models in transportation studies have most commonly used a multiplicative function, such as

\[
q_u = \alpha_0 X^a_i X^b_j t^c_{ij} e_u. \tag{10}
\]

Since (10) is difficult to estimate directly, a logarithmic transformation is used to obtain:

\[
\ln q_u = \ln \alpha_0 + \alpha_1 (\ln X_i) + \alpha_2 (\ln X_j) + \alpha_3 (\ln t_{ij}) + \ln(e_u) \\
= \alpha^* + \alpha_1 (\ln x_i) + \alpha_2 (\ln x_j) + \alpha_3 (\ln t_{ij}) + \epsilon^*_u \\
= \alpha^* \ln(X_u) + \epsilon^*_u \tag{11}
\]

If we wish to estimate \( \alpha^* \), and hence \( \alpha \) by minimizing \( \sum (\epsilon^*_u)^2 \) subject to the constraints (3) to (6), this implies replacing the objective function (2) in the model with:

\[
\text{Min } Z = \sum_{(u,j) \in S^0} (\ln q^0_u - \alpha^* \ln X_u)^2 \\
+ \sum_{(u,j) \notin S^0} (\ln \hat{q}_u - \alpha^* \ln X_u)^2 \tag{12}
\]

This is convex in \( \alpha^* \), but is nonconvex in \( \hat{q}_u \). Hence, if we adopt the suggested decomposition algorithm to solve (3) to (6) and (12), then the linearly constrained minimization problem of Step 1 will no longer be convex.

Alternatively, suppose that we assume an additive error term, rather than a multiplicative error term as in (10), i.e.

\[
q_u = \alpha_0 X^a_i X^b_j t^c_{ij} + e_u. \tag{13}
\]

In this case the objective function (2) of the model becomes:

\[
\text{Min } Z = \sum_{(u,j) \in S^0} \big( q^0_u - \alpha_0 X^a_i X^b_j t^c_{ij} \big)^2 \\
+ \sum_{(u,j) \notin S^0} \big( \hat{q}_u - \alpha_0 X^a_i X^b_j t^c_{ij} \big)^2. \tag{14}
\]

This is convex in \( \hat{q}_u \), but is nonconvex in \( \alpha \), which is exactly the opposite of (12) above.

With multiplicative direct demand functions, the algorithm presented above may not converge to a global minimum of \( P0 \). Consequently, it is convenient to retain functional forms which are linear in the parameters,
$\alpha$. Of course, even when the direct demand function is linear in the parameters $\alpha$, it need not be linear with respect to the impedance and socioeconomic attributes, $x_i, x_j$ and $t_{ij}$.

5. SOLVING THE SUBPROBLEM $P_1$.

Most of the computation involved in each iteration of the Algorithm A above is due to problem $P_1$ in Step 1 of A. Problem $P_1$ can be solved using any general quadratic programming algorithm, but, in the case of large scale problems, the special form of the constraints in $P_1$ allow us to use more economical solution techniques.

5.1. Transportation Type Constraints

Assume for the moment that the constraints in $P_1$ are standard transportation type constraints, i.e. assume that there are no observations of type $q_{ij}^0$ and no constraints of type (5), so that $P_1$ reduces to $P_1^1$:

$$P_1^1: \text{Min} \{ z (\hat{q}) = \sum_{i,j} (\hat{q}_{ij} - x_{ij} \tilde{a})^2 \}, \text{ subject to}$$

$$\sum_i \hat{q}_{ij} = M_i, \forall \ i; \sum_j \hat{q}_{ij} = A_j, \forall \ j; \hat{q}_{ij} \geq 0, \forall \ (i, j).$$

The discussion will be extended to more general forms of $P_1$, in Section 5.2 below.

Problem $P_1^1$ is a transportation problem with a convex separable quadratic cost function. There are several general purpose nonlinear programming algorithms, quadratic programming algorithms and fixed point algorithms which can take advantage of the special form of the convex cost transportation problem (e.g. $P_1^1$), and which will thereby have greatly improved time and storage requirements, due to the special form of the problem. For example, Hadley$^{[16]}$ (Ch. 4.12) uses piecewise linear approximation, a form of decomposition and the simplex method; Rao and Shaftei$^{[20]}$ use Zangwill's convex simplex method and the primal transportation algorithm. Two other approaches are described in Beale$^{[3]}$ and Bachem and Korte$^{[1]}$. In Bachem and Korte the form of the objective function is $\sum_j (q_{ij} - (\text{constant})_{ij})^2$, which is exactly the form of $P_1^1$, and they report that problems having more than 2,500 variables are solvable in a few seconds.

To illustrate the adaption of existing algorithms to solving $P_1^1$, we will present here only one method, namely the Frank-Wolfe$^{[12]}$ algorithm, as set out in Zangwill$^{[25]}$ (Ch. 8.1). This algorithm will converge to a global optimal solution of $P_1^1$, since $P_1^1$ has a convex, continuously differentiable objective function and a compact convex constraint space (see Zangwill$^{[25]}$).

Applying the Frank-Wolfe algorithm to $P_1^1$ yields the following algo-
rithm. For notational convenience let \( q = [q_{ij}] \) and \( y = [y_{ij}] \) be \( n^2 \times 1 \) vectors, \( \alpha = [\alpha_i] \) be a \( m \times 1 \) vector and let \( X = [x_{ij}] \) be an \( n^2 \times m \) vector.

**Algorithm B**

**Step 0.** Let \( k = 1 \). Choose an initial feasible solution \( q^k \) for \( P1^1 \) and choose a convergence tolerance parameter \( \epsilon > 0 \).

**Step 1.** Replace the objective function of \( P1^1 \) with a local linear approximation (a tangent plane) at \( q^k \), denoting the flow variables vector by \( y \). This yields a standard linear transportation problem,

\[
\min \quad \{2(q^k - X\alpha)^T y, \quad \text{subject to} \quad \sum_j y_{ij} = M_i, \forall i; \sum_i y_{ij} = A_j, \forall j; y \geq 0\}.
\]

Solve this problem, and let \( y^k \) denote its solution.

**Step 2.** Solve the line search problem which consists of seeking to minimize the objective function of \( P1^1 \) along the straight line joining \( q^k \) to \( y^k \), i.e.

\[
\min f(\theta) = (q^k + \theta(y^k - q^k) - x\alpha)^T(q^k + \theta(y^k - q^k) - X\alpha)
\]

over \( 1 \geq \theta \geq 0 \), where \( \theta \) is a scalar. Let the solution be \( \theta^k \). Test for convergence: e.g., is \( |f(\theta^k) - f(\theta^{k-1})| \leq \epsilon \)? If satisfied, terminate; otherwise let \( q^{k+1} = (q^k + \theta^k(y^k - q^k)) \), add 1 to \( k \) and go to Step 1.

The single variable optimization problem of Step 2 above is easy to solve in one step, requiring no iterations. Setting \( df(\theta)/d\theta = 0 \) yields \( 0 = 2(q^k + \theta(y^k - q^k) - X\alpha)^T(y^k - q^k) \), hence \( \theta^k = (X\alpha - q^k)^T(y^k - q^k) / (y^k - q^k)^T(y^k - q^k) \). If \( \theta^k < 0 \) set \( \theta^k = 0 \) and if \( \theta^k > 1 \) set \( \theta^k = 1 \).

Thus the Frank-Wolfe Algorithm B above reduces to solving a series of standard linear transportation problems (in Step 1), each followed by a trivial calculation in Step 2. Relevant linear transportation algorithms appear in Wagner.[24]

### 5.2 Handling Observed Trip Volumes and Cut Constraints

We saw in Section 5.1 that if the constraints of \( P1 \) have a standard transportation problem form, as in \( P1^1 \), then this leads to very great computational advantages in solving \( P1 \). But even when the constraints of \( P1 \) do not initially have this form they can very often be recast in this form.

Constraints (3) and (4) of \( P1 \) can always be converted to standard transportation type constraints by letting \( M_i^* = M_i - \sum_{l(i,j) \in S^p} q_{ij}^0 \) and \( A_j^* = A_j - \sum_{j(i,j) \in S^p} q_{ij}^0 \), so that (3) and (4) become

\[
\begin{align*}
\sum_i q_{ij} & = M_i^*, \quad \forall i \tag{3'} \\
\sum_j q_{ij} & = A_j^*, \quad \forall j \tag{4'}
\end{align*}
\]
with a very high cost penalty imposed on the \( q_{ij}, \forall (i, j) \in S^0 \), so as to ensure that these \( q_{ij} \)'s will all equal zero in the optimal solution.

Data for the \( M_i \)'s and/or \( A_j \)'s will usually be inexact, hence we may sometimes wish to replace the equality constraints (3) and/or (4) with lower and/or upper bounds. For techniques by which these inequality constraints can be recast as standard equality type transportation constraints see a standard text, e.g. Wagner,\(^{[24]}\) and also Charnes and Klingman,\(^{[6]}\) Charnes, Glover and Klingman,\(^{[5]}\) and Wagner.\(^{[23]}\)

If the cut constraint set (5) is not empty, then converting the constraints of \( P1 \) to standard transportation type constraints is more difficult and may not always be possible. Wagner\(^{[22]}\) and later A. Manne (in Dantzig,\(^{[7]}\) pp. 382–383) show that transportation problems, with added cut constraints of certain types, can be reduced to enlarged standard transportation problems. Their procedures require that if there is more than one cut constraint then these must involve only disjoint or nested sets of variables. Glover, Klingman and Ross\(^{[15]}\) extend these results by devising a procedure which will reduce any extra linear constraint in a transportation problem to a cut constraint on variables associated with a single node, if and only if this is possible. Wagner's\(^{[23]}\) procedures can then be used to transform the reduced problem into an enlarged transportation problem.

None of the above techniques allow us to transform a transportation problem containing an arbitrary collection of added cut constraints into a standard transportation problem. Fortunately, there are other specialized techniques for solving such problems. These methods tend to treat the problem as two parts: a "favored" part, which is the pure transportation problem, and an "unfavored" part which can be any arbitrary collection or "bundle" of linear constraints. We then take advantage of efficient transportation algorithms for solving the favored part. An important method in this context is the double-reverse method of Charnes and Cooper,\(^{[4]}\) Appendix F, which was later extended and refined in Bakes\(^{[2]}\) and Klingman and Russell.\(^{[17]}\) An alternative, basically different, approach is to use the Dantzig-Wolfe\(^{[9]}\) decomposition principle, though Bakes\(^{[2]}\) suggests that this is likely to be somewhat less efficient than the double-reverse method. Another alternative is to use the generalized upper-bounding approach of Dantzig and Van Slyke,\(^{[8]}\) though this handles a more restricted class of added constraints.

The references in the above paragraph are concerned with linear problems, but the approaches which they describe can be extended to problems, such as \( P1 \), having nonlinear objective functions, since we can choose a nonlinear programming algorithm which retains the linear constraints and replaces the objective function with a piecewise linear approximation or sequence of local linear approximations (see Section
5.1). These linear problems can then be solved using the approaches suggested in the previous paragraph.

6. CONCLUSIONS

The proposed methodology is a practical method for developing origin/destination trip matrices from incomplete, aggregate data such as traffic or turnstile counts. Its primary limitations involve the adequacy of the direct demand model embedded in the model and the computational burden imposed by solution of a quadratic programming problem with linear constraints. Of course, the limitation imposed by the selection of a demand function is shared by virtually all other explicit demand modeling techniques; as with other techniques, the adequacy and appropriateness of socioeconomic and impedance data as well as the perspicacity of the analyst will influence the accuracy of the demand model.

As for the computational burden, any quadratic programming algorithm can be used in the case of small to medium-sized problems, and the methods discussed in Section 4 and 5 to take advantage of the structure of the constraint set should be adequate to handle all but very large intrametropolitan problems. Further work is in progress to improve the computational efficiency of the algorithms in Section 5.

Balancing these limitations, however, the solution methodology has several advantages. In particular, the requirement that all volume estimates must be consistent with all applicable flow observations seems a desirable property. Moreover, even if for the moment we exclude applications to very large scale metropolitan level planning problems, there is likely to be a large number of problems for which the methodology is suited.

The authors are engaged in using the estimation approach presented in this paper to estimate the travel volumes among the townships, employment centers and central business district of Salisbury, Zimbabwe. The results of this study will be reported elsewhere.

APPENDICES

Appendix 1. Finiteness, Convexity and Uniqueness of the Solutions of Problems $P_0$, $P_1$ and $P_2$

The following lemma shows that in normal circumstances the problem $P_0$ will have a finite (bounded) solution, or solutions, for $z$, $\dot{q}$ and $\alpha$.

**Lemma 1.** Problem $P_0$ yields finite solution values for $z$, $\dot{q}$ and $\alpha$ if;

(a) The constraints of $P_0$ are not inconsistent, and

(b) The least-squares values of $\alpha$ and $f(\cdot)$ obtained from (1) are finite when both $X = [x_{ij}]$ and $q = [q_{ij}]$ are finite.
Proof. The variables $q_{ij}$ are bounded, since (a) and the constraints ((3), (4), (6)) ensure that $0 \leq q_{ij} \leq \min \{ M_i, A_i, V_k \mid (i, j) \in C(k) \}$. For any given $q$, the problem $P0$ reduces to an unconstrained least squares estimation of $\alpha$ from the demand functions (1). It follows immediately from (b) that $\alpha$ and $f(\cdot)$ are finite, hence $\varepsilon(\cdot)$ is finite, which completes the proof.

The following lemma states that $P0$, $P1$ and $P2$ have convex objective functions. Hence, since the constraint sets are also convex, any local minimum of $P0$, $P1$ or $P2$ is also a unique global optimum.

Lemma 2. The objective functions of $P0$, $P1$ and $P2$ are convex.

Proof. The objective functions of $P0$, $P1$ and $P2$ consist of sums of squares of scalar valued linear expressions. Consider an arbitrary scalar valued linear expression, say $f(x) = a^T x + b$. The Hessian of $(f(x))^2 = (a^T x + b)^2$ is $2aa^T$, which is positive semidefinite, hence $(f(x))^2$ is a convex function. A sum of convex functions is a convex function, hence the lemma follows immediately.

In most cases, but not all cases, the objective function of $P0$ will be strictly convex, and in these cases $P0$ will have a unique global optimum. We show this as follows.

Lemma 3. If the $m$ columns of $X$ are linearly independent (this is the normal case) and $q$ has at least $m$ pre-fixed elements, i.e. $S^0$ has $\bar{m} \geq m$ members, then $P0$ will almost always have a unique optimal solution ($\bar{q}, \bar{\alpha}$).

Proof. Let $\varepsilon = (Iq - X\alpha)$. Rearrange $P0$ by writing its objective function as $\varepsilon^T \varepsilon$ and including $\varepsilon = Iq - X\alpha$ as an explicit constraint. The objective function $\varepsilon^T \varepsilon$ is strictly convex in $\varepsilon$ and the constraint set is convex, hence $P0$ has a unique optimal solution, at $\bar{\varepsilon}$. This $\bar{\varepsilon}$ will imply a unique $(\bar{q}, \bar{\alpha})$ from $\bar{\varepsilon} = Iq - X\alpha$, if the columns of $[I^0, X]$ are linearly independent. $I^0$ is the matrix obtained by deleting from the identity matrix $I$, the columns corresponding to the constants $(q^0_0, s)$ in $Iq$. The conditions specified in the lemma will almost always be sufficient to ensure that the columns of $[I^0, X]$ are linearly independent. The lemma follows immediately.

The Hessian of the objective function of $P1$ with respect to all variables in $P1$ (i.e. the $q_{ij}$s) is $2I$, where $I$ is an identity matrix of appropriate dimension. Thus $P1$ has a strictly convex objective function, and a convex constraint set, hence any local minimum of $P1$ is the unique global optimum.

The Hessian of the objective function of $P2$ with respect to all variables in $P2$ (i.e. $\alpha$) is $2X^TX$, which is at least positive semidefinite. Normally
the columns of \( X \) will be linearly independent, in which case \( Xa \neq 0 \) for any \( a \neq 0 \) hence \( a^T(X^TX)a > 0 \). This implies a strictly convex objective function, and since the constraint set of \( P2 \) is convex, any local minimum of \( P2 \) is then the unique global optimum.

**Appendix 2. Convergence of Algorithm A**

It might seem that Algorithm A in Section 4 above is an application of a generalized Benders type decomposition (Geoffrion[13]), and that convergence could therefore be proven by quoting existing relevant theorems. Unfortunately this is not so. The variables in the given problem \( P0 \) decompose naturally into two sets, \( q \) and \( \alpha \); but only \( q \) (and not \( \alpha \)) appears in the constraints, hence a Benders type approach breaks down.

A global convergence theorem (Theorem 1) will be stated below. First we need the following four lemmas.

Denote the objective functions of problems \( P0^i, P1 \) and \( P2 \) by \( z(q, \alpha) \), \( z(q, \alpha^h) \) and \( z(q^h, \alpha) \), respectively, where the vector \( q = [q_{ij}, q_{ij}^0] \) has \( n^2 \) elements.

**Lemma 4.** The sequence of points \( (q^h, \alpha^h), h = 1, 2, \ldots, \) generated by the Algorithm A lie in a compact (convex) set.

**Proof.** The constraints on \( q_{ij}^h \) in A (i.e. (3)–(6)) represent a closed convex set, and ensure that \( 0 \leq q_{ij}^h \leq \min \{M_i, A_j, V_k\} (i, j) \in C(k) \) for all \( (i, j) \), so that \( q^h \) is bounded, hence \( q^h \) lies in a compact convex set. The computation of \( \alpha^h \), in Step 2 of A, shows that if \( q^h \) lies in a compact convex set (which it does), then \( \alpha^h \) lies in a compact convex set.

**Lemma 5.** Any set \( S \) is compact if and only if every (infinite) sequence of points in \( S \) contains a subsequence which converges to a point in \( S \).

**Proof.** See a topology or nonlinear programming theory text (e.g., Lipschutz[18]).

**Lemma 6.** The sequence of \( z \) values generated by (Steps 1 and 2 of) Algorithm A converges to a single limit point \( z^* \). (This does not imply a unique \( (q^*, \alpha^*) \) corresponding to \( z^* \).)

**Proof.** The sequence of points \( (q^h, \alpha^h), (q^{h+1}, \alpha^h), h \in H = \{1, 2, \ldots, \infty \} \), generated by A lies in a compact set (Lemma 4) hence this sequence or some subsequence, say \( (q^h, \alpha^h), h \in H' \subseteq H \), thereof, converges to a limit point \( (q^*, \alpha^*) \), (Lemma 5). Then since \( z(q, \alpha) \) is a continuous, single-valued function we have \( z(q^h, \alpha^h) \rightarrow z(q^*, \alpha^*) = z^* \). But the sequence of \( z \) values generated by Steps 1 and 2 of A is nonincreasing and is bounded
below \((z \geq 0)\), hence, since some subsequence of this sequence converges to \(z^*\), the entire sequence must converge to \(z^*\).

**Lemma 7.** The Algorithm A generates a sequence or subsequence of points converging to \((q^*, \alpha^*)\), at which point \(P0^1\) has a global optimum \(z(q^*, \alpha^*)\).

**Proof.** Some subsequence \((q^{h'}, \alpha^{h'})\), \(h' \in H' \subseteq H\), of the points generated by Step 2 of A in a sequence of iterations, has a limit point \((q^*, \alpha^*)\), (by Lemmas 4 and 5). But Step 2 of A ensures that \(z(q^{h'}, \alpha^{h'}) \leq z(q^h, \alpha), \forall \) feasible \(\alpha\), hence since \(z(q, \alpha)\) is continuous we have at a limit point,

\[
z(q^*, \alpha^*) \leq z(q^*, \alpha), \forall \text{ feasible } \alpha \tag{1.1}
\]

Further, since the subsequence \((q^{h'+1}, \alpha^{h'})\), \(h' \in H' \subseteq H\), generated by Step 1 of A is contained in a compact set (Lemma 4), it or a subsequence \((q^{h''+1}, \alpha^{h''})\), \(h'' \in H'' \subseteq H'\), thereof has a limit point \((q^{**}, \alpha^*)\), (Lemma 5). But Step 1 of A ensures that \(z(q^{h''+1}, \alpha^{h''}) \leq z(q, \alpha^{h''}), \forall \text{ feasible } q\), hence since \(z(q, \alpha)\) is continuous we have at a limit point,

\[
z(q^{**}, \alpha^*) \leq z(q, \alpha^*), \forall \text{ feasible } q \tag{1.2}
\]

But, from Lemma 6, the \(z\) value is the same at all limit points, hence \(z(q^{**}, \alpha^*) = z(q^*, \alpha^*)\), hence (1.2) yields

\[
z(q^*, \alpha^*) \leq z(q, \alpha^*), \forall \text{ feasible } q \tag{1.2'}
\]

Since \(z(q, \alpha)\) is everywhere differentiable, (1.1) and (1.2') imply

\[
0 \leq \nabla_\alpha z(q^*, \alpha^*)^T d_\alpha, \forall \text{ feasible directions } d_\alpha \text{ at } (q^*, \alpha^*) \tag{2.1}
\]

\[
0 \leq \nabla_q z(q^*, \alpha^*)^T d_q, \forall \text{ feasible directions } d_q \text{ at } (q^*, \alpha^*). \tag{2.2}
\]

Adding (2.1) and (2.2) gives,

\[
\nabla_q z(q^*, \alpha^*)^T d_q + \nabla_\alpha z(q^*, \alpha^*)^T d_\alpha \geq 0
\]

which states that the directional derivative of \(z(q, \alpha)\) at \((q^*, \alpha^*)\) is positive in all feasible directions. But this is a sufficient condition for a global optimum of \(P0^1\) at \((q^*, \alpha^*)\), since \(P0^1\) consists of minimizing a convex differentiable function \(z(q, \alpha)\) over a convex feasible region.

**Theorem 1.** The sequence of points \((q^h, \alpha^h)\) and \((q^{h+1}, \alpha^h)\), \(h = 1, 2, \ldots\), generated by the algorithm A converges to a global optimum of \(P0^1\).

**Proof.** This follows immediately from Lemmas 6 and 7 above.

**Acknowledgments**

The authors wish to thank two anonymous referees for helpful comments.
REFERENCES


Received October 1979