TOWARD AN EFFICIENT HIGHWAY SERVICE ESTIMATOR FOR CONGESTED NETWORKS

Hubert Le Menestrel and Bruno R. Wildermuth, Wilbur Smith and Associates, Columbia, South Carolina

This paper presents a new approach to obtain estimates of highway services on congested networks. Specifically, a 1-pass incremental capacity-restraint procedure is detailed that generates logical paths for reuse with the average network to obtain travel-time and travel-cost estimates. The method is based on the multiple-routing principle for path generation and Schneider's 1-pass capacity-restraint technique. The operation of the procedure is illustrated with an example, and results of several analyses are also presented. Suggestions are made for further research to establish reliable procedures for the estimation of average speeds and to investigate further the validity of the capacity-restraint paths in relation to the loading sequence used by the procedure.

•THE RAPIDLY growing expenditures for urban facilities at all levels of government have been paralleled by increased competition for public funds and by public awareness of the need to set community goals and objectives within which the full range of possible alternatives must be evaluated. Consequently, more demanding evaluation and decision criteria need to be established to facilitate complete consideration of the impacts of alternative transportation systems on the total urban system.

The transportation planning process provides the basic structure with which the characteristics of alternative networks can be measured and their respective abilities to satisfy objectives and goals evaluated. It provides the travel demand forecasts that serve to evaluate the potential cost and benefit patterns. However, this process is often limited and cumbersome in its capabilities for estimating service aspects of alternative transportation systems.

Within the past 15 years, urban transportation planning has relied increasingly on mathematical models and electronic computers as its basic tools. During this period, both the models and the computer systems have steadily been improved, although advancements in the design of computers have far outstripped those of analytical models.

Primary emphasis in the development of models and associated computer programs has been on analysis and estimation of travel demands and capacity evaluations. Increasingly, added emphasis is also placed on the modal-choice decision process and systems evaluation procedures.

A recent report (1) identified the following means that should be provided by transportation systems evaluation procedures:

1. Measuring and valuing the effects of transportation system changes on travelers and on community residents;
2. Estimating the system costs of transportation changes; and
3. Relating transport changes to high-level goals and to overall framework for decision-making.

To operate efficiently, service estimators and other systems evaluation models must be designed to accept travel demand projections and network descriptions from other transportation analysis models. Furthermore, their outputs must be structured for complete compatibility with other models to facilitate the explicit incorporation of service variables into trip generation, distribution, modal-split, and assignment models. Only then will it be possible to supply these models with measures of transportation services (e.g., travel times and travel costs), which are consistent with the anticipated levels of network congestion.

To date, major emphasis in evaluating alternative transportation systems has been on travel benefits. Estimates of transportation systems services, therefore, have been limited largely to measures of place-to-place travel times. Such estimates are also used most often to provide inputs and feedbacks to many travel demands and modal-choice projection models.

Not infrequently, however, analysts have ignored the feedback effects of congestion on the generation of demand and the projection of modal-choice decisions. The main reason for ignoring these feedback effects undoubtedly are the cumbersome and costly efforts required to obtain service estimates from existing analysis tools. Existing packages for urban transportation planning provide a number of programs that, when properly sequenced, can produce estimates of zone-to-zone travel times and travel costs for congested networks.

For highways, the programs required are capacity-restraint, link-costing, and skim-impedance-path programs. Capacity restraint in itself requires a series of 3 programs to be run in repeated succession. However, application of the capacity-restraint series of programs will only produce the data that, when averaged, will reflect the network conditions under the given level of congestion. This network can then serve to determine minimum-impedance paths from which zone-to-zone travel times and travel costs can be estimated.

The remainder of this paper describes a fast and efficient method that produces reliable estimates of zone-to-zone travel times and travel costs for congested highway networks. This procedure, which is easily operated at a fraction of the cost of other methods, has been tested on a number of highway networks for different urban areas.

METHODOLOGY

To obtain estimates of highway services for congested networks requires an efficien' capacity-restraint method. The method presented here produces, with 1 application and without iterations, an average network description (reflecting the level of congestion) and a set of valid "best path" trees along which zone-to-zone travel times and travel costs can be accumulated from the average network. The technique is based essentially on 2 concepts: a recently developed "multiple-routing" best path procedure that utilizes a stochastic approach and a modified version of Schneider's incremental 1-pass capacity-restraint model.

Multiple-Routing Principle

A procedure has been devised that eliminates the most critical problem of the minimum-path procedure and routes directional paths between pairs of nodes over all routes whose times differ from that of the shortest route by less than a given proportion. This procedure, called multiple routing, has made it possible to ensure proper balance of flows between alternate routes of equal merit. Theoretically, the multiple-routing principle allows the directional path between any pair of nodes to be different for each path generated. In practice, of course, the number of different paths is limited by the acceptable alternatives available within the network configuration. Multiple routing assumes that drivers associate a "perceived" travel time with each link of the network and that not all drivers will necessarily perceive the same time for a link.

Stochastic Approach—The multiple-routing principle is based on a stochastic approach that adds a long-sought element of realism to traffic assignments. Use of the stochastic process for the multiple-routing principle is limited to the generation of perceived link times that are drawn at random from a normal distribution of times for each link. Thus, maximum use can be made of existing concepts by relying on the fact that in the average network each pair of nodes will be traversed many times because of the large number of centroids normally used.

This approach requires very little additional computer time because only 1 tree is generated for each centroid. The only extra time is that required to calculate each time a link is being considered, a random sample from a normal distribution with a mean equal to each link's given impedance value.

Because most routes consist of a string of links, it is not possible for the stochastic process to generate unreasonable routings. The sum of perceived link times will differ on the average from the sum of mean link times by a smaller percentage for long routes than for short routes with few links. This corresponds well with the real world. For example, a driver might choose a route that takes 7 min, when the best available route takes only 5 min, and thus accept a 40 percent increase in time. On the other hand, it seems much less likely that anyone would choose a route that takes 70 min when the best route available takes only 50 min. Clearly, the multiple-routing principle reflects likely actual route choices by drivers.

Effect of Alternate Routes on Link Flow—In the application of the multiple-routing principle as presented here, all trips from 1 centroid to another centroid will use the same path; in effect, this is the all-or-nothing principle. However, not all trips passing from 1 node to another will use the same path, provided, of course, that the network consists of a reasonable number of centroids. The incidence of multiple-source centroids, therefore, serves to distribute the traffic flow onto all acceptable alternative routes. In the method described here, the route chosen for a particular trip may to some degree be arbitrary, in the sense that it is determined by a process of chance. The total traffic flows, however, on each link are clearly insensitive to the stochastic process.

One-Pass Incremental Capacity-Restraint Procedure

Previous research and experience with capacity-restraint procedures have indicated that to achieve stable speed-flow relations in congested networks requires either an incremental procedure or an average of the results of several iterations with the all-or-nothing method. Stability of speed-flow is attained when the use of one or more iterations or increments would not significantly alter the flow or the speed of any link in the network. The application of the multiple-routing method of building best paths will reduce the amount of imbalance generated for each assignment step. Hence, a reduction in the number of assignment increments or the number of iterations can be justified.

Alternatively, in combination with multiple routing, it is possible to devise a more efficient way of selecting increments of the total traffic demand for assignment at each step. The 1-pass incremental procedure based on Schneider's basic approach presents such an alternative.

Schneider's Assignment Concept—Schneider's basic philosophy was founded on the assumption that each unit of traffic seeks to travel along the best available path. Furthermore, the addition of each unit of flow to a link would, in effect, reduce the speed of travel on that link and, therefore, create a different best path situation for each additional unit of traffic demand desiring to move through the network.

The method, originally developed for the Chicago Area Transportation Study (2) combined trip distribution and traffic assignment into a single comprehensive process. Only the assignment concept, which substantially reduces the sometimes troublesome effects caused by changes in the loading sequence, is used here. It is a 1-pass incremental process whereby the fraction of the total traffic demand loaded during each increment is the total demand from 1 origin centroid.

Trips from the first origin centroid in the randomly selected loading sequence are loaded on the minimum-path tree determined on the basis of the input travel times. Link times are then updated according to an exponential time versus volume-capacity function. The travel demand from the next centroid in the random sequence is then loaded along the minimum path determined from the revised network times. This process is repeated until the entire traffic demand matrix has been assigned. A randomly selected loading sequence is used to minimize the effects of restraint on routings and travel times for traffic of the same geographic origin.

An Approach Using Multiple Routing—Schneider's incremental method was efficient in its use of only 1 tree per origin centroid. However, it required an excessively large number of network modifications (1 per origin centroid). Apart from conceptual reasons, the frequent modifications served to reduce the problems associated with the use of minimum-path, all-or-nothing procedures. Multiple routing has inherent balancing qualities that allow, in combination with the incremental 1-pass technique, a substantial reduction in the frequency of network modifications without loss of stability.

The increment of traffic demand that can be allocated between successive network updates, without excessively increasing the flow on any 1 link, can readily encompass traffic demands from several centroids as long as drastic changes in travel times are being avoided. The probability of any link being used by 2 or more successive trees is very small if a random-loading sequence is used because each origin centroid is normally separated spatially from the preceding and the following centroid. In addition, the characteristics of multiple routing further reduce the probability that successive trees will use the same path between pairs of nodes.

As the flow on many links reaches the practical capacity level, it is more important to load smaller fractions at each increment. This is readily controlled by selecting fewer origin centroids from the predetermined sequence. The total number of network modifications, which account for the increase in computer time above that of the all-or-nothing technique, is reduced by a factor of 10 or more compared to Schneider's original concept.

Speed-Flow Relation—Schneider used an exponential function to modify travel times. This assumption does not conform to actual speed-flow relations observed on highways but yields increasingly lower speeds with increasing volume-capacity ratios. A "real-life" speed-flow function, on the other hand, would approach a vertical tangent at maximum attainable capacity. The restraint function, however, must not be allowed to preclude altogether the assignment of flow beyond the maximum capacity level. This is because one of the functions of traffic assignments is to identify potential capacity deficiencies within travel corridors of transportation networks.

The function used must first of all satisfy the principle of the method rather than direct itself to a precise simulation of the real world. Links that are loaded with a volume reaching capacity after only a small percentage of the total demand has been processed require more drastic modifications than links that reach the same volume-capacity ratio at the end of the loading process. In short, the model must anticipate the effects of the additional demand still to be loaded and react in a way that will prevent (to a certain extent) its volumes from exceeding its capacity by 2 or 3 times.

Stability and Convergence of Procedure—The 1-pass incremental method permits estimates of speeds and flows in congested networks with a minimun of computer time. In addition, it avoids the nonconvergent unbalanced estimates of flows that characterize iterative methods, as illustrated by the following example.

Figure 1 shows a portion of a larger network in which many trees are routed between points A and B. Three alternative paths exist between A (a bridge head) and B (the main access point to a major center). Path 2 is an old road with a directional capacity of 2,500 vehicles, while paths 1 and 3 are newer, wider facilities with directional capacities of 6,000 vehicles each. Travel times at practical capacity are 36, 34, and 37 integer time units respectively (most network analysis computer programs work with integer values). Volumes entering from 5 assumed source-centroids are as shown.

Table 1 gives the progression of calculated travel times and volumes loaded on each path for 4 iterations of the FHWA capacity-restraint procedure. The first iteration (free flow) loads the total demand of 5,500 vehicles onto path 2, the old narrow road. This leads to drastic overloading and consequently a substantial increase in travel time for path 2. Times for paths 1 and 3, on the other hand, are reduced because of the absence of flow. The second iteration loads all traffic onto path 1. This leads to a downward adjustment of travel times for paths 2 and 3. In iterations 3 and 4, traffic is loaded alternatively on path 3 and 1, indicating the beginning of a nonconvergent shifting process. The example clearly establishes the need to average the flows of all iterations to achieve reasonably stable flows. In fact, for the particular example, 4 iterations are not sufficient to attain stability.

Figure 1. Network configuration with 3 alternate paths between A and B.

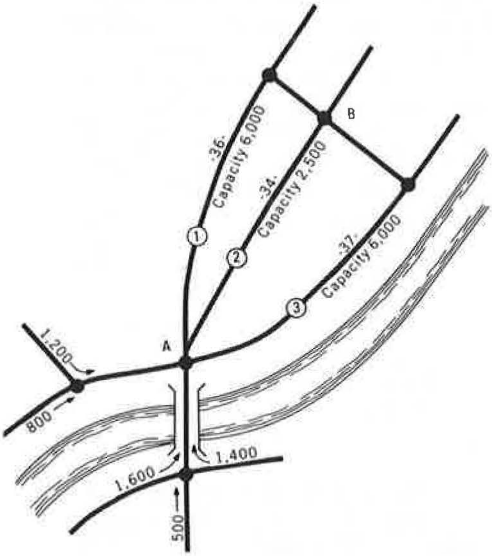


Table 1. Path times and flows for 4 iterations with averaging method.

	Path 1		Path 2		Path 3	
Iteration	Time	Flow	Time	Flow	Time	Flow
1	36	0	34	5,500	37	0
2	35	5,500	59	0	36	0
3	35	0	57	0	35[a]	5,500
4	34	5,500	55	0	35	0
Avg		2,750		1,375		1,375
Final V/C ratio		0.46		0.55		0.23

[a] It is assumed that, in the case of equal times, path 3 rather than path 1 will be chosen. If path 1 were chosen first, path 3 would be selected in iteration 4. In either case, overall results would not be affected.

Table 2. Path times and flows for 5 centroids with 1-pass incremental method.

	Path 1			Path 2			Path 3		
Increment	Time	Flow	Cumulative Flow	Time	Flow	Cumulative Flow	Time	Flow	Cumulative Flow
1	31	0	0	30	1,200	1,200	32	0	0
2	31	800	800	38	0	1,200	32	0	0
3	32	0	800	38	0	1,200	32	1,600[a]	1,600
4	32	500	1,300	38	0	1,200	36	0	1,600
5	34	1,400	2,700	38	0	1,200	36	0	1,600
Final V/C ratio		0.45			0.48			0.27	

[a] If the stochastic process would have selected path 1 instead of path 3, final flows would be 2,400 for path 1 and 1,900 for path 3.

Table 3. Path times and flows for 100 centroids with 1-pass incremental method.

	Path 1			Path 2			Path 3		
Increment	Time	Flow	Cumulative Flow	Time	Flow	Cumulative Flow	Time	Flow	Cumulative Flow
1	31	385	385	30	605	605	32	110	110
2	32	440	825	34	165	770	32	445	605
3	34	495	1,320	36	55	825	33	550	1,155
4	36	330	1,650	36	330	1,155	35	440	1,595
5	36	495	2,145	38	220	1,375	37	385	1,980
Final V/C ratio		0.36			0.55			0.33	

Table 2 gives the progression of volumes and travel times with the 1-pass incremental method. For the example, with only 5 centroids, each increment's traffic demand is assumed to be that of 1 centroid. The initial times for all paths are originally set equal to 0.87 times the value given at a flow level equal to practical capacity. (This value corresponds to the ratio of the travel time under free-flow conditions versus the travel time at a flow equal to practical capacity as defined by the Highway Capacity Manual.) Centroids are assumed to be selected for loading from left to right. A flow of 1,200 vehicles is assigned first to path 2 and results in a substantial increase in time on that path. The demand for increment 2 is loaded onto path 1 and increases that path's travel time. The third increment will choose either path 3 or path 1 depending on the random samples generated. This results in a condition where the other path will receive the flows from the remaining 2 increments as illustrated.

This example demonstrates the economical operation of the proposed method. With only 5 trees and 4 network modifications, a speed-flow condition was achieved that appears more stable than that obtained with 20 trees (4 iterations with 5 trees each) and 3 network modifications. Because network modifications require only marginal amounts of computer time compared to building sets of trees, it is obvious that the 1-pass method achieves substantial savings. The added cost of building multiple-routing best path trees (approximately 15 percent more than minimum path trees) does not substantially affect these cost savings.

The smoothness with which the 1-pass method increases the volumes on each path from 0 to their final flow is barely indicated by this example. In a real network with several hundred origin centroids, the smoothness of increase in flows would be much more pronounced because the volume loaded along each tree would represent a smaller fraction of the final flow. If there were 100 centroids instead of 5 and an average volume of 55 trips per centroid, the loading process for the sample network would proceed as given in Table 3. With 20 trees being generated and loaded during each increment, the benefits of multiple routing are now clearly apparent. Modifications of travel times are more gradual and each path receives some flow in every increment.

Calculation of Average Link Speed—The average speed on any link in a congested network is determined from its flow and the total vehicle time traveled on the link. Total vehicle time can be expressed by the sum of the products of the number of vehicles assigned during each increment (between 2 network modifications) and the travel times at which those vehicles have been assigned. This is not easy to calculate because core storage limitations preclude keeping a separate accumulator for each link.

Most capacity-restraint assignment procedures solve this problem by calculating the sum of the products of the total assigned volumes by an average travel time calculated from the input time and the time used during the last increment or the last iteration. Analysis of this problem suggests that the area under the time-versus-volume curve would yield an acceptable approximation of vehicle time. Use of the area under the curve will result in a slight overestimation of the correct value caused by the assumption of a smooth curve instead of the resulting step function (Fig. 2); however, with proper adjustment, the calculated value presents a reliable estimate.

Travel-time calculations for paths 1, 2, and 3 of the 5-centroid example previously illustrated would result in average values of 34, 33, and 33 time units. For the 100-centroid example (Table 3), average values would correspond to 33, 34, and 34 time units respectively. The corresponding estimates for the iterative method, derived from the average volumes and the fourth-power relation, are 32, 30, and 32. If the estimates are made on the basis of speeds used for each iteration weighted by the volume loaded, the values would be 35, 34, and 35 time units. The resulting values for the 1-pass method clearly fall within the range of estimates from the iterative method, indicating a need for further research on this topic.

Fraction of Demand Assigned by Each Increment—The accuracy of assignment results is related to the amount of traffic assigned to each path and the fraction of the total traffic demand assigned between network modifications. This fraction in turn is controlled by the volume originating at each centroid and the number of centroids. The procedure, therefore, allows the user to choose the number of trees built between modifications.

Figure 2 shows the accumulation of volume for a typical link located centrally within a small network. The volume buildup is shown for assignments loading 5 and 20 trees respectively between network modifications. Results obtained in each case are quite similar, although a more reliable average speed estimate would have been expected from the loading based on 5-tree increments.

The program has been prepared with an additional routine to decrease the fraction of traffic demand loaded at each increment whenever the accumulation of flow exceeded a set level between successive network modifications. This procedure evaluates, after each modification, the average factor by which link times have been modified. Whenever the difference between this factor and the one obtained in the previous increment reaches the set level, the number of trees is decreased by a percentage calculated from a parabolic function of the difference.

Validity of Individual Trees

The 1-pass incremental procedure, in addition to its efficient operation, generates trees that follow realistic and logical routes and can be used for subsequent purposes such as analysis of selected links and accumulation of service parameters along their paths. As a result of the randomly selected loading sequence, most links start receiving flows during early increments. In a typical run with 19 increments, more than 45 percent of all links showed a flow with only 8.2 percent of the total demand loaded. With 20.2 percent of the volume processed, almost 67 percent of all links had received flows; and with 39.1 percent of the total demand loaded, more than 80 percent of all links showed flows.

The multiple-routing procedure, as illustrated in the 100-centroid example (Table 3), will load flows on competing routes at an even pace and thereby ensure the smoothness of the 1-pass capacity-restraint method and the validity of its individual trees.

Comparison of Highway Service Estimates—Service estimates accumulated along the paths generated by the capacity-restraint procedure were verified by comparing zone-to-zone travel times and travel distances with estimates obtained along the minimum-time paths of the average network. Values accumulated along the capacity-restraint paths were expressed as a percentage of those obtained from minimum-time paths and grouped according to 6 value ranges.

The data from a 165-centroid network with a high degree of congestion showed travel-time values that were 1 percent higher on the average for capacity-restraint paths. More than 75 percent of all capacity-restraint paths had travel-time values that were within 2 percent of minimum-path times. Only 12 paths contained differences greater than 5 percent with extreme values of 94 percent and 114 percent respectively.

Distance values accumulated along capacity-restraint paths were found to be 3 percent higher than those along the minimum-time path. However, this does not necessarily cause concern because assignments based on minimum-time paths have been found to underestimate vehicle-miles of travel. Table 4 gives the frequency of capacity-restraint paths by percentage difference of travel-time and travel-distance values.

Of more interest perhaps than the variation in average values is the occurrence of differences within the loading sequence of the capacity-restraint procedure. Figure 3 shows a plot of the average percentage difference of time and distance for each group of 10 successive paths. The plot of travel-time differences seems to indicate a slight but definite trend toward increasing values for the paths generated in the second half of the loading sequence. The same trend, but much more pronounced, is found for the distance values.

Detailed analysis shows that group 16, for which the accumulated distances over the capacity-restraint paths are 10 percent higher than those of the minimum-time path, contains 2 of the 3 most extreme paths. Another extreme path is located in group 15 and, like the other two, originates from a zone bounding the external cordon. Further investigation of the few extreme paths has not yet been undertaken, but a preliminary analysis indicates that peculiar network conditions may be the major cause of the differences. The findings suggest, however, a definite need for further detailed analysis, including plots of individual path traces.

Figure 2. Progression of V/C for loading of centrally located link for different numbers of trees built between network modifications.

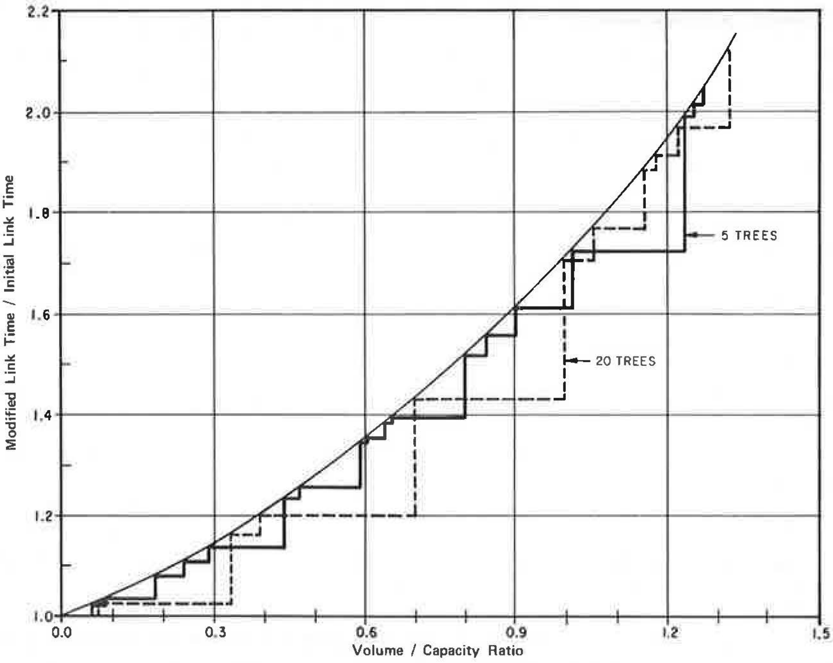
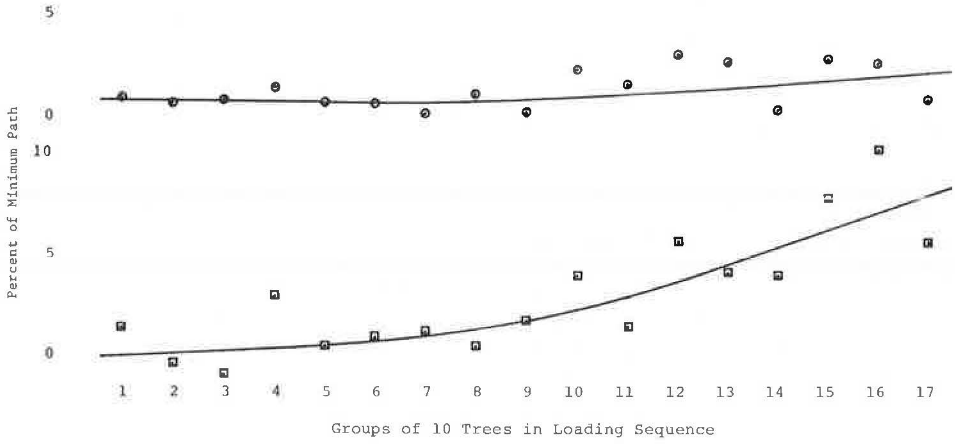


Table 4. Frequency of capacity-restraint paths by percentage difference in time and distance.

Percent of Minimum Path	Travel Time (number of paths)	Travel Distance (number of paths)
<95	5	2
95	0	1
96	0	2
97	2	3
98	0	1
99	6	16
100	42	17
101	52	29
102	27	25
103	9	16
104	8	10
105	5	9
106	1	9
107	1	10
108	2	3
109	1	3
>109	4	9
Total	165	165
Mean	101.4	103.1

Figure 3. Percentage difference in accumulated time and distance for capacity-restraint path by loading sequence groups.



CONCLUSIONS

The 1-pass incremental capacity-restraint program presented in this paper provides a flexible and economical tool for the estimation of highway service measures on congested networks. Computer costs experienced with this method indicate that stable estimates of speeds and flows can be obtained at costs that are, at most, 40 percent more than those required for a minimum-path all-or-nothing assignment. This compares with an increase of 350 to 400 percent when other capacity-restraint methods are used, not counting the cost to generate final sets of trees that properly reflect the level of congestion.

A number of areas have been identified that suggest the need for additional research including the calculation of average speeds for congested links and the detailed analysis of individual tree traces in relation to the loading sequence and the level of congestion.

The 1-pass procedure provides an efficient and realistic method to obtain service estimates for the evaluation of networks and modal choice under conditions of congestion. Because this procedure does not require cumbersome sequencing of many different programs, it should be of interest to those concerned with analyses and evaluation of alternative transportation networks.

REFERENCES

1. Transportation Network Evaluation: Costable Criteria. Stanford Research Institute, Nov. 1969.
2. Data Projections. Chicago Area Transportation Study, Vol. 2, July 1960.