# Extensions of Stochastic Multipath Trip Assignment to Transit Networks

Alan J. Horowitz

A procedure for applying stochastic multipath trip assignment to transit networks is described. The procedure extends an existing traffic assignment algorithm by (a) establishing strict criteria for transit zone definition; (b) using a comprehensive measure of disutility of transit trips; and (c) reconstructing the transit network so that all passenger movements are explicitly represented. The assignment procedure was tested on a large section of the Milwaukee County Transit System, which was specifically chosen to reveal any undesirable properties in the procedure. The assignment procedure was found to be free of those problems previously associated with applications of stochastic multipath traffic assignment in automobile networks.

Multipath trip assignment procedures have not yet been incornorated into the more widely used transit ridership forecasting models, such as the Urban Transportation Planning System (UTPS). Recent research on two lesser known models, EMME II (1) and the Transit Ridership Forecasting Model (TRFM) (2), has suggested that the validity of forecasts from UTPS-type models could be greatly improved with multipath trip assignment. The justification for still using a version of all-or-nothing assignment (3) in UTPS-type models is that existing multipath assignment procedures are extremely inefficient for large networks or will produce implausible assignments in some commonly encountered network structures.

The purpose of this paper is to demonstrate that a plausible and efficient multipath procedure can be built from existing theory. The basis of the procedure explored here is a stochastic, multipath trip assignment algorithm that originally was developed by Dial (4). Although it is considered to be efficient, Dial's algorithm has been correctly criticized for inaccurately representing travel behavior in many situations [see references (5) and (6) among others]. The research presented here shows that the undesirable properties of Dial's algorithm are of little consequence when transit networks are properly reconstructed as part of a larger multipath assignment procedure.

# STRUCTURE OF TRANSIT NETWORKS

Much of the criticism of Dial's algorithm concerns its performance in automobile networks. However, some obvious facts about transit networks, which distinguish them from networks of other modes, are presented here. Transit networks consist of many relatively independent routes (or lines). The routes are not physically interconnected; passengers wishing to use more than one route must change buses (or trains). Access to transit networks is typically accomplished by walking. Different routes often share portions of their alignments, and a passenger has the choice of a route for a single leg of the trip. Most transit networks have a number of difficult route designs, such as oneand two-way loops, turnbacks, branches, and skip stopping.

There are also two less obvious facts that affect multipath trip assignment. First, passengers dislike transferring and will avoid as many transfers as possible (7). Second, passengers also dislike long walks at either end of their trips; thus, there is an industrywide standard for a service area of one-quarter mile to either side of a route. Alternative paths through the network that are too arduous, because of particularly long walks or excessive transfers, will never be considered. Riders will look for a better path, choose a different mode, choose a different destination, or entirely forego the trip.

With these facts in mind, an inspection of any transit map will lead a casual, but objective, observer to conclude that there are only a few passengers who have a reasonable choice of alternative paths. This conclusion markedly contrasts with path choice on automobile networks, where almost everyone has many path choices. Consequently, a transit, multipath trip assignment algorithm must first determine those passengers who may have an acceptable choice of paths. Although there are numerous exceptions, these passengers generally have both trip ends within the service areas of two different routes (2). Next, the algorithm must split these passengers among a small set of reasonable paths, based on their relative merits.

Because the primary criterion of whether a passenger actually has a choice is the location of trip ends, a multipath trip assignment algorithm must have an amicable set of zones. It has been argued (2) that improper zone definition is the major source of error in UTPS-type transit assignments. A fundamental characteristic of good zones is that their boundaries must coincide with service area boundaries. Thus, any given parcel of land can be immediately categorized by its proximity to the various routes. An example set of zones is shown in Figure 1. Zones of this nature are required for TRFM and are not technically difficult to create. There are a number of ways to produce such a set of zones, although shifting data from a set of traffic analysis zones (TAZs), or a set of census tracts, can be problematical. Recently, methods have been developed for redefining existing zones and will be discussed in forthcoming sections.

Zones based on service areas may be smaller than zones based on typical TAZ criteria. A normal consequence of smaller zones is an increase in the number of origin-destination (O-D) pairs and an increase in execution time. This increase in execution time is mitigated by the significant fraction of zones (hatched areas of Figure 1) that are not proximate to any route and, thus, can be ignored by the assignment algorithm.

There is conventional wisdom among researchers that path

Center for Urban Transportation Studies, University of Wisconsin-Milwaukee, P.O. Box 784, Milwaukee, Wis. 53201.

#### Horowitz



FIGURE 1 Hypothetical transit network and associated set of zones.

choice in transit networks is governed by travelers' perceptions of the relative merits of their available alternatives. Direct evidence in support of this notion is largely anecdotal; strong, indirect evidence is derived from other travel choice processes, such as mode split. Accordingly, this notion has been implemented into recent multipath trip assignment procedures (including EMME II and TRFM) with reported good results. Because there is no accepted method of finding all the necessary parameters in a fully configured path choice model, the various parameters must be adopted from another source typically a mode split equation.

Another potential source of parameters for a path choice algorithm is psychological scaling (8,9). It has been shown that psychological scaling can produce a set of parameters consistent with those found from statistical estimation of mode split equations. In a psychological scaling experiment, subjects are asked to rate various trip descriptions. The ratings are independent of any choice process (mode choice or route choice). A particular advantage of psychological scaling is that subjects can be asked to evaluate infrequently encountered alternatives. Thus, it is possible to systematically build a much more complete model than is possible through other methods. For instance, TRFM uses essentially the same choice model for both mode split and trip assignment, and all but one of the default parameters are derived from a psychological scaling experiment.

The current research uses the disutility equation of the TRFM choice model because its source is well documented and it has undergone extensive testing as part of a complete modeling package. It should be pointed out that TRFM differs from UTPS-type models (the subject of this paper) principally because it is designed to forecast ridership on a single route. The conclusions of this study would probably remain unchanged if another, equally reliable, disutility equation were substituted.

The disutility of a transit trip may be represented in terms of weighted components of travel time and penalties for various actions (9). The units of disutility can conveniently be taken to be travel time, typically in minutes. Thus, when access to transit is by walking:

Disutility = (access walking time) (walking weight) + initial waiting penalty + (waiting time) (waiting time weight) + riding time + (transfer time) (transfer time weight) + (transfer penalty) (number of transfers) + (egress walk time) (walking weight) + (fare)/(value of time) (1)

The weights and penalties vary according to the environmental conditions for the particular trip component. For example, the transfer penalty has been noted to be considerably smaller for a timed transfer than for a normal, uncoordinated transfer (7). TRFM's default value of the transfer penalty under normal conditions is 23 min. This penalty is a conservative estimate; there is substantial evidence to suggest that the penalty should be larger—perhaps as high as 45 min. It should be noted that the full effect of a transfer in mode or route choice also includes the transfer time multiplied by the transfer time weight (defaulted at 1.6 for TRFM).

If path choice is to be made, strictly on the basis of disutility, then it is evident that the most important element of a transit trip is a transfer. It is, therefore, important that transfers be carefully represented in a multipath trip assignment procedure. This cannot be accomplished by simply using a more elaborate path choice model. Rather, it is necessary to perform a major reconstruction of the whole transit network. Network reconstruction will be discussed after a brief review of Dial's algorithm.

#### **REVIEW OF DIAL'S ALGORITHM**

Dial's algorithm is a clever modification of the standard Moore algorithm for finding the shortest path through networks. It requires some extra calculation and memory, but like the Moore algorithm, Dial's algorithm has a computation time that is roughly proportional to the number of links in the network. A plot of computation time against the size of network for path building and loading to a single trip destination from all trip origins is shown in Figure 2. For the record, these times were measured on an IBM-PC/AT, without a math coprocessor, running Turbo Pascal. A math coprocessor improves computation time by about 20 percent.

Dial's algorithm simulates the behavior of many people attempting to travel from a single origin to a single destination. An example trip is illustrated in Figure 3. (For those already familiar with Dial's algorithm: the sequence of events along this trip follows the backward pass.) As the travelers progress through the network, they encounter a number of intersections. Each intersection is a decision point. The algorithm assumes that the travelers have good, but not perfect, knowledge of what lies ahead. Any traveler will choose a direction (for example, a single link) on the basis of the shortest path disutility to the destination, given that direction. For example, at Intersection A



FIGURE 2 Computation time from all origins to a single destination.

it is found that there are two choices of direction. Link 1 has the lowest shortest path disutility to the origin (20 min versus 24 min), therefore this direction will draw most of the travelers. Some of the travelers progress to Intersection B. Here there are technically three choices of direction. However, Dial's algorithm eliminates Link 5 as a possible direction because travelers are taken farther away from their destination. Whatever paths are chosen, all travelers eventually reach their destination.

The behavioral underpinnings of Dial's algorithm may be succinctly stated: (a) travelers choose a direction (or link) at every intersection, (b) the only directions that will be considered are those that permit travelers to get closer to their destination, and (c) travelers are more likely to choose a direction that has a smaller shortest path disutility to the destination.



FIGURE 3 Example of path choice in Dial's algorithm.

Dial's specific functional form for handling choice of direction computes the probability  $(p_{ij})$  that a rider, presently at node *i*, chooses to travel to node *j*. That is:

$$p_{ij} = W_{ij} / \sum_{\substack{\text{all} \\ \text{feasible} \\ j}} W_{ij} \tag{2}$$

where

$$W_{ii} = \exp\left[\theta(d - d_i - d_{ii})\right] \tag{3}$$

and where

- $\theta$  = a calibrated parameter,
- $d_i$  = shortest path disutility from node *i* to the destination,
- $d_j$  = shortest path disutility from node *j* to the destination, and
- $d_{ii}$  = disutility on the link between node *i* and node *j*.

For feasible paths,  $d_i$  will always be larger than  $d_j$ , at most by the amount  $d_{ij}$ . It is important to note that the value of  $W_{ij}$ attains a maximum value of 1 for any direction that is on the shortest path to the destination.

Dial's algorithm considers only a subset of the paths between an origin and destination: those that always take travelers closer to their destination. As will be explained more fully, this subset is often quite small for transit networks. Typically, only one direction can be chosen at any intersection.

The most serious criticism of Dial's algorithm concerns a situation that occurs often in transit networks, for example, when there are more than two choices of direction. This situation is illustrated in Figure 4. Here a traveler is faced with a choice between one superior direction (Path 1) and four inferior directions that are just minor variations of the same path (Path 2). Dial's algorithm could assign many more travelers to all parts of Path 2 than to Path 1.



FIGURE 4 Illustration of an argument that Dial's algorithm is biased by the number of alternative directions.

#### Horowitz

A second look at this criticism is worthwhile. It can be argued that Dial's algorithm is indeed closely representative of normal travel behavior of transit riders. In transit networks, choices of direction are only possible at stops or at potential transfer points. Assume for this example that the differences in disutility between the two paths are due entirely to differences in in-vehicle time; headways for all directions are identical. The traveler, when making a choice of direction, is standing on the curb. Choice is largely related to chance. The traveler will most likely board the first bus that arrives at the transfer point. Because headways are equal, buses leading to Path 2 will arrive with four times the frequency of Path 1. Consequently, it is logical to expect Path 2 to be more heavily used. The aforementioned criticism of Dial's algorithm appears, at worst, to be a minor nuisance rather than a fatal flaw.

Users of UTPS are familiar with the concept of frequency split. When each of two (or more) bus routes entirely serves the same O-D pair, UTPS can be directed to split the trips between the routes according to their respective frequencies. It has been seen that Dial's algorithm, if properly implemented, will do exactly the same thing.

However, Dial's algorithm will not, in general, properly perform a frequency split if alternative paths involve more than a single route. Consider the situation, shown in Figure 5, of a choice between Path A and Path B, both of which involve a transfer. All headways and in-vehicle times are identical. It is external representation is the one that the network designer provides to the multipath assignment procedure, and the internal representation is the one that is actually used for path building and loading. For transit assignment, a strong case can be made for these two representations to be made distinct from each other.

The purpose of the external network is to permit the user to accurately transmit all relevant data to the mathematical model. The trend in recent years has been to show and edit the network graphically on a CRT display. Both TRFM and EMME II have this feature. Ideally, the external network should be free of extraneous detail; it should not contain artificial network elements (transfer links and centroid connectors); and it should be to scale. In other words, the external representation should look much like a system map that is provided to riders. Because there are only a few easily understood rules for drawing an external network, there is a strong likelihood that the network will be free of serious structural deficiencies.

The test network used for this research is shown in Figure 6. It is part of the Milwaukee County Transit System (MCTS). The network was drawn, and numerical data was entered through the General Network Editor (GNE) developed at the University of Wisconsin-Milwaukee. GNE is a graphics editor and network data-base manager that is dynamically configurable to nearly any type of transportation application.



FIGURE 5 Hypothetical network showing lack of symmetry in Dial's algorithm.

clear that Dial's algorithm will produce a 50/50 split at node U, even though Path A should be preferred. The better set of transfer opportunities at transfer point y is not reflected in the choice between Route 5 and Route 1 at the origin.

## TRANSIT NETWORK RECONSTRUCTION

It is rare that anyone considers the differences between the internal and external representations of a transit network. The



FIGURE 6 Test network—northeast portion of MCTS.

The purpose of the internal network, on the other hand, is to permit an accurate simulation of the behavior of travelers. It would include all the necessary artificial network elements, and thus would be substantially more complex than the external network. The way in which the external network is reconstructed into an internal network is based on assumptions about path choice decision processes. Consequently, the process of creating the internal network is part of the assignment model and not just an innocuous manipulation of data.

A highly formal reconstruction procedure has the additional advantage of standardization. Dial's algorithm is known to be sensitive to the way in which the actual system is represented as a network (10). With a formal reconstruction procedure, the algorithm is likely to yield the same result, regardless of who draws the external network.

Of particular concern to the current research is the reconstruction of transfer points. A reconstruction scheme is shown in Figure 7. The three-way transfer point is replaced by a subnetwork of four nodes and six two-way links. The original node is isolated from the rest of the network by three links that represent out-of-vehicle time. Three additional links represent the six possible transfers between the three routes. In general, an N-way transfer point requires (N + 1) nodes and [N + N(N - 1)/2] two-way links. It is interesting to note that when UTPS was first written for the IBM 7090 computer (3), transfer point reconstruction was considered but was rejected as being too computationally inefficient.



Three-way Transfer Point



Multiple-Route Links FIGURE 7 A network reconstruction scheme.

Also shown in Figure 7 is a reconstruction of links that each represent more than a single route. Such links occur when two or more routes share the same alignment; but there are other times when multiple-route links are useful. For example, routes with two branches are most accurately shown as two distinct routes. The joined portion of the route must then be described by a series of multiple-route links and transfer points. The transfer points ensure that riders cannot travel between the two branches without transferring.

As might be expected, the internal network can be considerably larger than its external counterpart. The transit network of Figure 2 has 130 nodes and 187 links. The internal network is almost three times as large, with 325 nodes and 574 links.

A straight application of Dial's algorithm on a reconstructed network can produce a small but annoying amount of frivolous transferring. That is, a few travelers will appear to make two or more transfers at a single transfer point. Frivolous transferring is easily eliminated by amending Equation 3 so that

$$W_{ii} = 0, \text{ if } -d_i + d_i + d_{ii} > a_k$$
 (4)

The value of  $a_k$  is set to be slightly less than the disutility on any transfer link at a given transfer point k. This constraint has an additional effect of eliminating otherwise feasible paths that are extremely poor choices.

### **TESTS OF THE ASSIGNMENT PROCEDURE**

The purpose of the following tests is to determine if the assignment procedure performs as expected: that it is conservative in its generation of paths; that it is unbiased with respect to the number of alternative directions at different points in the network; that it is symmetrical where it should be; and that it consistently applies a unified model of travel behavior. In other words, the tests should determine if the full procedure has properly dealt with the criticisms of Dial's algorithm.

The test network contains approximately one-fifth of the MCTS. This particular system was selected because it is essentially a grid with a few radial routes. A grid system provides a maximum of transfer possibilities and therefore offers the most demanding case for a multipath assignment procedure. This particular section of MCTS contains routes with branches, routes with one-way loops, other one-way sections, and multiple routes sharing the same alignment. The section also contains 26 multiple-transfer points, one of which is the intersection of five different routes. The test network has a service area with a population of about 200,000. All headways and running times are for the midday period. None of the transfers are coordinated.

Using GNE, data input procedures were designed to ensure that the full detail of the actual transit system was preserved. All possible path alternatives (rational or irrational) that were available to actual riders were available to the assignment procedure. The value of  $\theta$  (0.06) was established during an earlier study by running TRFM on two subnetworks of MCTS (2).

The tests consisted of 30 runs of the procedure, each assigning 100 riders from a single origin to a single destination. By inspecting the loadings on the links, it was possible to determine the feasible paths and the split of riders at each transfer point. The trips (O-D pairs) were not chosen at random. Instead, they were chosen in an attempt to force the procedure into producing odd results. The selected trips required an unrepresentatively large number of transfers. Twelve of the trips could be made without transferring; fifteen of the trips could be made with a minimum of one transfer; and three trips needed a minimum of two transfers. The trips were also unrepresentative in the number (15 out of 30) that had both their origin and their destination at transfer points. Most of the trips were selected either to follow a string of multiple-route links or to start, end, pass near, or pass through one of the more complex multipletransfer points.

It is difficult to obtain a clear idea of the performance of the algorithm without inspecting the CRT display of each result. The example trips shown in Figure 8 and the summary statistics given in Table 1 support the general conclusions that are made here.

All paths of two separate trips are shown in Figure 8. Trip A (going between Origin A and Destination A') has one of the more complex set of paths among the 30 trips. Origin A is just a regular stop, while Destination A' is a five-way transfer point. This trip requires a minimum of two transfers and, thus, would



be quite unusual. Even though there are many possible paths, the maximum number of feasible alternative directions faced by any rider is only two. There are four distinct paths involving seven different routes. None of the paths had more than two transfers.

The second trip (B to B') in Figure 8 is much more typical of the 30 test runs. This trip requires a minimum of one transfer. Origin B is a four-way transfer point. Only two paths, both quite reasonable, were generated. Any other path would have required more than the minimum of one transfer or would have required considerably more in-vehicle time. The difference in ridership (70 percent to Route 22 and 30 percent to Route 60) between the two paths is due entirely to the relatively small transfer time from Route 22 to Route 30.

Statistics on path generation for all 30 trips are summarized on Table 1. Note that there is a strong relationship between the number of paths generated and the minimum number of transfers required. Trips that required no transfers, with just two

TABLE 1SUMMARY OF PATH GENERATION FOR 30 TESTRUNS

Minimum No. of Transfers	No. of Test Runs	Mean Paths Generated	Mean Transit Disutility	Mean Automobile Disutility
0	12	1.17	44.8	10.8
1	15	1.87	87.2	14.6
2	3	2.67	137.7	14.2

exceptions, had only one feasible path. More than one-half (8 of 15) of single-transfer trips had only one feasible path, although there were trips with as many as six paths. Trips that required at least two transfers averaged less than three paths. Because of large disutility differences between automobile and transit modes, multiple-path trips are less likely to be chosen by potential transit riders.

#### **DISCUSSION OF THE PROCEDURE**

The assignment procedure requires more computation time than a UTPS-type, all-or-nothing assignment algorithm. The increase in computation time depends on the number of multiple-transfer points and the sizes of the original zones. Based on the MCTS network, increases in computation time of between 200 to 400 percent should be anticipated.

It is important to recognize that the described assignment procedure cannot be readily applied to most existing transit networks or to most existing sets of traffic analysis zones. Incompatibility of networks is not particularly serious, given the new generation of network editors (such as GNE). For example, the UTPS network for MCTS could be completely redrawn in approximately 1 person-week. Incompatibility with existing sets of zones is less easily solved because (a) the zones required for the assignment may differ from those required for other model steps, and (b) all travel data has already been aggregated to the original zones. Reaggregation of data can be prohibitively expensive.

Areal interpolation is a promising method of developing a new set of zones from an old set (11,12). Inputs to an areal interpolation program include boundaries of both source and target zones and a statistic (number of households) on each source zone. The program then estimates the statistic for each target zone. For example, Tobler's Pycnophylactic Histospline Interpolation Model has been successfully used at the Center for Urban Transportation Studies for a comparable problem moving demographic data from census tracts to TRFM zones (13). With areal interpolation it is possible to retain the original TAZs for other parts of the travel demand forecast, while adopting entirely new zones for transit assignment. In this case, it is only necessary to interpolate the transit trip table.

The multipath assignment procedure should not affect current iterative methods of handling user equilibrium in transit networks (14). Larger-than-planned link volumes manifest themselves by (a) increasing boarding and alighting times, (b) increasing the amount of standing and thereby increasing negative perceptions about riding time, and (c) prompting the transit operator to add buses or trains. The combined effects can be quite complex; in many cases an increase in ridership reduces disutility as additional buses are supplied. Much more research is required before a practical and completely integrated equilibrium or multipath trip assignment procedure can be assembled.

## CONCLUSIONS

The assignment procedure described in this paper retains the speed and memory efficiency of Dial's algorithm (3). However, the procedure requires a specially configured system of zones based on service area boundaries, and a network on which every possible transfer and waiting period are explicitly represented by links. These artificial links are not seen by the network designer but are created during a network reconstruction step in the procedure.

The extensions produced an assignment procedure that behaved in a manner consistent with current understanding of how transit riders choose paths. Tests of the procedure, under unusually harsh network conditions, failed to reveal any of the undesirable traits that had been attributed to Dial's algorithm. For the most commonly made trips, the procedure was properly conservative in path generation. The procedure revealed only one best path between an origin and destination, unless there were close alternatives. When alternative paths were generated, the split of riders among the paths closely followed the widely accepted principle of frequency split. Because of the dominating effect of out-of-vehicle time on path disutility, Dial's algorithm works well on transit networks, even though it had been dismissed as unusable by other researchers.

## ACKNOWLEDGMENTS

Several comments were made by the following on various aspects of the trip assignment procedure as it was being developed: Joe Ossi, Ron Jensen-Fisher and Tom Hillegass, UMTA; Edward Beimborn and Peter Lindquist, Center for Urban Transportation Studies; and Roger Tobin, GTE Laboratories. GNE was coauthored by Anand Pithavadian. The research was funded by UMTA's University Research and Training Program.

#### REFERENCES

- 1. H. Spiess. On Optimal Route Choice Strategies in Transit Networks. Centre de Recherche sur les Transports, Université de Montréal, Canada, 1983.
- A. J. Horowitz and D. N. Metzger. Implementation of Service Area Concepts in Single-Route Ridership Forecasting. In *Transportation Research Record 1037*, TRB, National Research Council, Washington, D.C., 1985, pp. 31–39.

- R. B. Dial. Transit Pathfinder Algorithm. In *Transportation Research Record 205*, HRB, National Research Council, Washington, D.C., 1967, pp. 67–85.
- R. B. Dial. A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration. *Transportation Research*, Vol. 5, 1971, pp. 83–111.
- R. L. Tobin. An Extension of Dial's Algorithm Utilizing a Model of Tripmakers' Perception. *Transportation Research*, Vol. 11, 1977, pp. 337–342.
- C. F. Daganzo and Y. Sheffi. On Stochastic Models of Traffic Assignment. *Transportation Science*, Vol. 11, No. 3, Aug. 1977, pp. 253-273.
- A. J. Horowitz and D. J. Zloscl. Transfer Penalties: Another Look at Transit Rider's Reluctance to Transfer. *Transportation*, Vol. 10, 1981, pp. 279–282.
- A. J. Horowitz. The Subjective Value of the Time Spent in Travel. Transportation Research, Vol. 12, 1978, pp. 385–393.
- A. J. Horowitz. Subjective Value of Time in Bus Transit Travel. Transportation, Vol. 10, 1981, pp. 149–164.
- Y. Sheffi. A Note on the Turn and Arrival Likelihood Algorithms of Traffic Assignment. *Transportation Research*, Vol. 13B, 1979, pp. 147-150.
- N. S. Lam. Spatial Interpolation Methods: A Review. The American Cartographer, Vol. 10, 1983, pp. 129–149.
- W. R. Tobler. Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association*, Vol. 74, 1979, pp. 519–530.
- P. S. Lindquist. Histospline Interpolation for Data Input into the Transit Ridership Forecasting Model. *ITE Journal*, Vol. 56, No. 11, Nov. 1980.
- P. R. Stopher, W. A. Davidson, L. K. Tamny, and G. S. Spivack. Using Regional Forecasting Models of the Urban Transportation Planning System for Detailed Bus Route Analysis. *Transportation Research Record 1037*, TRB, National Research Council, Washington, D.C., 1985, pp. 52–58.

Publication of this paper sponsored by Committee on Public Transportation Planning and Development.