HIGHWAY RESEARCH RECORD

Number 6

Traffic Assignment

4 Reports

Presented at the
42nd ANNUAL MEETING
January 7-11, 1963

HIGHWAY RESEARCH BOARD
of the
Division of Engineering and Industrial Research
National Academy of Sciences—
National Research Council
Washington, D. C.
1963

Department of Traffic and Operations

Fred W. Hurd, Chairman
Director, Bureau of Highway Traffic
Yale University, New Haven, Connecticut

COMMITTEE ON ORIGIN AND DESTINATION

Garland E. Marple, Chairman
Chief, Urban Planning Division
U. S. Bureau of Public Roads
Washington, D. C.

Charles F. Barnes, Jr., Secretary
Alan M. Voorhees and Associates
Bridgeport, Connecticut

John E. Baerwald, Professor of Traffic Engineering, University of Illinois, Urbana
Howard W. Bevis, Logistics Department, The RAND Corporation, Santa Monica, California
Glenn E. Brokke, U. S. Bureau of Public Roads, Washington, D. C.
J. Douglas Carroll, Jr., Deputy Director, Tri-State Transportation Committee, New York, New York
Nathan Cherniack, Economist, The Port of New York Authority, New York, New York
Donald E. Cleveland, Assistant Research Engineer, Texas Transportation Institute, Texas A & M College, College Station
M. A. Conner, President, Mel Conner and Associates, Inc., Tallahassee, Florida
E. C. Cross, Jr., Bureau of Highway Planning and Programming, New York State Department of Public Works, Albany
Thomas J. Fratar, Partner, Tippetts-Abbett-McCarthy-Stratton, New York, New York
Harold W. Hansen, Senior Planning Engineer, Portland Cement Association, Chicago, Illinois
Walter G. Hansen, Alan M. Voorhees & Associates, Washington, D. C.
Joseph Kates, Traffic Research Corporation, Ltd., Toronto, Ontario, Canada
Louis E. Keefer, Director, Transportation Study, Southeastern Wisconsin Regional Planning Commission, Waukesha, Wisconsin
Norman Kennedy, Professor of Transport Engineering, Institute of Transportation and Traffic Engineering, University of California, Berkeley
Warren Lovejoy, The Port of New York Authority, New York, New York
John T. Lynch, McLean, Virginia
William R. McGrath, Director, Department of Traffic and Parking, New Haven, Connecticut
John K. Mladinov, Director, Puget Sound Regional Transportation Study, Seattle, Washington
Albert J. Mayer, Institute for Regional & Urban Studies, Wayne State University, Detroit, Michigan
W. L. Mertz, Tri-State Transportation Committee, New York, New York
Harold L. Michael, Associate Director, Joint Highway Research Project, Purdue University, Lafayette, Indiana
Frank J. Murray, Engineer of Planning Survey, Ohio Department of Highways, Columbus
R. Parker, Head, Research Section, Planning and Research Division, Arkansas State Highway Department, Little Rock

William S. Pollard, Jr., Partner in Charge, Harland Bartholomew & Associates, Memphis, Tennessee
Lloyd A. Rivard, Planning and Programming Engineer, Automotive Safety Foundation, Washington, D. C.
Paul Shuldiner, Northwestern University, Evanston, Illinois
S. S. Taylor, General Manager, Department of Traffic, Los Angeles, California
Alan M. Voorhees, Alan M. Voorhees and Associates, Washington, D. C.
Lawrence S. Waterbury, Consulting Engineer, New York, New York
George V. Wickstrom, Director, Transportation Division, Penn-Jersey Transportation Study, Philadelphia, Pennsylvania
David K. Witheford, Bureau of Highway Traffic, Yale University, New Haven, Connecticut
Martin Wohl, Transportation Consultant to Assistant Secretary for Science and Technology, U. S. Department of Commerce, Washington, D. C.
J. E. Wright, Traffic Survey Manager, Texas Highway Department, Austin
F. Houston Wynn, Wilbur Smith and Associates, New Haven, Connecticut

Contents

TRAFFIC ASSIGNMENT ANALYSIS AND EVALUATION
David K. Witheford 1

A COMPARATIVE DESCRIPTION OF A CAPACITY-RESTRAINED TRAFFIC ASSIGNMENT
Robert B. Smock 12

A CAPACITY-RESTRAINT ALGORITHM FOR ASSIGNING FLOW TO A TRANSPORT NETWORK
Walter W. Mosher, Jr. 41

A DIRECT APPROACH TO TRAFFIC ASSIGNMENT
Morton Schneider 71

Traffic Assignment Analysis and Evaluation

DAVID K. WITHEFORD, Pittsburgh Area Transportation Study

Traffic assignment by computer is one of the most useful tools presently available to the transportation planner. But like all tools, traffic assignment is only as good as the manpower behind it. To use it most advantageously requires understanding of the capabilities and limitations not only of the tool but also of the inputs and related assumptions that go hand in hand. Properly employed, the assignment is invaluable, and not just in producing volumes on networks.

Assignment analysis should be designed and carried out with the following purposes in mind: (a) establishing validity of assignment results; (b) systematically producing workable data for evaluation (including economic evaluations, further general planning, design volumes); (c) permitting evaluation of internal system performance (identifying good and weak points in system, spotlighting deficiencies, etc.); (d) permitting comparative evaluations with other outputs to aid in planning and design toward the "best" system; and (e) permitting evaluation and interpretation of results for use by highway designers. These points are discussed both in general and with particular reference to PATS procedures. Emphasis is also given to the application and interpretation of results within the study for its own benefit and also for the benefit of participating agencies.

•TRAFFIC assignment is a process of allocating trips to a system to ascertain probable loads on part or all of the system. Although this can be accomplished in many ways, in this paper assignment refers to the process as done by electronic computer. Even though there are a variety of computer assignment techniques and disagreements concerning the most suitable approach, there is general similarity in both the end products and the end uses. Further, this paper is not concerned with trip forecasting or trip distribution techniques, which also exist in some variety and which frequently are associated directly with the computer assignment of trips to a system.

What is discussed here are the applications of computer outputs, their values and limitations, and the consequent responsibilities which thereby appear to devolve on the agency making the assignments. These comments are centered principally on the experiences and procedures reported elsewhere (1, p. 159). However, they are probably applicable to other groups similarly engaged, despite the different trip distribution or assignment techniques that may be employed. It is hoped, therefore, that this paper will be mutually useful toward the exchange of information and ideas between those involved in making assignments and also useful toward increasing understanding of computer assignment among other users of the output data.

CONTRIBUTION OF TRAFFIC ASSIGNMENT

The principal benefits of traffic assignment (including, as needed, the preceding steps of trip forecast and distribution) are its speed and low cost as a means of assessing traffic volumes. This may be done in terms of one ramp, one route, or an entire metropolitan area. It permits the highway designer to test many alternate schemes for an interchange or route. It permits the planner to design transportation systems using something more substantial than intuition.

Paper sponsored by Committee on Origin and Destination.

Assignment procedures can be and obviously are varied according to the specific need at hand. But a general characteristic of all assignment techniques is that they provide, as an output, volumes on all parts of the system being tested. Variations in techniques are designed principally to give additional related output data or to include differing degrees of complexity in loading trips onto the network. Assignment philosophy may accept "minimum path" loadings or diversion curves, may or may not employ capacity restraints. Whatever the case, output data could include "trees," could show the origins of trips on selected links, could include travel costs, etc. However, these matters are generally subsidiary to the desired end of network loadings. These network loadings are needed by two types of users: the transportation system planner and the route designer (either highway or transit). The nature of application by the two types of users differs, and this is perhaps why problems sometimes arise. Although traffic assignment by computer developed as a tool for both the system planner and the designer, it has characteristics that are more appropriate at the planning level than at the design level.

TRAFFIC ASSIGNMENT AND THE TRANSPORTATION SYSTEM PLANNER

Assignment Values

The value of traffic assignment to the transportation planner is sometimes misconstrued. It does not take the place of planning; it merely enables the planner to determine the salient good and bad characteristics of various preconceived plans. The planner is concerned with providing a total transportation system that maximizes benefits and minimizes costs, according to whatever criteria might have been established. The traffic assignment technique that will be most useful, therefore, is that which will permit the planner to determine how well his goals have been achieved by a given plan, to determine the plan's shortcomings, and to provide information that may guide him toward a better plan. But even then, this is just one part of the planning process. The performance of a plan as it can be measured by assignment does not necessarily reflect its performance in meeting other criteria based on design standards or, more importantly, on other community objectives, such as the preservation of residential neighborhoods, the economic health of a business district, or encouragement of industrial or recreational development. The merits and limitations of a transportation plan reach beyond the aspects that are measurable through traffic assignment. These considerations, however, do not detract from the value of assignments as a tool in the development of a sound plan.

Useful Outputs from Traffic Assignment

The basic output from assignment is a listing of all the segments or "links" in a network with the traffic volumes that have been assigned individually to them. Such an output, which can be arrived at through a variety of techniques, is almost always supplemented with additional data. These might pertain to summaries useful as validity checks on the "run," to network usage from selected zones, to the source and/or destination of trips using certain selected links, or to other special subroutines of assignments, such as travel cost data or overflows from expressways to arterials.

A typical output from the assignment program in use at the Pittsburgh Area Transportation Study (PATS) contains the following data: on-line printouts from computer storage, principally various summary data; printouts obtained from off-line tapes of link loads and selected characteristics of zonal interchanges; and two separate decks of punch cards, totaling about 3,500 cards, containing data by zone and by link. This total volume of output, requiring little more than brief-case space, is sufficient to generate up to 50 man-days of work in analysis and presentation procedures (2).

The test of one network generally will involve the following work: a "free" assignment, more or less a trip desire line map fitted to the network regardless of link capacity; a "restrained" assignment, which forces a pattern of loading reflecting the capacity limitations of links in the network; one or more "trees," which indicate the paths followed to all other zones by trips from selected origins; and a conversion of the assigned travel characteristics of each link to the costs of travel. Of course, the design of the outputs is influenced primarily by the type of analysis to follow and, secondarily, by the desire to facilitate processing on EAM equipment.

Analysis Procedures

For discussion purposes, analysis procedures may be put into three categories: validity checks, manual and graphic procedures, and machine analysis of punch card outputs. The first need is to confirm the validity of the computer outputs. If some input or operating error, not apparent during the assignment run itself, caused an error in the results, then it needs to be found before other analysis proceeds. The reasonableness of results is initially established largely from over-all summaries. These can be judged against summaries of vehicle-miles and vehicle-hours of travel, travel costs, screenline and cordon crossings, and various other measures from previous comparable assignments. Because the PATS trip distribution and assignment are coupled together, a plot of the trips distributed to and from each zone is a performance check generally made of the distribution technique. Accounting machine checks are made at the same time, to verify card totals against printout totals and to insure that no omissions or errors have occurred in the tape-to-card conversion. This series of initial checks can be completed within a day after the computer work is accomplished.

Following these operations, the task of posting link volumes is begun. The arterial network and freeway system are depicted on a series of maps covering the 420-sq mi study area at a scale of 1 in. to 2,000 ft. These maps are assembled from the culture plates of standard USGS quadrangles on which have been drafted the basic network with complete numbering of nodes and interchanges (3). Volumes assigned to each link are posted on a reproducible set of the maps so that they can be distributed later as needed. Thus, complete detail of the network loading is available. Figure 1 is a section of a typical assignment map, showing the degree to which the network is coded. The arterial network comprises the major streets, in this case at fairly close spacing because of the travel density in the part of Pittsburgh illustrated. Nodes prefixed on this map with double zeros are where trips are loaded on the network from the surrounding origin zone. On the right is shown typical coding of a freeway interchange, where individual ramps are coded in detail.

The manual posting of volumes is clearly a time-consuming and tedious task. With over 1,500 arterial links and almost the same number of freeway and ramp links when a freeway system is superimposed, over ten man-days often are required to complete the entire posting for one network test. The use of data plotters and node numbering

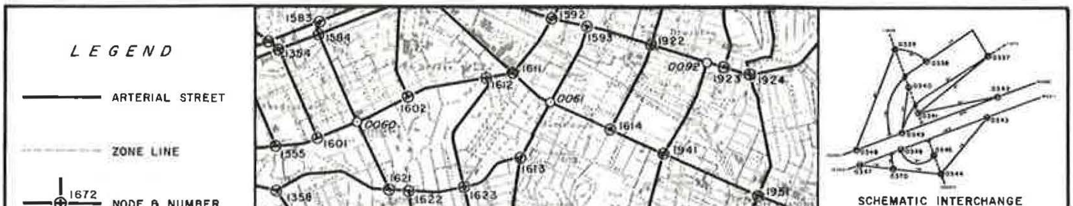


Figure 1. Typical network coding on base map (source: 1).

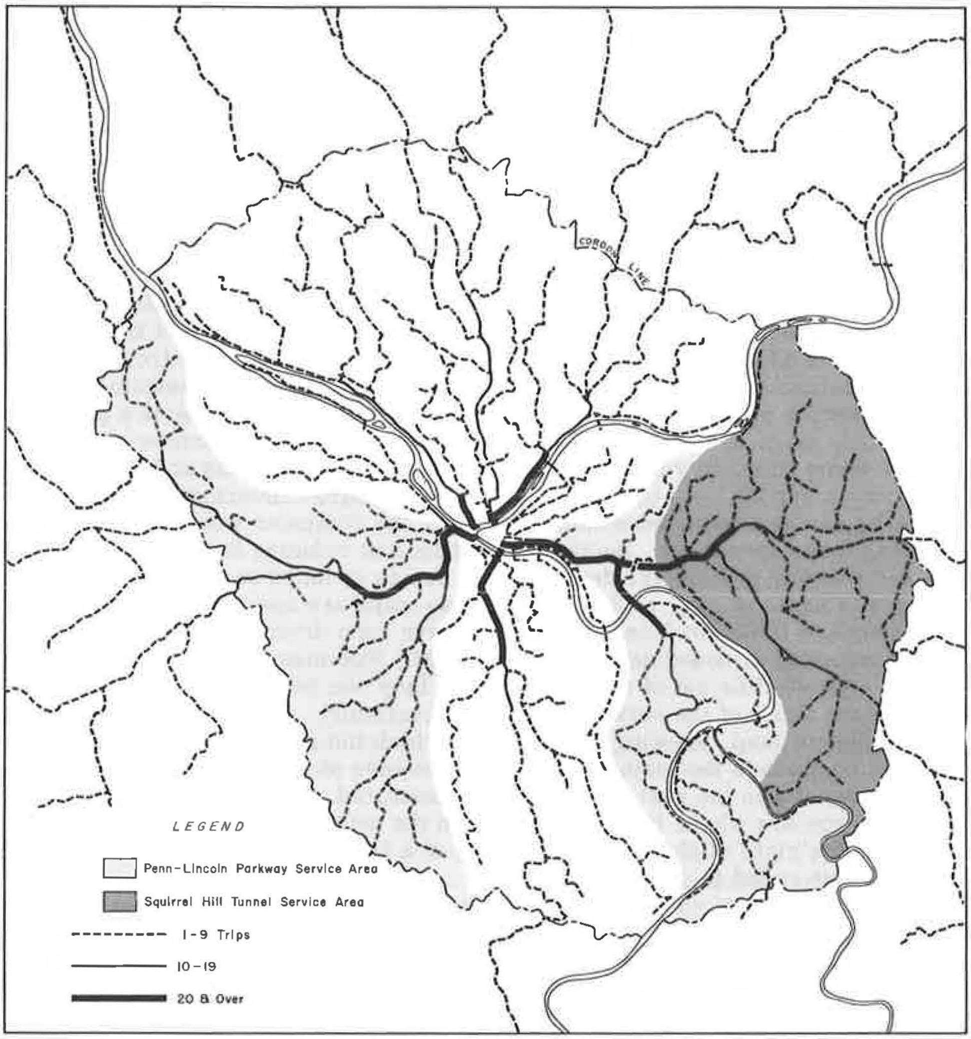

CORDON LINE
LEGEND
Penn-Lincoln Parkway Service Area
Squirrel Hill Tunnel Service Area
1-9 Trips
10-19
20 & Over

Figure 2. Minimum path tree for Golden Triangle to all other zones (1958 network).

based on a coordinate system is being developed by some studies to accelerate the process. But even so, manual work will still probably be necessary to show desirable detail at the scale of interchange layouts.

The value of the maps arises primarily from the fact that a permanent, detailed record of each assignment is available, not only for analysis of small areas or for taking off information used to make other study area evaluations, but also for comparisons with assignments made earlier or that may follow subsequently.

The maps themselves are used in a variety of ways for further analysis. For example, screenline checks can be summarized readily or examined by area. Volumes on links crossing the outer cordon line in different sectors of the study area can be taken off readily. The raw information is available to construct traffic flow maps for the benefit of technical or advisory committees or for public presentation. In such cases, volumes may be shown for the freeway system only or on both arterials and freeways by color coding and band widths. Such maps can be prepared best from the detailed assignment maps.

The volumes posted for planning purposes are generally those from both "free" and "restrained" assignments. Free assignments, as used in Pittsburgh, assign all trips via the minimum time-paths between zones based on freely moving travel speeds. Some links develop loads far beyond their capacity, and others develop no loads at all. Such results are of particular value in planning because they show whether the preconceived network tends to satisfy the travel desires of present or future users. Analysis of free network volumes after posting presents a good opportunity to evaluate a network, because both underloaded or overloaded routes indicate network deficiencies that should be corrected.

Equally valuable is the restrained assignment. In this type of assignment zonal transfers still follow minimum paths, but path times are increased as the ratio of assigned volume to link capacity increases in the assignment of trips. Original minimum paths increase in time until another path may become faster, the net effect being more evenly distributed travel and more realistic simulation of traffic flows. What may have appeared as a highly loaded route in a free assignment may appear as a group of less heavily loaded routes. The restrained assignment, which recognizes the realistic capacity limitations of network segments, will show where freeways are relieving arterial capacity problems as well as those areas where congestion may be expected as a result of inadequate design. Because both types of assignments are useful in system planning, all network tests at PATS include an assignment with each technique.

Another graphic version of output analysis can be obtained from "trees." Here the paths followed by trips from selected origins to all destination zones are plotted. The principal objectives in tree plotting are either to check the reasonableness of network coding or to demonstrate the part played by a freeway system in serving the origin zone. Posting minimum path trees tends to point up travel time deficiencies in network coding which might otherwise pass unnoticed. Devious routes that appear as minimum paths, or heavily loaded paths, may suggest either error in or unreasonable establishment of link travel times.

However, of greater interest is the use of trees in depicting the role of a system. For example, Figure 2 shows a somewhat abstracted version of a tree. In this case, an east-west freeway serving the Golden Triangle (GT) clearly carries the burden of GT trips. Shaded in is the "watershed" of this one freeway as a distributor of GT trips. By itself, the illustration is not overly significant; added to other assignment results, however, it provides further insight into system performance.

An illustration of zone-to-zone movements using a selected link, or of the origin zones contributing trips to a link of special interest, is another useful form of presentation for analysis purposes. But of more value than graphics is the machine-accounting evaluation of assignment results.

The link output cards are the source of most tabulations prepared from assignment data. They are summarized and processed to produce results at several levels of interest. Analyses by ring and sector (the largest units of the study area), by district, by zone, and by link, all play a part in the total evaluation of assignment results. Starting with the link card data, more than thirty tabulations are often developed by the PATS machine room.

Study area values are selected first. These are part of the early checks and summaries of network performance. Of particular value here are the travel costs incurred on the network—a function of speeds on each link after the capacity restraint is applied (4). Although over-all volume-capacity relationships also are compiled, the principal value is in summarizing travel costs for later use in benefit-cost analysis.

Investigation into performance by ring, sector, and district follows similar lines. There are 54 internal districts in the PATS study area, and in these it is possible to compare the amounts of travel on arterials and freeways with the total capacity provided by each type of route. Travel costs can be studied to assess where the system is performing efficiently or inefficiently. For more detailed inspection, the 226 zones can be used as the base point.

An example of evaluations at the zone level is shown in Figure 3, which illustrates the degree of service provided by freeways. One of the assignment outputs on punch cards can be used to determine the percent of total trips from each zone that used freeways. Grouping of zones into three levels of freeway trip percentages (0 to 19, 20 to

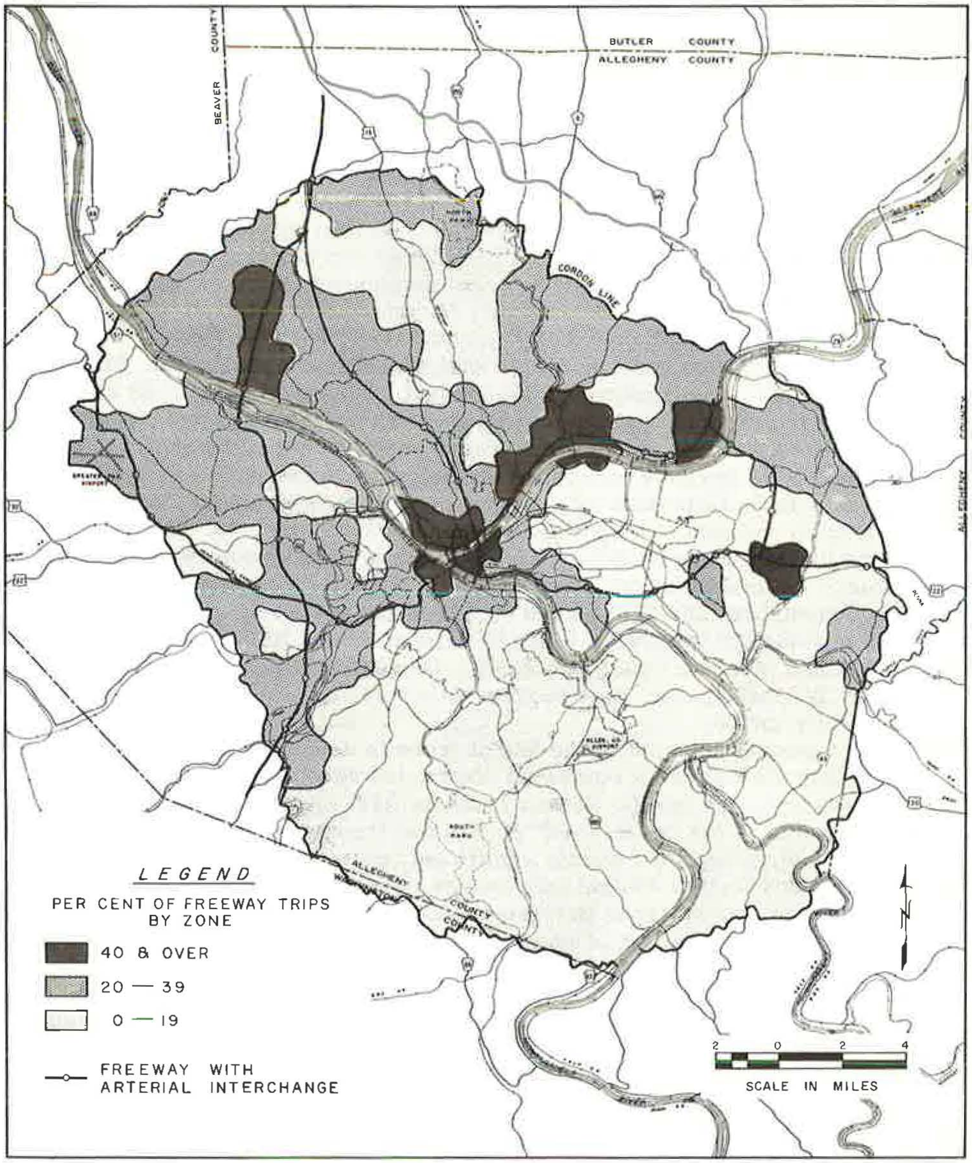


Figure 3. Map showing degree of freeway service (source: 2).

39, and 40 percent and over) provides a dramatic illustration, in this case, of service inequalities. One would not expect, for example, that a suburban area would make as extensive use of freeways as would a downtown commerical district. The illustration points out the obvious but provides a measure of the inequities, suggesting the degree of necessary additional facilities in various area. To a lesser extent, the effectiveness of proposed interchange spacing can be evaluated. A similar graphic summary, also compiled from punch card data, can show the relationship of travel to network capacity by zone. Some areas, because of low trip densities or relatively high amounts of freeway mileage may show unused capacity. Others may show a major deficiency in street capacity. Here, again, the map points out the obvious but, at the same time, shows both the degrees of deficiency and the generalized locations. This type of information, coupled with the data on assignment maps, forms a measure of what may be needed to bring the transportation system to an adequate level of service.

Lastly, the punch card output data can be aggregated by individual links to assess performance. Average speeds can be computed; the amounts and costs of travel on different classes of arterials can be assessed, and the underloaded and overloaded proportions of the network can be derived. Figure 4 shows a cumulative curve plot of arterial street mileage against volume-capacity ratios resulting from three different assignments. A ratio of 1.0 represents a balance between assigned volume and link capacity. Obviously, extremes of underloading or overloading are equally inefficient even if underloading represents a high quality of service. The three curves represent conditions on arterials in 1958, conditions in 1980 without highway improvements, and conditions in 1980 with PATS recommended plan. In all cases,'there are a few links with volumes more than double their capacity. This result is a reflection of computer mechanics, particularly on links connecting with loading nodes, and is an example of one of the difficulties that can arise from evaluating assignment data at the individual link level. The indicated overloads do not disturb the system planner who is aware of their cause—the necessity of loading trips at a zone centroid rather than distributing them over local streets to and from their actual terminals.

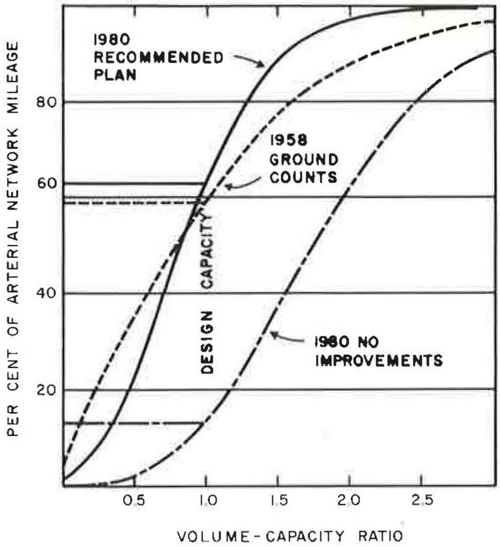


Figure 4. Cumulative percentage curve of arterial network by volume-capacity ratio (source: 1).

The situation just described points out the paradox in the use of traffic assignments for transportation planning. The assignment produces a wealth of details, some of which could be termed minutiae; for example, the cost of vehicle operations on a link $^1/_{10}$ mi long in the business district. But the planner cannot and must not accept this degree of detail. Instead, after measuring performance with an apparently fine degree of precision, he must apply his evaluations much more broadly in solving a metropolitan area transportation problem. If the assignment shows a need for improvement in a specific area, the planner must then apply judgment as to whether such improvement should be provided by tightening freeway spacing, by providing more freeway capacity on existing routes, by adding another freeway, by improving arterial access to freeways, or simply by adding arterial capacity. Individually balancing capacity and volumes on overloaded links is probably the least likely solution even though it might appear appropriate from examination of assignment link loads.

The objective behind the analysis of minutiae is to sift out the travel-related advantages and disadvantages of a transportation system for a given level of travel movement. The planner may want to develop a shift of through traffic away from the downtown area and may test various proposals accordingly. He may find that too extensive a freeway system is producing too small a return in travel cost reduction and is leaving existing arterial capacity unused. He may find an imbalance in freeway loadings because freeway spacing does not fit the underlying trip desires. The assignment provides this type of information, enabling the planner to improve his designs and to work toward the solution that best satisfies his planning criteria.

What can be accomplished by assignment is naturally dependent on the computer programs and the design of the output data. The more supporting detail the assignment can offer, the more soundly and quickly can a plan be achieved and the less intuitive and subjective will be its development and evaluation. The traffic assignment, because of its low cost and speed, makes possible the testing of many alternate schemes and provides results in a degree of detail that permits sound evaluations. It does not provide all the answers, even on just the traffic-related criteria of system planning; it does provide direction, however, to the planning effort. It is when assignments as used by the planner are modified or adapted to the more specific ends of the highway designer that the technique calls for even more discretion in its application.

TRAFFIC ASSIGNMENT FOR THE HIGHWAY DESIGNER

Assignment Values

The principal utility of traffic assignment to the highway designer appears to be in obtaining design traffic estimates on the facility with which he is immediately concerned. Traffic assignment tests the facility as a part of a system, and thus represents a tremendous technical gain over past procedures for estimating traffic on a single route. The assignment works with trip inputs for a whole area and a network of highways to provide, in effect, a working model of the route under consideration. A given route can be tested at various stages: for example, as partly built under stage construction or as a complete facility at different periods when it may be subject to the influence of other proposed projects. Traffic volumes for each situation can be predicted.

Assignment can produce traffic volumes with design variations for the individual route. Alternate alignments may be tested, or interchanges relocated or dropped, to determine the shifts in traffic flow on both the route and adjacent facilities; or the assignment can produce, in a difficult weaving area, for example, data on the origins, destinations, and volumes of weaving movements.

The outputs of greatest value to the designer are, therefore, the individual link loads, whether these give truck percentages and design-hour volume directly or require further manipulation. Though the designer does not need to know all about assignments or the values of different outputs to the system planner, he does need to know the qualifications that should be applied to the posted numbers appearing on an assignment map. These qualifications, which arise in part out of the design of assignments for planning purposes, must be made clear.

Some Dangers in Using Assignments for Design

Assignment outputs developed for use in system planning cannot be applied to design problems without discrimination. Assignment techniques and network design developed principally for regional transportation studies are concerned with planning on an area-wide scale. The degree of detail in network design, trip interchange, and trip assignment is thereby conditioned. The sensitivity of the computer technique employed in PATS and other transportation studies is more than sufficient for area planning but, perhaps, is insufficient for the degree of detail desired by a designer.

The matter of scale in network design and in its representation to the computer is the source of some difficulties. In PATS' case, link lengths are coded to tenths of a mile; travel speeds are represented in 5-mph increments; and capacity in a series of increments is calculated from standardized intersections. Such characteristics are not refined sufficiently to use in determining individual intersection performance.

Added to this are the gross techniques (from a designer's viewpoint) of loading trips from an area of major trip generation at one intersection within the area. Many thousands of trips to and from a zone, consequently, may appear on just one link within the zone or on a series of links leading to a ramp being studied. A prime example of this occurred in Pittsburgh when one ramp developed volumes exceeding 20,000 vehicles per day, whereas the adjacent ramp showed almost no traffic. This was a result of loading node placement. Obviously, trips, in reality, are generated on a plane surface rather than at a point; and ramp volumes should be distributed accordingly. But the scale of network design and the mechanics of assignment for system planning really do not need to reflect this degree of detail when the concern is to develop a transportation system for an area of several hundred square miles.

What is less evident, and perhaps of greater concern to the system planner when his assignment results are used for design, are qualifications that need to be applied because of programing or other techniques. Here, again, the techniques—even though providing detail—are designed for broader analysis than that of the designer. Everything leading up to the individual link load in the computer output is designed for regional rather than local use. This is true of forecasts, mode splits, trip inputs, network design, trip distribution techniques, and the assignment itself. Developers of trip distribution models and assignment programs are far from agreement on what constitutes the best approach to traffic forecasting. It is not unreasonable, as a result, that the transportation planner should shrink from standing pat on an assigned volume of 4,256 vehicles on ramp X of a downtown interchange.

The significance of predicting turns or delays at arterial intersections in 1980 is even more questionable. For example, in PATS' assignments, the intrazonal trips do not even appear on the network, and all trips between adjacent zones cross the common boundary on one link only. In either case, this is not a simulation of movement in reality. How then can turning movements be predicted for design purposes? This gets to the fundamental questions of how much accuracy of detail is needed and how it is to be achieved.

It would be unseemly for transportation planners to suggest the level of detail needed by the designers. But, as pointed out in a recent report (5, pp. 45-56) the reporting of design-hour volumes as a neatly packaged, arithmetically balanced presentation on design drawings seems neither necessary nor appropriate. Design-hours on the ramps and roadways of a freeway do not necessarily occur in the same hour of an evening or morning peak. In some cases, they may not coincide with any of the traditional peak periods. Perhaps the designer should work more within ranges of volume than with specific values. In PATS' experience, at least, it appears that designers expect too much detail and that planners (as assignment makers) obligingly provide it.

Assignments can, and some do, provide peak-hour turning movements at arterial intersections, and perhaps it is best that they do so. Despite the limitations that exist within any assignment technique, traffic assignment does provide the most sound basis on which to establish design-hour volumes. However, computer-produced numbers in such detail are almost certainly spurious if taken literally, even though they may appear realistic. Computer outputs, thus, are of most value in design when used comparatively, with judgment based on knowledge of the forecasting and other limitations inherent in the processes.

In effect, the designer should examine closely the details available and then form his conclusions from a broader perspective in the same sense as the system planner must. To insure this approach, there is a need for interpretive judgment somewhere between the computer laboratory and the drafting room. The agency generating assignments would appear most responsible for providing such advice.

Responsibilities for Assignment Interpretation

Without the intermediate step of interpretation, there are too many opportunities for misunderstanding, misguidance, or misuse of assignment data by the many agencies that need this type of information. The problem may not arise when the assignment and its application to design is performed within, and is wholly restricted to, one organization. But in most cases, the assignment makers and the design users are organizationally removed from one another. At PATS, it has become clear that transmittal of assignment data to other agencies must incorporate some means of conveying judgments related to application of the data. Several procedures could be followed.

First, the assignment-making agency can transmit data to designers with accompanying explanations and qualifications on how the data can be used. However, such a procedure generates much report writing and places the burden of translating numbers plus qualifications into design-hour volumes on the designer—who is not likely to be interested in this additional work. In addition, the assignment maker should feel responsible to follow up and ascertain whether the data have been properly interpreted.

A second possibility would be for the assignment makers to provide analysis and interpretation for all data requests. Considering the variety of these requests and their sources, this could be a really burdensome task. It is possibly the best solution in terms of getting the most meaningful interpretation of assignment results, but the time and manpower involved would add considerably to budgets. A drawback might be that the assignment agency may not be properly qualified to assess other important aspects of the requesting agency's problem, such as the influence of right-of-way costs in one design location versus another, or the planner's need for balancing street capacity requirements against requirements for public open space or other activities.

A third possibility is for the agency making assignments to designate a staff member to work closely with the requesting agency as long as assignment data are being used. Such a procedure certainly would be effective in providing good guidance and understanding on both sides. But this takes a lot of staff time, and qualified staff is generally in short supply, at least on most transportation studies. In a similar but reversed manner, the requesting agency could assign a designer to work with the assignment makers. All requests for traffic estimates for design purposes could be channeled through this individual, who would be able to assess the designer's needs and provide information accordingly.

A fourth alternative, which is being practiced in the Pittsburgh area with some success, is to bring in an intermediary person who is familiar with the problems of both the assignment makers and the designers. Using PATS' assignment data as the starting point, a local traffic consultant is providing the Pennsylvania Department of Highways with traffic estimates for design purposes. Introducing a third party obviously means added direct cost, but it also means that both PATS and design personnel are enabled to concentrate on their individually specialized tasks. The consultant, through experience with assignment techniques, is familiar with the capabilities and limitations of assignment data. He has a knowledge of local conditions and problems which are beyond the scope of a regional study and is familiar, also, with the geometric and structural considerations of design. A principal benefit is minimization of the duplicated work that might occur if the study staff had to explore local conditions which were already known elsewhere. Furthermore, because a local highway engineer was quoted as "viewing PATS with a mixture of awe and suspicion," it appears to be beneficial for the Department of Highways to obtain an outside view on the meaning of assignments. The consultant, of course, eventually may find himself viewed dubiously from both directions.

Regardless of the procedure followed (and each of the preceding is practiced at PATS according to the nature of the data request), all require initiative on the part of the agency making assignments. This group must provide information on the significance of output data, because it is generally the only group qualified to do so. The assignment makers should feel, also, a responsibility to follow through and ascertain that assignment outputs have been used properly. This is partly for their own self-protection but more importantly to prevent what, otherwise, might be costly mistakes. In addition to the Department of Highways and its consultants, organizations that have made use of PATS' assignment results include the City of Pittsburgh's City Planning Department, Traffic Engineering Bureau, and Urban Renewal Authority; the Allegheny County Redevelopment Authority; Pittsburgh Regional Planning Association; and consultants with clients ranging from the Federal government to local department stores. Most, if not all, of these users wish to make local or detailed application of assignment information. All need to be advised of the limitations on assigned link-load values for such purposes.

The use of assignment results in design applications must be regarded, by transportation planners at least, as a secondary benefit to be obtained from what is primarily a tool for regional planning. Assignments can be devised to specify turning movements at individual intersections, but it must be remembered that the accuracy of such predictions is as much a function of forecasting and trip distribution techniques as it is of assignment technology. Because assignment results are sought out for design or other local applications, the assignment makers have a clear responsibility to assure proper evaluation and avoid misuse on the part of other organizations.

SUMMARY

The end uses of traffic assignment are, of course, directly related to the characteristics of computer programs. But regardless of the diversity that exists in the techniques, all assignment outputs have basic similarity in their production of individual link volumes after distributing trips over a network of highway or transit routes. This discussion has not been concerned with areas in which improvement in technique would be desirable, only with the uses that can be made of assignment output data, based on the experience of one transportation study.

It is felt that traffic assignment is most beneficial as a planning tool in the development of area-wide systems. Its employment and value in design (for example, in developing design volumes on interchange ramps) is a collateral benefit. Traffic assignments are undoubtedly the best means available today of reaching design-hour traffic estimates, but computer outputs cannot be applied without qualification and considerable study. Possibly, assignments for design should provide the maximum level of detail attainable, so that the designer can examine the details but ignore them as such, extracting the broader conclusions only in the same sense as the system planner must. The danger lies in the assumption of accuracy merely because results are detailed. Because the computer does produce detailed results quickly and cheaply, its virtues lie in permitting more testing of alternates, more freedom of choice, and the bringing to bear of more complex relationships, all of which require judgment in evaluation and in making subsequent decisions. These aspects, of course, represent the principal benefit of traffic assignment for design purposes. The agency making assignments for system planning purposes thus has a dual responsibility. First, in using assignments to derive a plan, the agency must apply the judgments needed for its own purposes and then must weigh the traffic advantages with the other criteria essential to system planning. Second, the assignment makers must be certain that the proper judgments are made by the many other agencies which expect to, and can, benefit from traffic assignment techniques.

REFERENCES

1. "Forecasts and Plans." Pittsburgh Area Transportation Study, Vol. 2 (1962).
2. "Traffic Assignment Analysis." PATS Research Letter, Vol. 4 (Jan.-Feb. 1962).
3. "The Assignment Map." PATS Research Letter, Vol. 2 (Jan. 1960).
4. Haikalis, G., and Hyman, J., "Economic Evaluation of Traffic Networks." HRB Bull. 306, 39-63 (1961).
5. "Inner Belt and Expressway System, Boston Metropolitan Area." Commonwealth of Massachusetts (1962).

A Comparative Description of a Capacity-Restrained Traffic Assignment

ROBERT B. SMOCK, Detroit Area Traffic Study, Wayne State University

• THIS PAPER specifically presents a step-by-step case history of a "capacity-restrained traffic assignment" and compares it to three other kinds of assignment that are in use today. It has also the more general purpose of starting to move the capacity-restraint method out of the research institutes and into the State highway departments and local planning offices where it will do the most good. Traffic assignment in general is the technique used to determine in advance how well a proposed highway system will work. Its components are (a) a table giving traffic volumes between traffic analysis zones in some future year, (b) a highway network proposed for the same year, and (c) a program for routing all the interzonal trips through the network and totaling the volume "assigned" to each link. Testing alternate proposed systems in this way, it is possible to produce one that will function without undue traffic congestion.

Doing away with traffic congestion requires a number of steps in addition to traffic assignment, including steps in the planning and design areas, as well as all those in the areas of finance, construction, maintenance, and operation. However, without reliable traffic assignments it is not possible to pursue any transportation plan with confidence that it will succeed.

Before the introduction of traffic assignment, highway planning tended to proceed on the basis of estimates of percentage increases in the loads to be carried by individual existing roads. This approach has been made rather obsolete by changes such as the growing recognition of individual highways as parts of highly interdependent systems, and by the large-scale additions of new facilities to existing systems—particularly freeways. The start of highway-system planning, at least within urban areas, might be dated from the introduction of the origin and destination (O-D) survey. This kind of research produces "trip tables" which indicate the interzonal trips people make independent of the routes they select. Once there is a trip table, there can be traffic assignment.

A typical situation today might include a town which has had an O-D survey and has projected its trip table to 1980. On the basis of the projection it is concluded that two or more freeways will be needed, and a total highway system is proposed for construction by 1980. At this point at least three important questions can be raised: (a) Does the proposed network provide sufficient highway capacity where it is needed, so as to minimize congestion? (b) Does it provide no more than the minimum amount of necessary highway capacity to be paid for from the public purse? (c) How is the network to be designed in terms of lancs and turning movements? These are questions that can be answered by traffic assignments.

This hypothetical town can consider four different types of assignment. It can trace a path through the network by hand for each interzonal connection, and assign the appropriate volumes to each link for each path. This "hand-trace assignment" could take a great deal of time, however, and if an electronic computer is available, its use is likely to be considered.

The major planning problem is posed by new facilities—especially freeways--and for some years when computers were new to this field, their contribution was limited to assigning parts of interzonal volumes to freeways alone. The computer "freeway assignment" was a giant stride forward, especially in large cities where hand tracing was almost prohibitively time consuming, but it left unanswered the important questions

Paper sponsored by Committee on Origin and Destination.

about the adjustment of the other parts of the network to the new facilities.

Only in recent years has it become possible to program a computer to trace paths through an entire major street system. The products of the simplest of such programs are called "desire assignments" because every interzonal volume is assigned to the best path that could be desired, regardless of the capacities of the streets. Such an assignment was made in Detroit for 1958, and it was found that more than 300,000 vehicles desired to use some individual links of the freeway network. This was an interesting statistic, but not very useful for planning purposes because it was not considered practical to try to provide that much capacity in a single facility.

The development of "capacity-restrained assignment" is a very recent achievement but its advantages are so obvious that future assignments may be expected to be of this type just as quickly as computers with their programs, and the necessary know-how, can become available to highway planners. Once capacity-restrained assignment is in general use, that hypothetical town is not likely to choose hand-trace, freeway, or desire assignments. Instead, there is likely to be general agreement on the following six capacity-restrained assignment steps as constituting good highway system planning from the assignment point of view (1). For the most part these steps are applicable and desirable regardless of the particular assignment technique adopted.

First, present-day volumes are assigned to the present-day street network, and the results compared to traffic counts. This serves to validate the trip table and the O-D survey sample from which it is derived.

Second, at least two alternative future land-use plans are developed, perhaps one assuming a continuation of present trends and the other assuming some reasonable planned improvements. Each land use plan will produce a different future trip table, and both of them are assigned to the present major-street network to provide a starting point for the planning of future networks.

Third, the future-trip table based on the first land use plan is assigned to several alternative proposed future networks until a good one can be demonstrated.

Fourth, alternative networks are tested with the trip table based on the second land use plan until a final proposed network can be selected.

Fifth, morning and afternoon peak-hour volumes are assigned to the selected network as an aid to its geometric design.

Finally, five-year step forecasts between the present year and the future planning year are assigned as an aid to the determination of priorities for construction. Also, five years pass fairly quickly and these five-year interval assignments allow a periodic checking of forecasts against the actual situation.

These six steps call for 15 or 20 traffic assignments, which means considerable work and cost. The time and money for such research and planning are not a large part of the total costs of well-planned highway construction, however, and are actually small when compared to the high costs of poorly-planned construction with its possible waste and continuing traffic congestion. This report is intended to illustrate how useful a capacity-restrained assignment can be to good highway planning, as well as to demonstrate how it works.

ASSIGNMENT BACKGROUND

The Program

The essence of capacity-restraint assignment is a program for an electronic computer. Computers have proved themselves capable of performing extremely intricate tasks with great speed and accuracy, and it is some times overlooked that they can do absolutely nothing except what they are instructed to do. A set of computer instructions is a "program," and the development of a program for the performance of something as intricate as traffic assignment is a long, difficult, and costly job.

The one whose workings are described in this paper was developed by the Detroit Area Traffic Study and the Computing Center at Wayne State University. The original objective was to program a high-speed computer to assign traffic to the major streets of Detroit, but two interim objectives have developed. One is to prepare the IBM 650 to assign traffic to a 240-intersection network, which is probably large enough to handle

TABLE 1

TYPICAL SPEEDS AND CAPACITIES[a] ON FOUR-LANE HIGHWAYS

Area	Speed (mph)			Capacity (thousands)		
	Arterial		Freeway	Arterial		Freeway
	Standard	Major		Standard	Major	
CBD	10	20	45	22	27	78
City	15	25	50	16	19	56
Suburbs	20	30	55	12	15	42

[a] 24-hr average weekday; practical, not ultimate.

all Michigan cities except Detroit. The other is to specify every step of the capacity-restrained assignment procedure in an operating manual that will allow the Michigan State Highway Department to perform such assignments on a routine basis.

At its present stage of development, this method includes the following steps. First, every link in the network to be tested is classified as to its type, the area in which it is located, and its pavement width. A distinction is made between three types: freeways, major arterials (such as divided highways), and standard arterials; and three areas: the central business district (CBD), the rest of the city, and the suburbs.

Next, a capacity can be given every link by looking it up in a table on the basis of its type, area, width, and kind of intersecting streets (2). A typical speed can be given each link on the basis of its type and area. Table 1 gives typical speeds and typical capacities for four-lane streets. These generalized measures should be replaced by more exact measures when they are available.

Then, the length of each link is measured by the computer on the basis of a coordinate-coding of each intersection. From its length and speed, the computer determines a travel time for each link, and this time is the "value" (V) of each link which the computer uses to determine paths through the network. With this information and a trip table, the actual assignment can begin.

The assignment proceeds in a series of "passes" over the network. On each pass the computer determines the shortest time path between every pair of intersections which has been identified as the origin and destination of a trip. On the first pass, travel times are determined on the basis of typical speeds and all volumes are assigned to their quickest paths, thus providing a "desire assignment." At the end of the first pass, the computer determines the ratio (R) between assigned volume and capacity for every link, and this quantity is used to compute a new travel time on each link according to the conditions of congestion pictured by the desire assignment. Where R is high, speed is reduced below typical speeds and travel time is increased; where R is low, speed is increased above typical speeds and travel time is reduced.

The formula according to which these increases and decreases are determined is the keystone of the capacity-restrained program.

$$V_i = e^{(R_i - 1)} V_o \qquad (1)$$

in which

V_i = travel time on a link for a given pass;
e = 2.71828;
R_i = ratio of averaged assigned volumes (from all preceding passes) to capacity; and
V_o = original (typical) travel time on link.

The formula was derived both by mathematical logic and by trial-and-error experimentation (3).

For the second pass these travel times reflecting congestion oh original best paths are used in the computation of a second set of paths between every pair of origin-destination intersections. Each interzonal volume is then divided evenly between the two paths that have been traced for that trip, and the averaged volumes are used to compute new values of R and another set of travel times. For the third pass the same procedure is followed, and such passes can be repeated, dividing interzonal volumes over more and more paths, until capacity-adjusted speeds, on the average, come to approximate typical speeds. The two measures of speed will converge as assigned volumes converge on capacities. This particular type of "convergence" will happen quickly when the network provides sufficient capacity exactly in the places where it is needed, and will never happen completely if the network does not contain sufficient capacity for the volumes assigned to it.

This approach was tested by using it to assign a 1950 trip table to the 1950 arterial streets of Flint, Mich., and comparing the assigned volumes to traffic counts (4). It also has been the approach followed in the assignment to a proposed 1980 network for Flint.

Volumes and Network

The network proposed in Flint for 1980 was selected not because it was suspected of being poor and in need of testing, but because it was known to be the well-planned product of lengthy research by the City of Flint, by the city-planning firm of Ladislas Segoe and Associates, and by the State Highway Department. Douglas Carroll helped conduct one of the early O-D surveys in Flint in 1950, and further research in 1960 had forecasted its trip table to 1980 (5). Being able to rely on network and volumes was a great help in this continued experimentation with the assignment program: it was expected to be a "good" network, as defined later.

A part of the network shown in Figure 1 is the major street system of Flint in 1950. When a "desire assignment" of 1980 traffic volumes was made to this network, about two-thirds of its links were loaded beyond their capacity. Because these overloaded links tended to be the high-capacity ones, it is obvious that no amount of capacity restraint and the tracing of alternate paths could bring about "convergence" with capacity for these volumes on this network. In fact, the volumes were so far beyond the capacity of the system that one-third of the links were loaded to more than twice their capacity, and one out of every ten links was assigned a volume more than five times its capacity.

Figure 1 also shows the additional facilities proposed for 1980 Flint, and Figure 2 shows the types of facilities proposed. The major addition is a system of three freeways: an east-west one running just south of the CBD, a north-south one along the western edge of the metropolitan area, and a branch of the north-south one passing through the city just to the east of the CBD and rejoining the other at the northwestern edge (top of map). The freeways along two sides of the CBD carry only "through" traffic, as all turning movements to and from the CBD are carried by service drives connecting with the freeways only at the outer corners of the downtown section.

Flint is astride a major north-south route through Michigan, and traditionally its two major arterials have been Dort Highway which runs north-to-south on the eastern edge of the city, and Saginaw Avenue which is shown in Figure 2 as branching off from Dort on the south to run through the CBD, and rejoining Dort on the north. Other major arterials reach into the four quarters of the area, except to the far northwest, and are joined together by a grid of standard arterial streets.

The detailed design of this network is not complete, but tentative widths were obtained for every part of it. All the freeways sections were treated as if they were of four-lane width, and many of the existing arterials streets were treated as if they had been considerably widened from their width as of today. This, then is the network expected to be a "good" one, and in general it was not disappointing. Of course, to a large extent, its satisfactory nature was a function of the way in which "good" was defined.

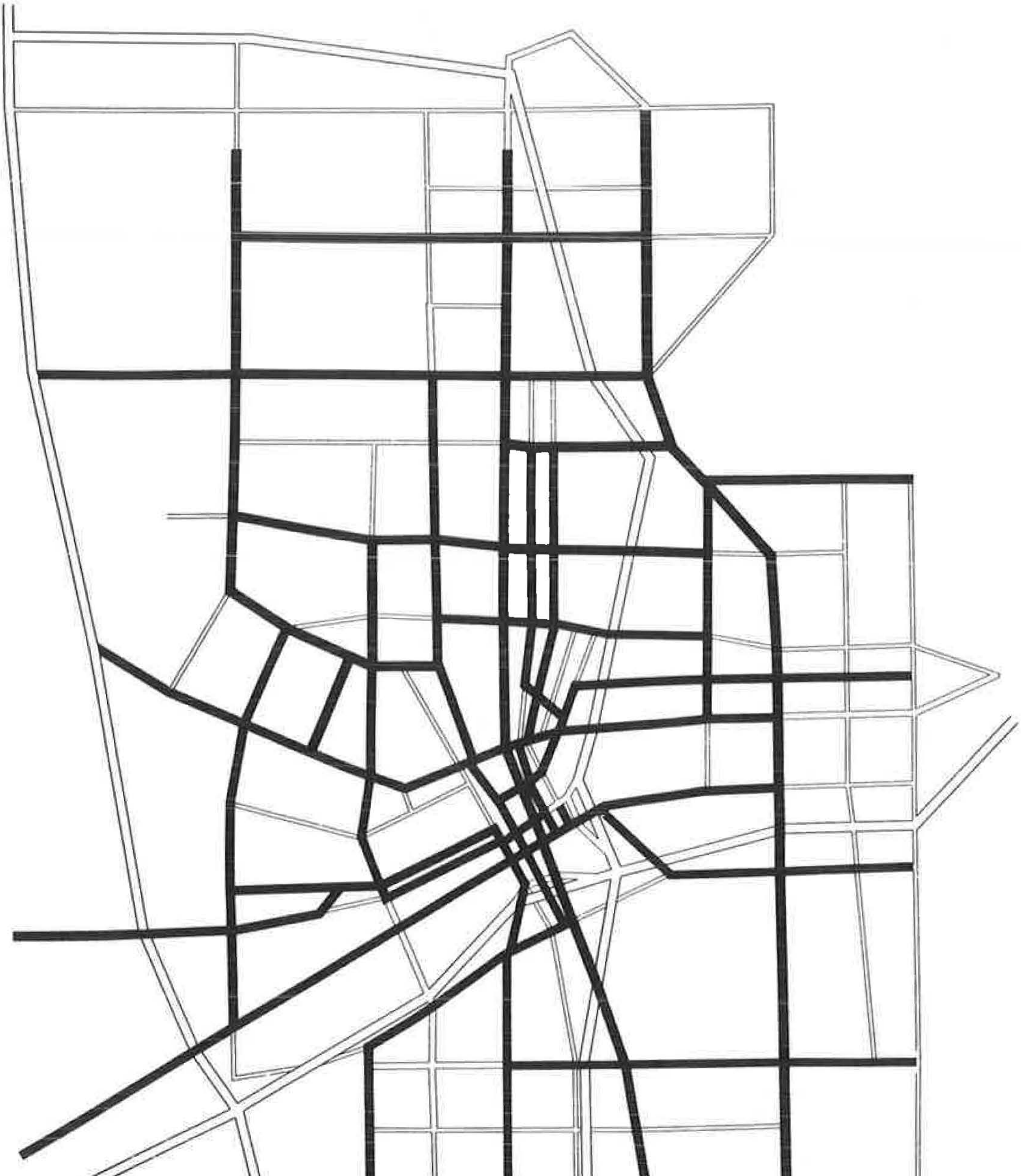


Figure 1. 1950 and 1980 networks.

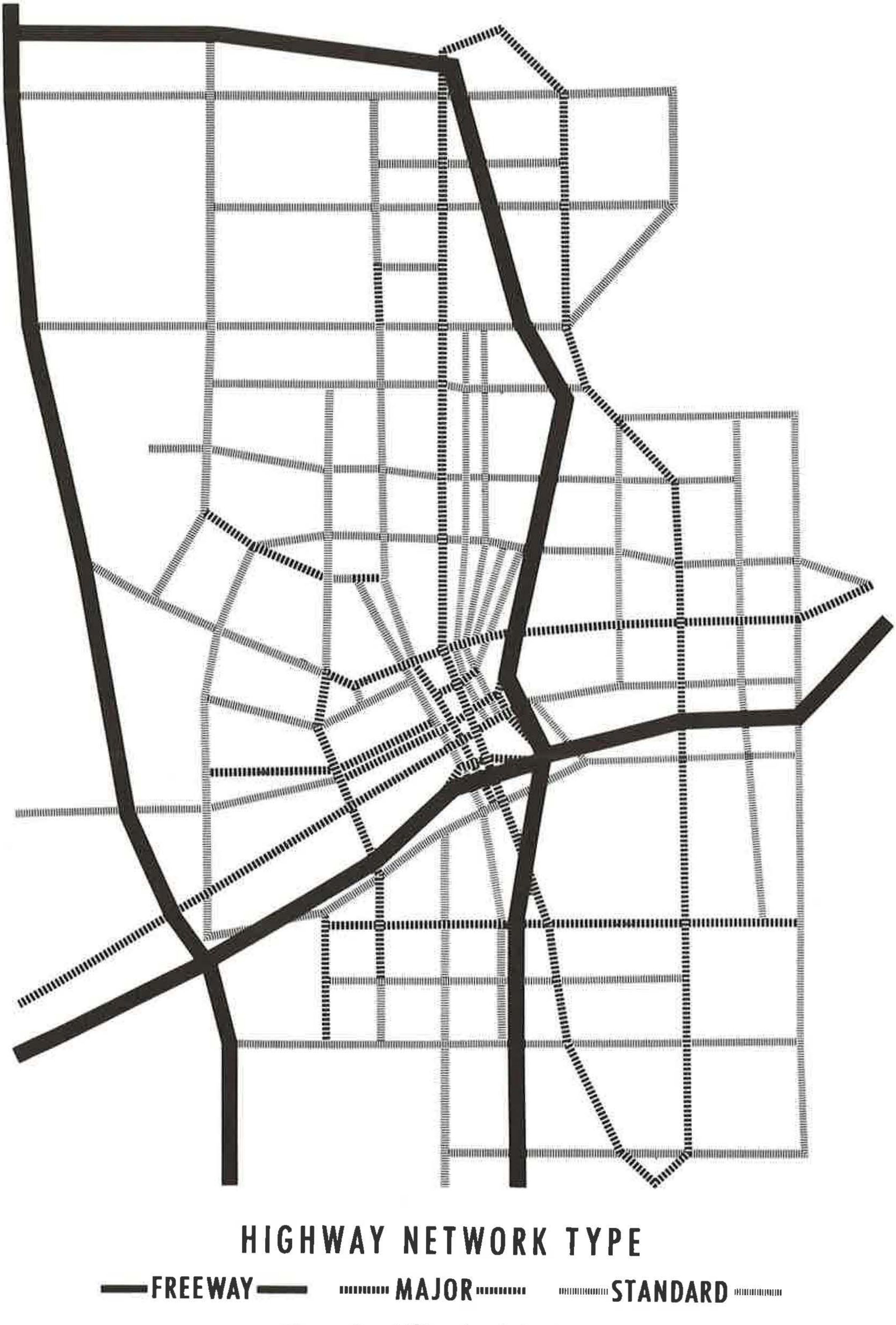


Figure 2. 1980 network by type.

A "Good" Network

A good highway system is one that provides sufficient capacity where it is needed to avoid undue congestion, without providing unneeded capacity. The idea of "sufficient capacity" has a relatively clear meaning, but the idea of capacity "where it is needed" is a little harder to define in practice. In American cities it is common for major corridors of vehicular movement to contain more vehicles than can be carried on a single facility, so that it is unavoidable that some vehicles will travel on paths through street systems that are different from the paths composed of the most direct facilities between their origins and destinations. On the other hand, it seems unlikely that more than two or three parallel facilities would be required for any such corridor in order to avoid forcing any trip to take an extremely indirect path. One exception to this is that the unusually long trip through a modern city should use a freeway, so long as the cost of the possible indirectness of the freeway path is offset by a savings in travel time.

These general considerations about what constitutes a good highway system can be stated in the form of a four-part operational definition of "good" in terms of the characteristics of a capacity-restrained traffic assignment. A good highway system allows the following:

1. "Convergence" of assigned volumes and capacity in three or four assignment passes, because two or three paths as alternatives to the "best" path should handle practically all trip volumes.
2. Total vehicle-miles of driving through the system which are not greatly in excess of the mileage required for all trips to take their "best" paths, otherwise capacity has not been provided where it is needed.
3. Total vehicle-hours of driving through the system which are not greatly in excess of the hours required for all trips to take their "best" paths at typical speeds, which is the same as saying that congestion-adjusted speeds should converge to be approximately the same as typical speeds in three or four assignment passes.
4. Freeways should tend to carry the longer rather than the shorter trips, because the costs of freeways are only justified by the benefits of removing "through" traffic from standard arterial streets.

The indefinite quality of phrases like "not greatly in excess of" is not the handicap it appears to be. The usual purpose of traffic assignment is to compare alternate networks, although that is not the case for the Flint assignment discussed here. When alternate networks are tested, "excess" vehicle-miles or vehicle-hours can be simply defined as the additional miles or hours required to travel through one network as compared to another. Each of these general principles can be illustrated in terms of the actual workings of the Flint assignment.

THE ASSIGNMENT

Convergence

Figure 3 shows the assigned volume on each link of the Flint 1980 network, expressed as a percentage of the capacity of that link. This is a picture of the situation at the conclusion of the first pass or "desire" assignment. At this point, all volumes have been assigned to the path between their particular origin and destination which can be traveled in the shortest amount of time at typical speeds. This kind of path is said to be the "best" path. The lightest links on this map were included in no best paths and therefore have zero volumes.

In connection with the next darker class of links, it should be noted that no traffic assignment assigns to a network every vehicular trip reported during an O-D survey. Those trips having both their origin and destination within a single traffic analysis zone cannot be assigned. Such intrazonal trips have to be considered "local traffic" and be kept off the arterial street system. This is not entirely realistic because local trips can make up a large proportion of the traffic on major streets in the CBD, and perhaps 20 percent of the traffic on other arterial streets in the city. The next darker links in

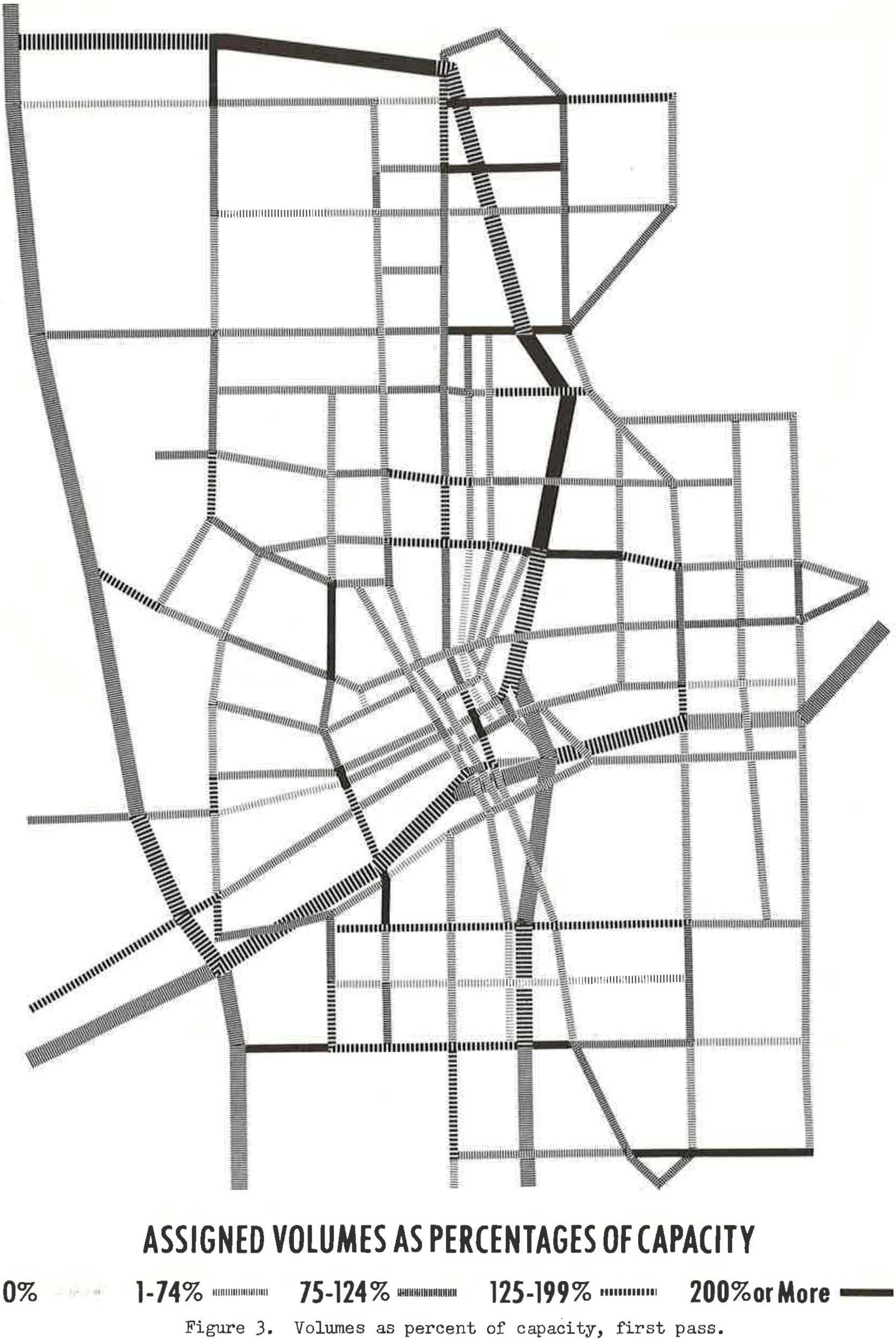


Figure 3. Volumes as percent of capacity, first pass.

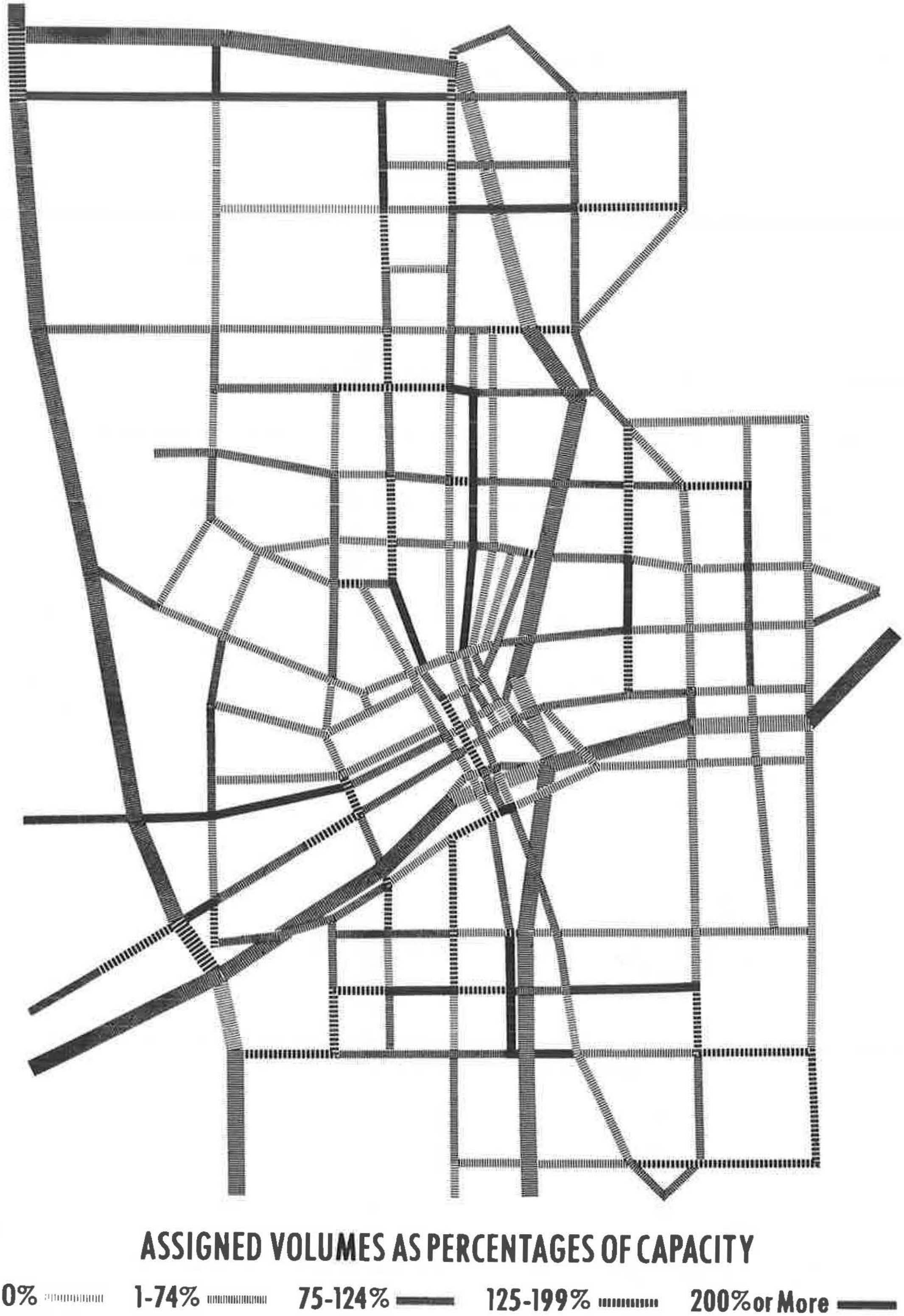


Figure 4. Volumes as percent of capacity, second pass.

Figure 3 have been assigned a volume at least 25 percent below capacity, and therefore a volume likely to move without congestion even when local traffic is taken into consideration.

The next class of links have been assigned a volume within 25 percent of their estimated capacity, and those in the class after that are definitely congested, having been assigned a volume between 125 and 199 percent of capacity. Links in black are heavily congested, having been assigned more than twice their capacity.

On this first pass, every link but one in the 25-link freeway system is loaded to capacity, and almost one-half the freeway links are congested, two of them heavily. Of the other heavily congested links in the system, about three-quarters are funneling traffic onto or off the freeways. A fact to be especially noted, however, and one that illustrates clearly the danger in using desire-assigned volumes for design purposes, is that the majority of these congested links are adjacent and parallel to underassigned links. This means that, were this network in actual operation, its links would not carry volumes like those assigned on this first pass. Furthermore, it might operate much more effectively than the number of congested links suggest because many vehicles could avoid congestion by taking paths along the underassigned adjacent and parallel links, which probably would be only slightly less direct than their "best" paths.

If the first pass is called a "desire assignment," then the second pass might be called a "rush-hour assignment" in that its paths have been traced through a network in which popular links have had their speeds slowed to the minimum. It can be considered the "rush-hour pass" only in the limited sense that, for the average trip, the path traced at this point is likely to be the one that deviates the farthest from the "best" path, and this is analogous to the real situation in which drivers are most likely during rush hours to seek the suggested alternatives to their own choices of best paths.

During the second pass, volumes are divided evenly between the two paths that have been traced for each trip. The averaged volumes on each link are expressed as percentages of capacity, with results in Flint that are shown in Figure 4. Here the distribution of heavily congested links is considerably different from that at the conclusion of the first pass. Only two links originally assigned more than twice their capacity are still in that class after the second pass. These two links are parts of standard arterial streets in the suburbs carrying traffic to and from freeway ramps; they were assigned more than four times their capacity on the first pass. Therefore, they are left with at least a double load from their averaged assigned volumes even if they were zero-volume links on the second pass. In addition, 21 links which were not heavily congested after the first pass are heavily congested after the second, although they have been assigned only one-half the volume of any particular trip.

On the whole, there are more links assigned volumes near their capacity after the second pass than after the first, so that it can be said that convergence has begun. On the other hand, a few links that were congested after the first pass are underloaded after the second, and the opposite is also true. The third and subsequent passes are "balancing assignments," then, and convergence should proceed rapidly. Figure 5 shows the situation at the conclusion of the third pass. Now there are no zero-volume links, and the number of heavily congested links has been reduced by two-thirds.

It is after the third pass, however, that there is the first hint of a few trouble spots in the network, in the form of congested links that are not paralleled by underassigned links. After the fourth pass, a considerable amount of convergence has taken place (Fig. 6). A number of links continue to be congested, however, and four links are loaded to more than twice their capacity at this point. Again, they are links providing access to the freeways, specifically to the branch north-south freeway passing through the city.

At this point there was an inclination to consider the assignment complete, and to view the remaining degree of lack of convergence as an indication of points in the proposed network that were in need of change. However, a fifth pass was made to check this conclusion, with the results shown in Figure 7. There continue to be four heavily congested links, and they continue to be links providing access to the branch north-south freeway.

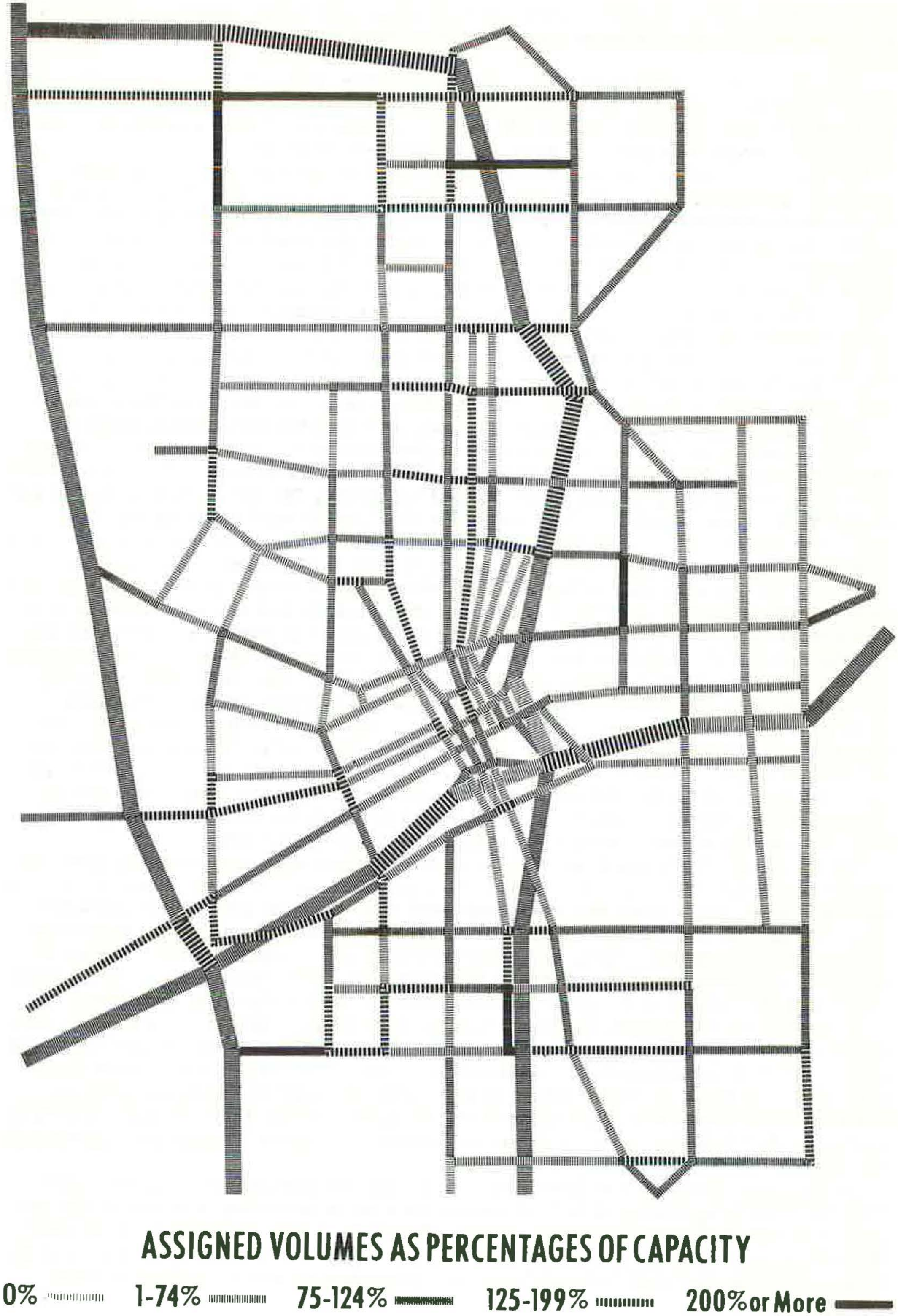


Figure 5. Volumes as percent of capacity, third pass.

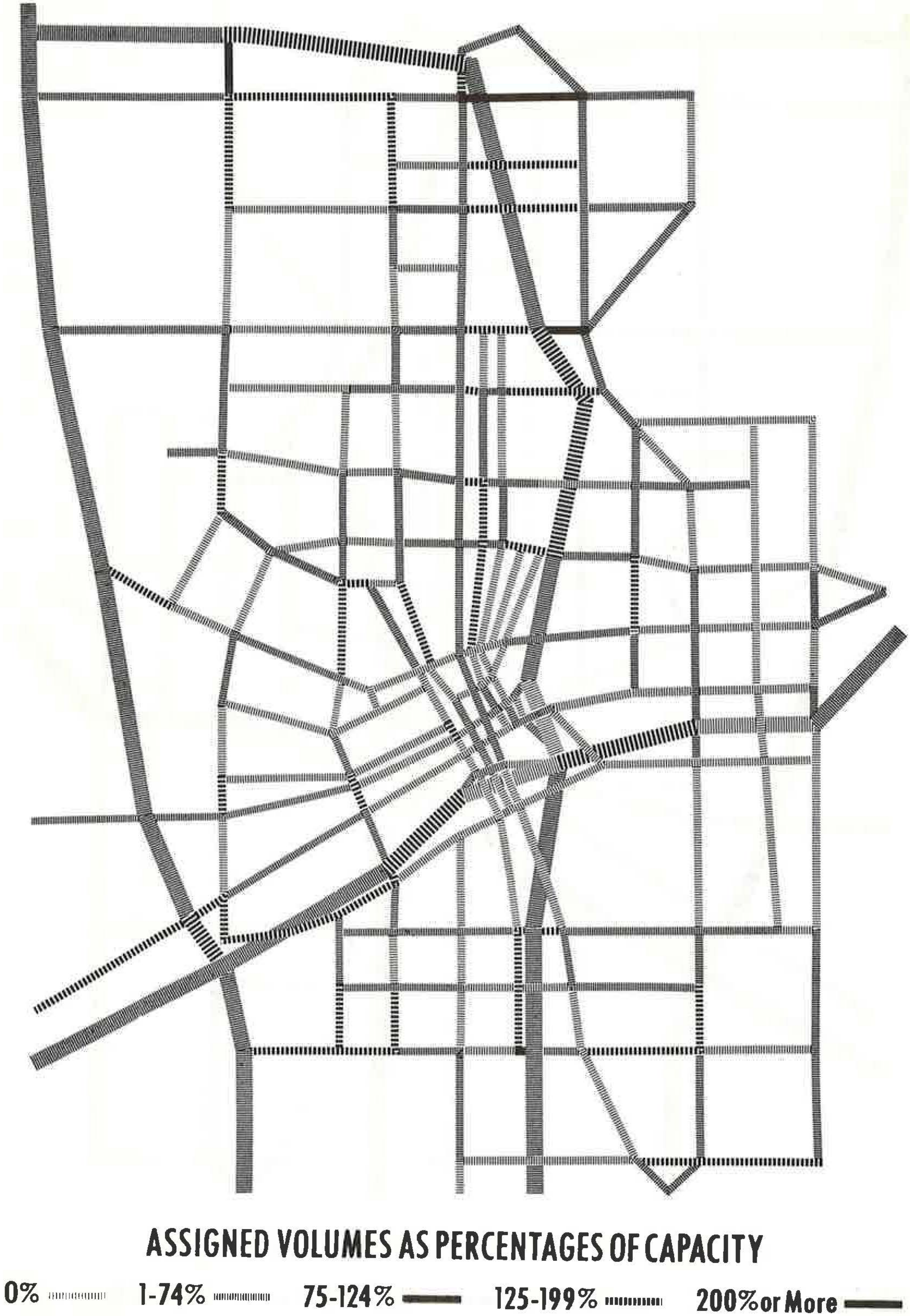


Figure 6. Volumes as percent of capacity, fourth pass.

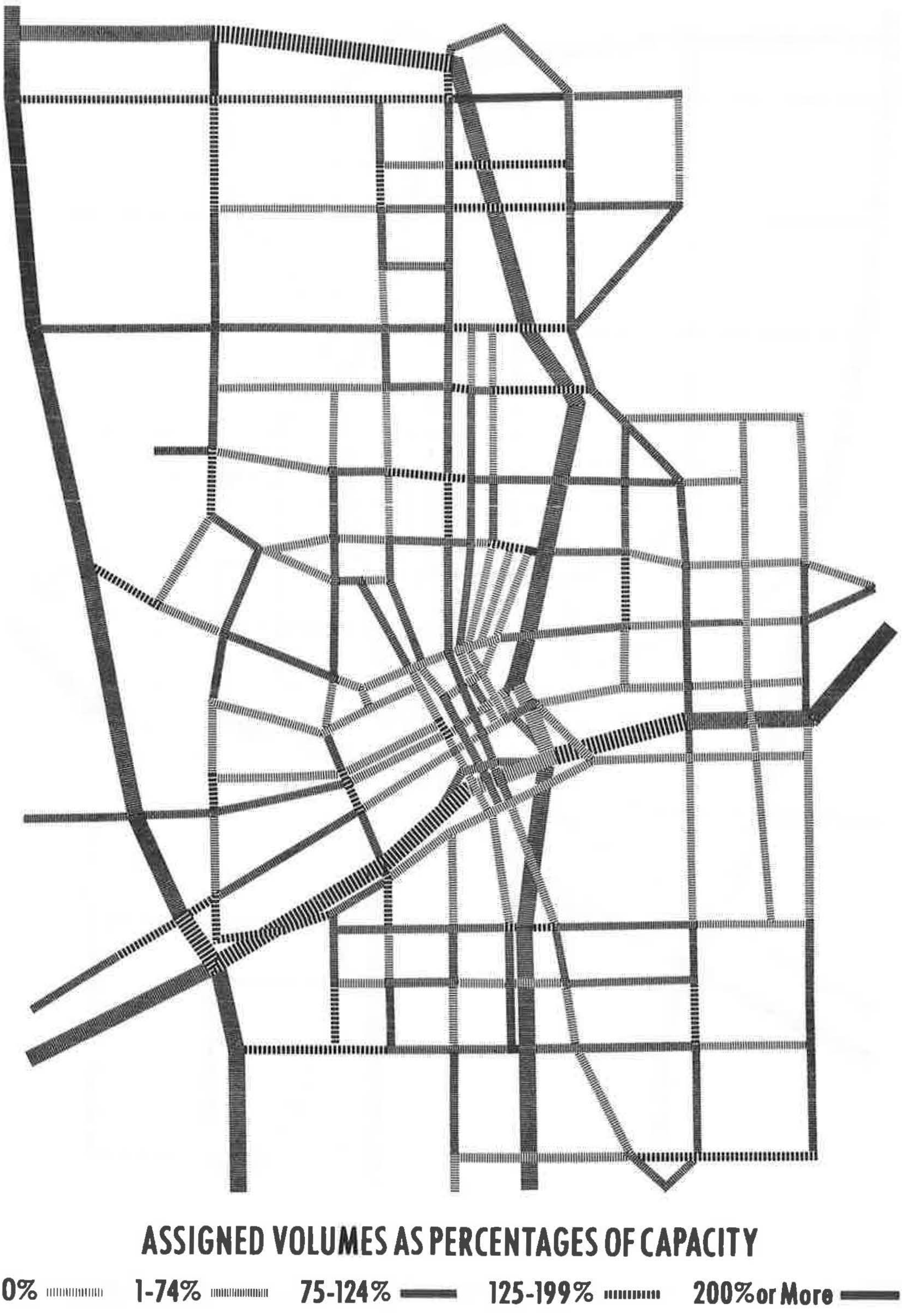


Figure 7. Volumes as percent of capacity, fifth pass.

TABLE 2

LINKS BY ASSIGNED VOLUME AS A PERCENTAGE OF CAPACITY AT BEGINNING OF CAPACITY RESTRAINT

Capacity (%)	First Pass	Second Pass					
		0%	1-24%	25-74%	75-124%	125-199%	200% +
0	21	2	3	2	3	3	8
1-24	66	0	11	27	13	11	4
25-74	85	0	1	42	29	8	5
75-124	60	0	0	36	16	6	2
125-199	44	0	0	8	26	8	2
200+	21	0	0	0	10	9	2
Total	297	2	15	115	97	45	23

One important question is whether the balancing process is continuing with different links being congested on different passes, or whether links that are congested on one pass continue to be congested on the next. This question is answered in Tables 2 and 3. Table 2 shows that a number of originially underassigned links are overassigned after the second pass. Also, 8 links originally overassigned were underassigned after the second pass. Table 3 shows that nearly all the links that were congested after the fourth pass had also been congested after the third.

Vehicle-Miles

At this point it could be concluded that, in terms of speed of convergence, this network needs certain revisions in street widths before its design is finalized, but that it is generally a good network. There are other criteria, however, and one of them is given in Table 4—vehicle-miles. This table shows that the 930,000 vehicles assigned to the network would travel a total of 3,928,000 vehicle-miles on their average weekday if all of them could travel over their "best" paths, for an average of 4.22 miles per vehicle trip.

The paths traced for the second pass are longer, however, requiring nearer 4,000,000 vehicle-miles of travel. When volumes are averaged over the first two sets of paths there are 31,000 additional miles of travel compared to the desire assignment. Paths

TABLE 3

LINKS BY ASSIGNED VOLUME AS A PERCENTAGE OF CAPACITY AT CONCLUSION OF CAPACITY RESTRAINT

Capacity (%)	Third Pass	Fourth Pass					
		0%	1-24%	25-74%	75-124%	125-199%	200%+
0	0	0	0	0	0	0	0
1-24	8	0	6	2	0	0	0
25-74	112	0	2	99	11	0	0
75-124	111	0	0	15	91	5	0
125-199	59	0	0	0	17	39	3
200+	7	0	0	0	0	6	1
Total	297	0	8	116	119	50	4

TABLE 4

VEHICLE-MILES

Type	Vehicle-Miles (thousands)				
	1st Pass	2nd Pass	3rd Pass	4th Pass	5th Pass
Total	3,928	3,959	3,997	3,988	3,975
Avg. trip-miles	4.22	4.26	4.30	4.29	4.27
Freeways	1,983	1,339	1,512	1,581	1,594
Hi-types	1,171	1,255	1,232	1,266	1,277
Lo-types	774	1,365	1,253	1,141	1,104
CBD	603	568	612	603	589
City	2,103	2,188	2,224	2,227	2,216
Suburbs	1,222	1,203	1,161	1,158	1,170

of the third pass are longer still, but there is some reduction on the fourth pass.

If the four passes are considered to be the complete assignment, then the effort to avoid congestion in this particular network requires some 60,000 additional vehicle-miles of travel on an average weekday, compared to the mileage required if all trips could travel their "best" paths. Whether this is excessive depends on whether the additional mileage could be less on an alternative network. Averaged over all the 930,000 vehicular trips, this appears to be a minimal increment.

Due to capacity restraint about 400,000 fewer miles are assigned to the freeways than would be the case if all vehicles took their "best" paths. About 460,000 more miles are required on other arterial streets: 100,000 on major arterials (divided highways, one-way streets, etc.), and 360,000 on standard arterials. In Flint, the additional miles are driven in the general city, rather than in the CBD or the suburbs.

In terms of vehicle-miles, then, this appears to be a generally "good" network at supplying capacity where it is needed. Its degree of goodness is measured by the difference between 4.22 vehicle-miles driven on the average interzonal trip over its "best" path, and 4.29 vehicle-miles for the average trip after the capacity-restraint adjustments.

TABLE 5

VEHICLE-HOURS AT TYPICAL SPEEDS

Type	Vehicle -Hours (thousands)				
	1st Pass	2nd Pass	3rd Pass	4th Pass	5th Pass
Total	135	166	161	156	154
Avg. trip-minutes	8.7	10.7	10.4	10.1	10.0
Freeways	39	26	30	31	31
Hi-types	48	52	51	52	53
Lo-types	47	88	80	73	70
CBD	21	27	27	26	25
City	79	101	97	94	93
Suburbs	35	38	37	36	36

TABLE 6

VEHICLE-HOURS AT CONGESTION-ADJUSTED SPEEDS

Type	Vehicle-Hours (thousands)				
	1st Pass	2nd Pass	3rd Pass	4th Pass	5th Pass
Total	388	311	186	164	155
Avg. trip-minutes	25.0	20.1	12.0	10.6	10.0
Freeways	71	24	29	32	32
Hi-types	56	51	46	47	48
Lo-types	261	236	111	85	75
CBD	19	34	24	20	19
City	145	175	103	93	88
Suburbs	224	102	59	51	48

Vehicle Hours and Speed

If all 930,000 vehicular trips are considered as being driven at typical speeds on every pass, regardless of congestion, the vehicular hours of driving time required on this network are as given in Table 5. On their "best" paths, vehicles would drive a total of 135,000 vehicle-hours on an average weekday, and the average trip would take a little less than 9 minutes. Capacity restraint adds about 21,000 vehicle-hours, however, and brings the average trip to about 10 minutes. This is not a realistic picture, of course, because congestion on "best" paths would reduce speeds below the typical.

Table 6 gives the influence of congestion on vehicle-hours. They are increased so much on the "desire" assignment that the average trip takes 25 minutes instead of 9. Nothing like this could occur in actuality, however, because all trips could not possibly be contained in their "best" paths. By the end of the fourth pass, a major degree of convergence has taken place, and the average trip in congestion-adjusted hours takes only about ½ minute longer than in unadjusted hours. The convergence is even greater at the conclusion of the experimental fifth pass but, as noted earlier, the fourth pass can be considered to end the assignment because the problems remaining at its conclusion are still there after the fifth pass.

This is particularly clear when speeds are considered, as in Table 7, where vehicle-miles have been divided by the two measures of vehicle-hours. At typical speeds, all these 930,000 vehicle-trips on an average weekday could be made at an average speed of 29 mph. The effect of congestion as measured by Eq. 1 is to reduce this average speed to 10 mph if all vehicles were to use their "best" paths.

On each successive pass, the congestion-adjusted speed comes closer to the typical speed for the same traffic distribution, and by the end of the fourth pass they are nearly the same. The convergence of average speeds for all trips is complete after the fifth pass, but after both the fourth and fifth pass it is clear that congestion is continuing on some standard arterial streets in the suburbs. Like the convergence of assigned volumes and capacity, then, this convergence of typical speeds and congestion-adjusted speeds suggests a generally good proposed network, but also indicates that certain types of streets in certain areas are in need of design revision.

Paths

The fourth criterion of a "good" network is that its freeways should tend to carry the longer trips. In the trip table assigned here, the 930,000 vehicle-trips are divided over 3,980 specific interzonal exchanges, and a path must be computed through the network for each of these 3,980 interzonal trips. Table 8 gives the number of such paths using the freeways, and the number not using them, by path length in miles. The gen-

TABLE 7

TYPICAL SPEEDS AND CONGESTION-ADJUSTED SPEEDS

Type	Speed (mph)									
	1st Pass		2nd Pass		3rd Pass		4th Pass		5th Pass	
	Typical	Adjusted	Typical	Adjusted	Typical	Adjusted	Typical	Adjusted	Typical	Adjusted
Total	29	10	24	13	25	21	26	24	26	26
Freeways	51	28	52	56	50	52	51	49	51	50
Hi-types	24	21	24	25	24	27	24	27	24	27
Lo-types	17	3	16	6	16	11	16	13	16	15
CBD	29	32	21	17	23	28	23	30	24	31
City	27	15	22	13	23	22	24	24	24	25
Suburbs	35	5	32	12	31	20	32	23	33	24

eral conclusion to be drawn is that the vast majority of short trips do not use the freeways, whereas the vast majority of long trips do.

There is another important point to be made about paths—the success of a traffic assignment in predicting actual traffic counts depends on two things: (a) the validity of the assigned trip table and (b) the realism of the selected paths. All other factors such as the character of the capacity-restraint formula, the accuracy of the estimates of capacity and speed, and the quality of classifications of street by type and area are of importance only as they lead to the computer simulation of actual driver behavior in the selection of paths.

A sample of paths was selected for examination from the tape record for the present assignment passes, although this operation is not a necessary part of capacity-restrained traffic assignment. The most typical situation which was found (Fig. 8) shows the paths traced on each pass between the external station at Saginaw and Dort on the south and the same two highways on the north. For its "best" path (from the desire assignment), the trip has been routed up Saginaw to the first street giving access to the branch north-south freeway, and thence to the freeway itself. It travels up the freeway to the northernmost ramp and then takes one link of Saginaw to its destination.

The so-called "rush-hour" path (second pass) avoids the congested freeway and

TABLE 8

LENGTH[a] OF FREEWAY AND NON-FREEWAY PATHS BY PASS

Miles	First Pass		Second Pass		Fourth Pass	
	Freeway	Non-Freeway	Freeway	Non-Freeway	Freeway	Non-Freeway
0-3	300	1,520	80	1,660	230	1,560
4-6	880	410	360	1,130	670	680
7-9	610	30	290	320	530	100
10+	230	0	100	40	200	10
Total	2,020	1,960	830	3,150	1,630	2,350

[a] Rounded to tens.

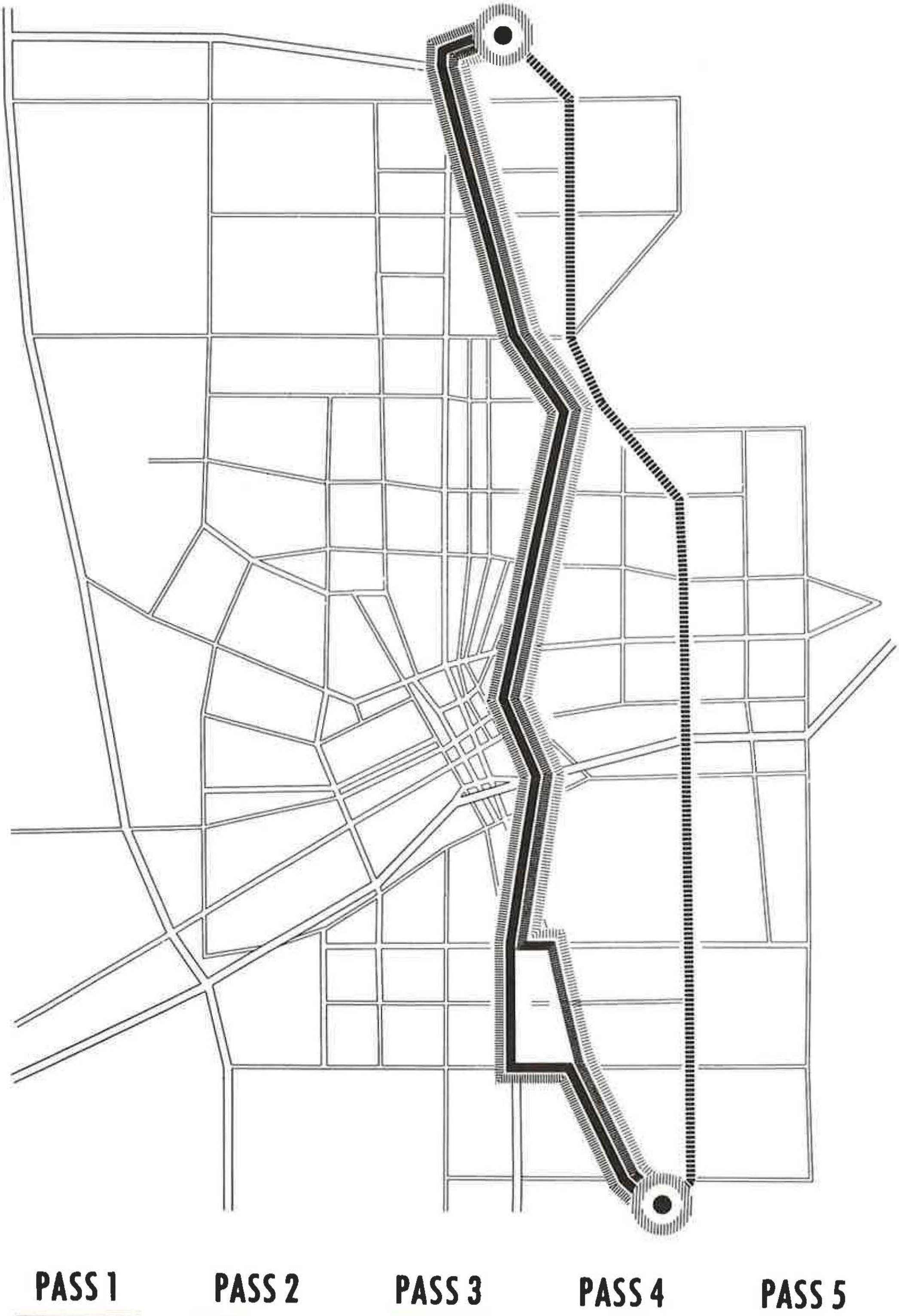


Figure 8. Set of paths A.

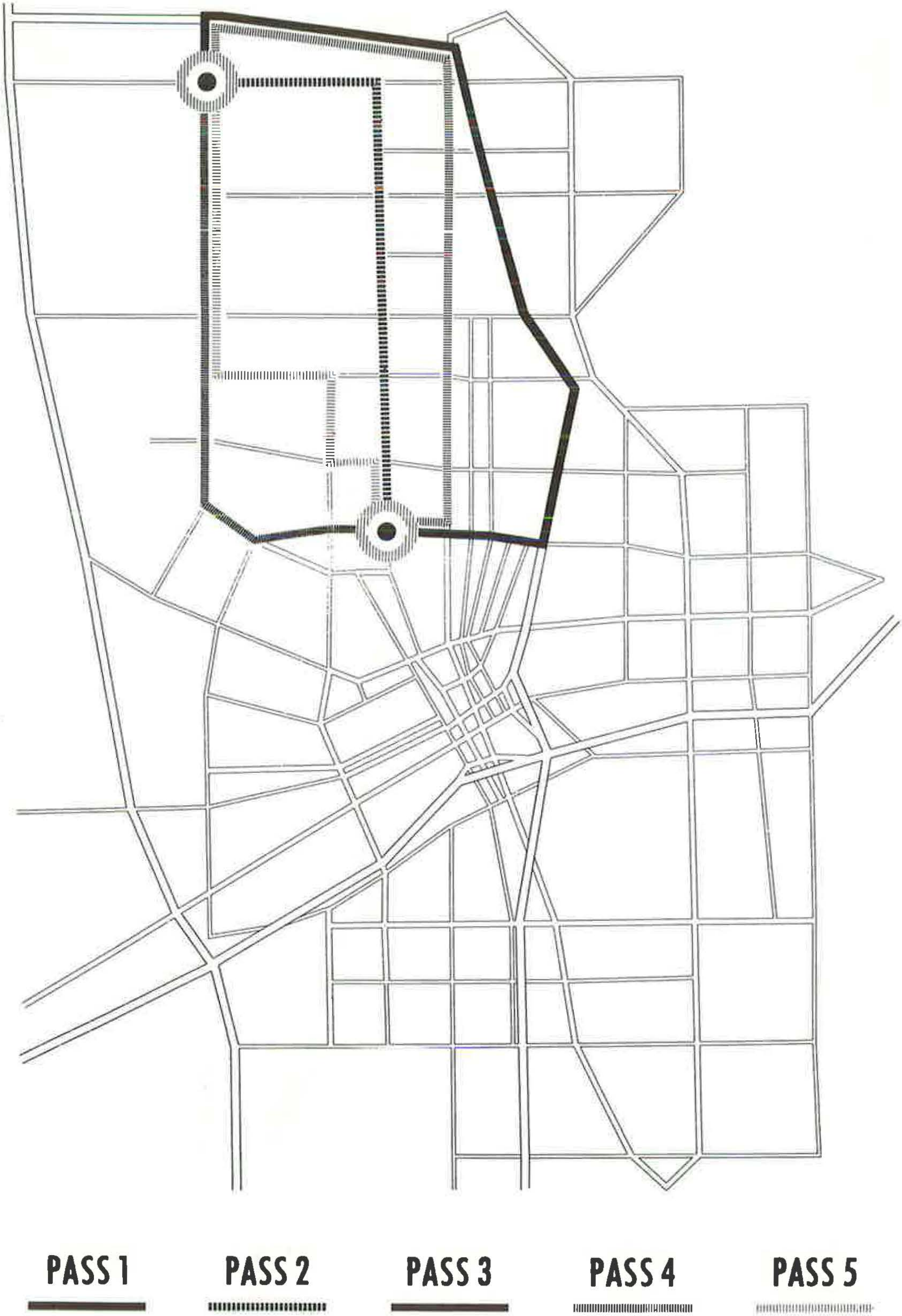


Figure 9. Set of paths B.

travels up Dort Highway all the way from origin to destination. When all volumes have been divided between their first two paths, congestion-adjusted speeds on the freeway are high enough again to bring the path back to it, with the exception of one link. This is shown by the third path, which indicates that the southernmost freeway link in the "best" path has remained congested enough, and the alternative is good enough, so that the path continues up Saginaw to the second street giving access to the freeway. From the point of entering the freeway, this third path is the same as the first path. The fourth pass is exactly like the first, and the fifth pass is exactly like the third.

Although this rapid convergence is typical of many parts of the network, another situation is shown in Figure 9. These paths connect two intersections in the northwest section where there is insufficient capacity in the network. No two of the five paths are exactly alike, and to some extent the situation is one of paths moving farther away from the congested north-south corridor through the center of the area. Of course, at the conclusion of the fifth pass, 20 percent of this trip's volume is on each one of these paths so that some of its drivers are assumed to prefer the freeway and some are assumed to prefer the more direct but slower routes over standard arterial streets.

Figure 10 shows another situation in the case of the paths connecting the intersections at the outer southeast and northwest corners of the network. Here the upper parts of three of the paths use the branch north-south freeway and two use the outer north-south freeway, whereas on later passes their southern parts avoid the congested central corridor by staying to the east until they can take the east-west freeway to its interchanges with the others.

As mentioned earlier, a more limited number of alternative paths than this is the usual situation. In fact, in the case of 3 of the 22 paths examined, exactly the same path was traced for all five passes. For example the trip between the southern and the northern terminals of the north-south freeway along the western edge of town do not leave that freeway on any pass.

The important point is that this proposed network has been tested by a capacity-restrained assignment method which has distributed volumes over paths reflecting familiar driver behavior. That is, they are distributed over reasonable and realistic paths. The consequences have been discussed in terms of convergence, mileage, speeds, and paths, with the conclusion that the network is generally a "good" one in terms of the locations of major facilities. All of these considerations refer to only one of the two uses of traffic assignment, however—the use for network testing. The other use is in connection with the design of individual facilities. The kind of information provided by a capacity-restrained assignment for this purpose can be illustrated by the link record.

Link Record

Table 9 gives the kind of information available for every link in the network before and after an assignment. Reading across are the fifteen crucial digits which constitute the complete description of the link that the computer must have before the assignment begins. These are the numbers of the two intersections the link connects, the type and the area of the link, its pavement width in feet, its practical 24-hr capacity in thousands of vehicles, and its length in miles.

Table 10 gives the link's travel time and assigned volume, pass by pass. At the far end of the first line is its original (or typical) speed for the desire assignment. This number varies with both type and area (Table 1). The length of the link divided by its speed gives the link's "value" (V) for path-tracing purposes, which is travel time expressed in thousandths of hours. Beside it is recorded the same travel time expressed in minutes, simply for ease of interpretation.

When the computer has the fifteen crucial digits plus V for every link the assignment can begin. Link record A is the link record for the curved segment of the branch north-south freeway which appeared as heavily congested in Figure 3 (after the first pass). On the line for Pass 1 of this record is the volume of 129,000 vehicles which was assigned to this link on a "desire" basis. This is followed by 2.30, which means that the assigned volume is 230 percent of the capacity of the link (that is, R = 2.30).

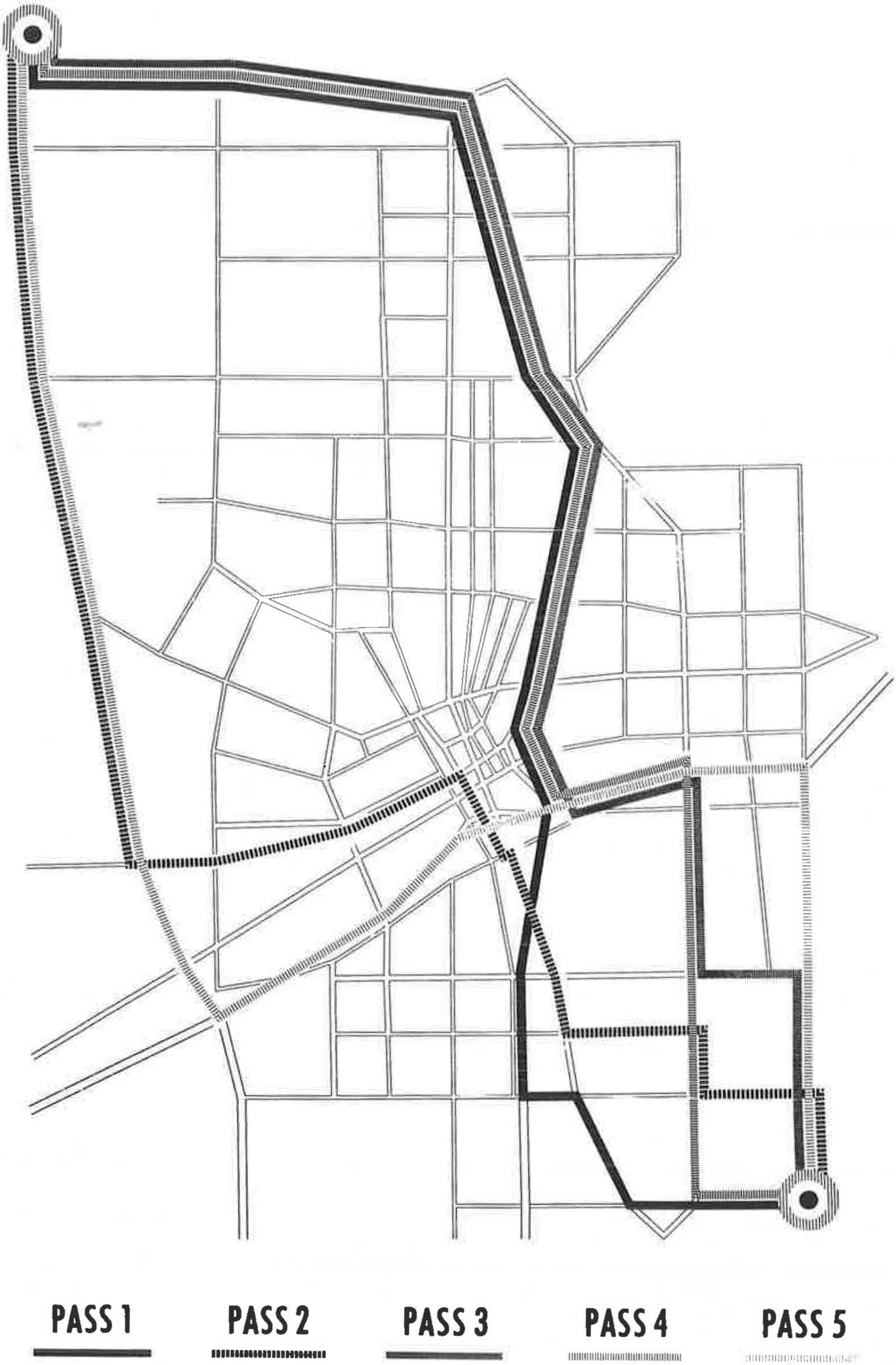


Figure 10. Set of paths C.

TABLE 9

LINK RECORDS

Link Record	Intersections		Type	Area	Width (ft)	Capacity (×1,000)	Length (mi)
A	150	152	1	2	48	56	2.09
B	177	178	3	2	44	14	0.55
C	38	125	3	2	44	16	0.55

At this point Eq. 1 is applied, and the "typical" travel time is increased by 2.7 raised to the R^{-1} power. In this case, the equation produces a V of 150, which means 9 minutes of travel time, which in turn means that the speed on the link is dropped to 14 mph for the second pass. This speed is only slightly below the typical speed of a standard major street in the city, but such streets tend to be underassigned on the first pass and to have higher than typical speeds for the second pass. In any case the speed of 14 mph was such that this freeway link was included in no path on the second pass and was assigned no volume. But one-half the volumes of the trips using this link on the first pass are still considered to be assigned to it, so its average assigned volume after the second pass is 64,000 vehicles, or 115 percent of its capacity.

This volume allows a speed of 44 mph on the link, according to Eq. 1, and this speed determines its travel time for the third pass. At that point the trips whose paths include this link have a total volume of 93,000 vehicles, but only one-third of this volume, along with one-third of previously assigned volumes, is considered to be actually

TABLE 10

ASSIGNED VOLUMES AND TRAVEL TIMES FOR LINK RECORDS

Link Record	Pass	Assigned Volume			Travel Time		
		Avg. (×1,000)	Per Pass (×1,000)	R	V (hr × 10^{-3})	Min	Speed (mph)
A	0	-	-	-	41	2.5	50
	1	-	129	2.30	150	9.0	14
	2	64	0	1.15	47	2.8	44
	3	74	93	1.32	55	3.3	38
	4	77	88	1.38	59	3.5	35
	5	70	39	1.24	52	3.1	40
B	0	-	-	-	37	2.2	15
	1	-	0	0	13	0.8	42
	2	33	66	2.36	144	8.6	4
	3	22	0	1.56	65	3.9	8
	4	17	0	1.18	44	2.6	13
	5	13	0	0.95	35	2.1	16
C	0	-	-	-	37	2.2	15
	1	-	2	15	15	0.9	37
	2	8	13	47	21	1.3	26
	3	14	26	85	31	1.9	18
	4	14	15	88	32	1.9	17
	5	14	14	88	32	1.9	17

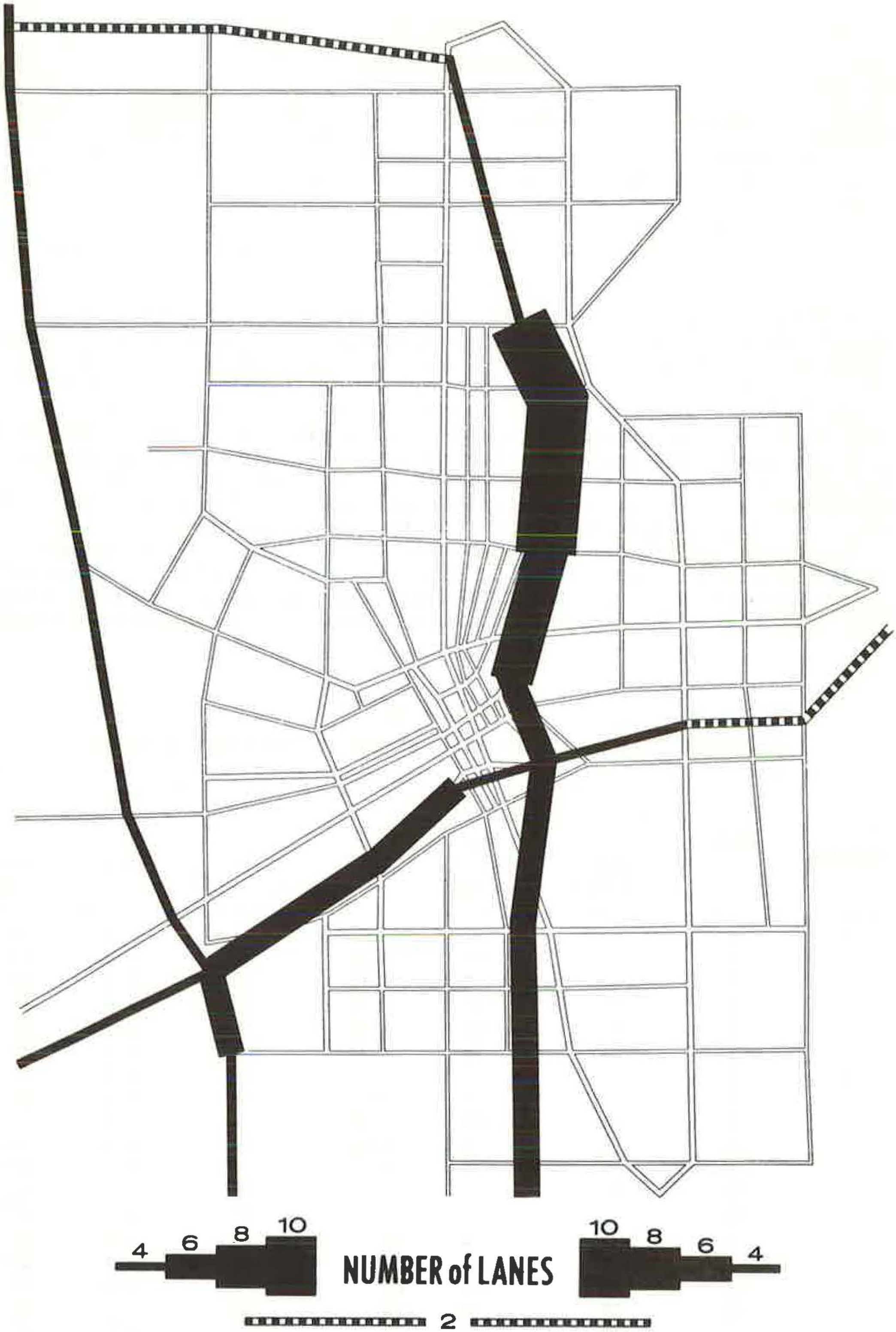


Figure 11. Freeway design, hand trace.

assigned to the link. Its average assigned volume, then, is 74,000 vehicles after the third pass, or 132 percent of capacity.

After the fourth and final pass, the average assigned volume is 77,000 vehicles, or 138 percent of capacity, which is the capacity-restrained assigned volume. It allows a speed of 35 mph on the link, and would lead the planner to consider increasing the width of the link from four lanes to six. (Assignment of peak-hour or design-hour volumes could be relied on to lead to the same conclusion only to the extent that such hourly volumes are some fairly constant percentage of the 24-hr volumes used here.) The assigned volume after the experimental fifth pass is somewhat closer to the original capacity but still represents a degree of congestion and a necessity for redesign.

Link record B is a link record of a different kind. It represents a link of a standard major street in the city which is about four lanes wide and can carry an estimated 14,000 vehicles per day without congestion, at an average speed of 15 mph. On the first pass, it was assigned no volume at all and its speed was increased to 42 mph. In this case, the increased speed led to an overassignment on the second pass and speed dropped to a mere 4 mph. No volumes were assigned to it thereafter, and its final volume was close to capacity.

Link records A and B both represent extreme situations, whereas link record C is the record for a more typical link. It appeared in few paths on the first pass and its speed was more than doubled, then reduced again on subsequent passes. By the third pass, its assigned volume was close to capacity and did not change again, so that it appears that the link is designed about as it should be for 1980 traffic volumes.

This completes the description of the kind of information available from a capacity-restrained traffic assignment. Consideration has been given to (a) the four tests of the total system in terms of convergence, miles, speeds, and paths, and (b) the test of every individual link with the related data of assigned volume, width, and speed. The following section compares this with the kind of information available from other types of traffic assignment.

COMPARISON TO OTHER ASSIGNMENTS

Freeway System

In the process of planning a Flint network, the State Highway Department made a hand-trace assignment to the version tested here as well as to two earlier versions. In addition, a diversion-curve assignment was made on the computer to the freeway subsystem of the network. Therefore, with the desire assignment and the capacity-restrained assignment described in this report, a comparison can be made of the results of four different assignments to this particular freeway layout. All of the assignments obtained both link volumes and turning movements, so that a mass of information is available for analysis, but it is summarized here in the form of four maps (Figs. 11 through 14).

Figure 11 shows the kind of freeway system that would be constructed if the volumes assigned by the hand-trace method were to be treated as design volumes. The broken line indicates that the assigned volume could be handled without undue congestion by a two-lane freeway, and the narrowest solid line indicates that four lanes would be required. The solid line is widened by ⅛ in. for each additional needed pair of lanes and the widest line indicates that a ten-lane freeway is called for by the assigned volume to that particular link.

The Highway Department does not intend to use these particular hand-assigned volumes for design purposes; they were produced as a general test of the freeway-system layout. For example the Department would not design and construct the ten-lane freeway section shown, but instead would conclude that ". . . there must be revisions in the original system to provide capacity on other facilities in the immediate vicinity." This is a useful approach and one clearly preferable to less systematic and more subjective methods of traffic estimating.

On the other hand, at the conclusion of this particular hand-trace assignment there were other facilities in the immediate vicinity which were loaded with volumes significantly smaller than their capacities. This raises the possibility that the system might

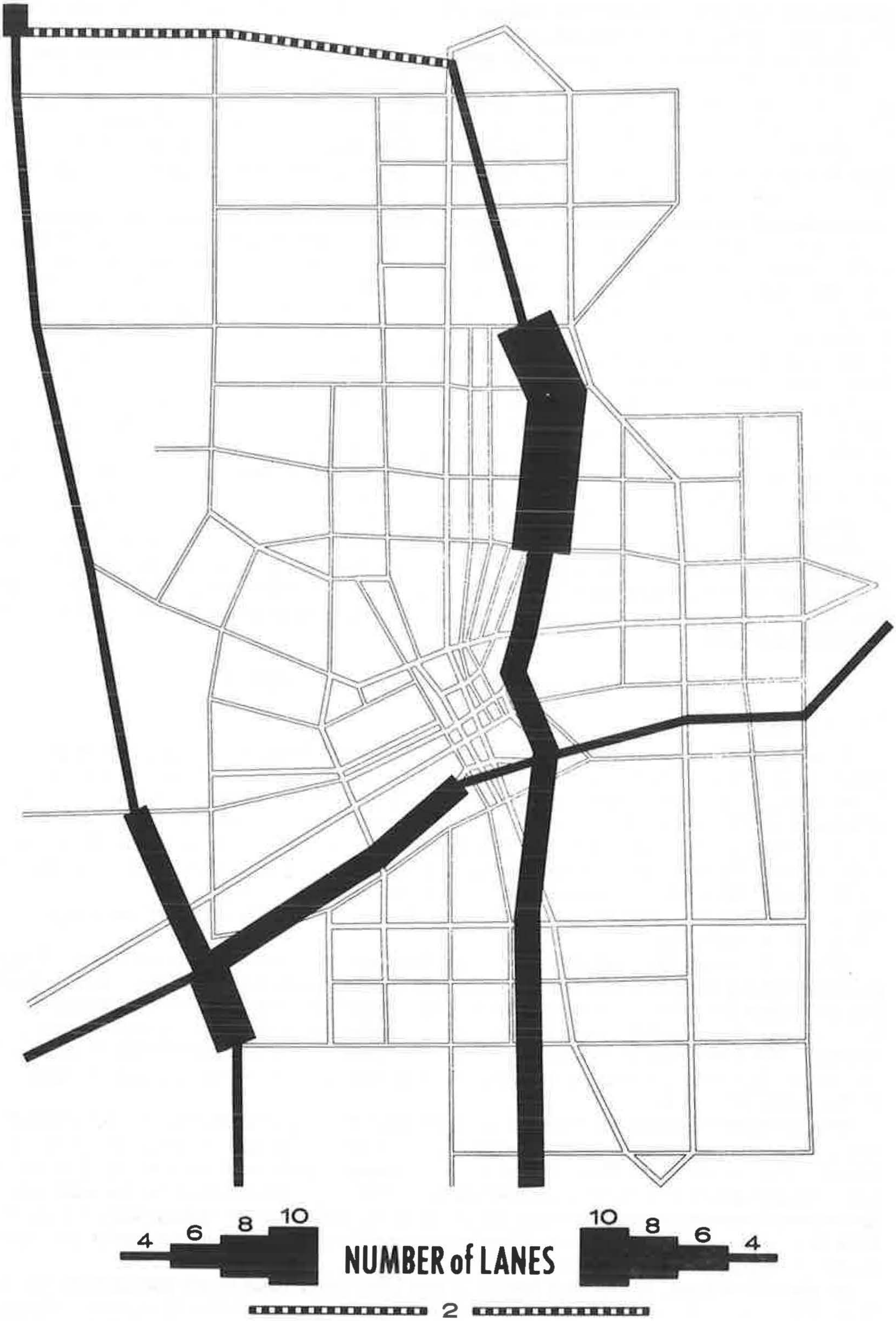


Figure 12. Freeway design, diversion curve.

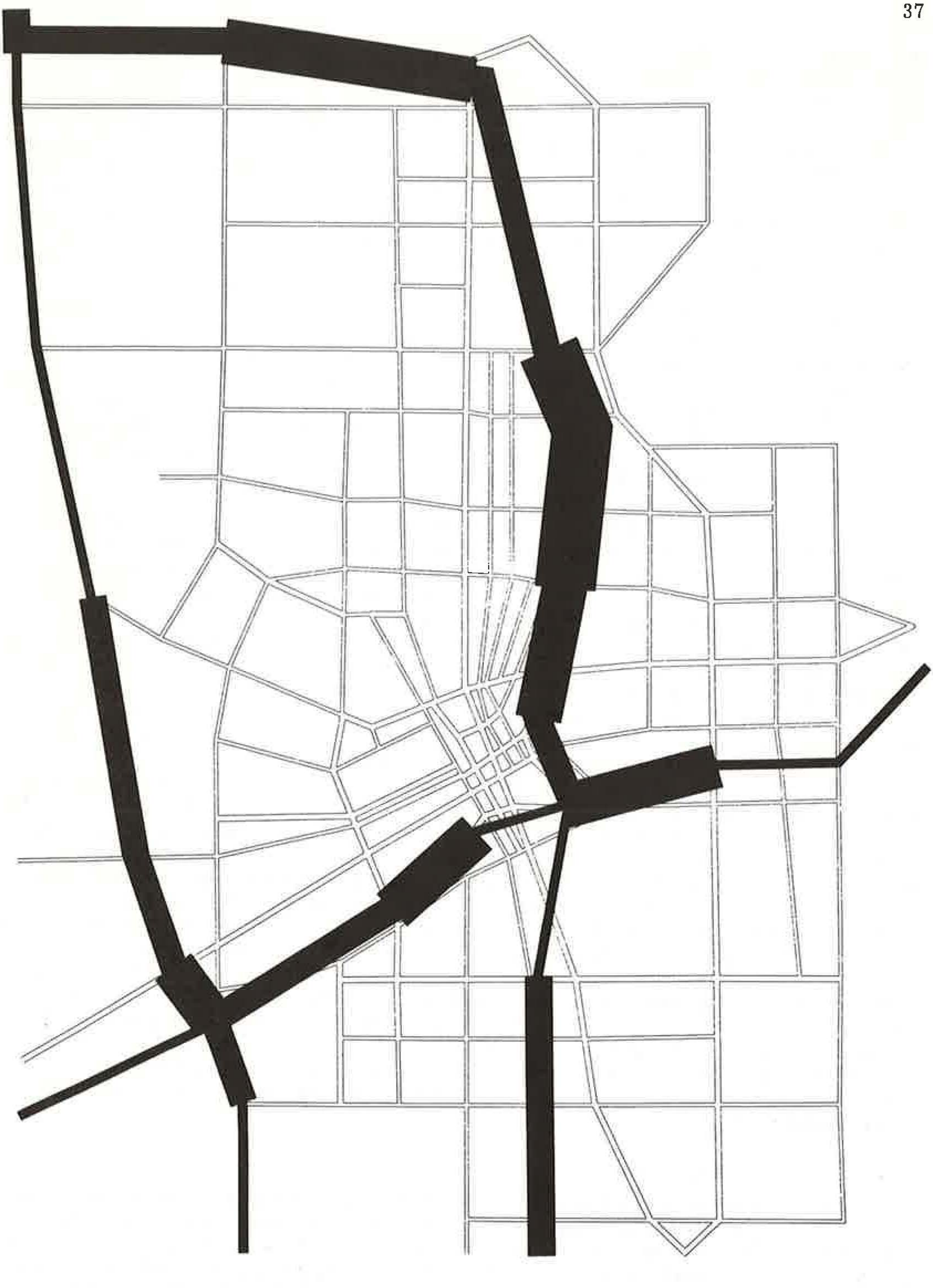


Figure 13. Freeway design, "desire."

not appear to need added capacity if the assignment technique allowed individual interzonal volumes to be divided between alternate "best paths." Such a procedure is recommended by the probability that all drivers making the same interzonal trip are not likely to select the same route if there are very many choices to be made along the way.

When hand-trace volumes are compared to the results of other kinds of assignments, another kind of question can be raised. For example, the hand-trace assigned 23,000 more vehicles to the most heavily assigned link than any of the other methods (to be specific, a total of 152,000 vehicles compared to 129,000, 115,000 and 77,000 by the other methods). What factors did the hand tracers take into account to produce this volume? They could be correct, but the computer volumes always have one advantage: the factors taken into account to produce them can be precisely specified.

Figure 12 shows the results of a computer freeway "diversion curve" assignment, mapped in the same way as the results of the hand-trace. The two assignments produce a similar pattern of results, and the computer assignment has the advantage already mentioned. On the other hand, the computer assignment, too, can be questioned on the basis of the desirability of alternate paths. In addition, the freeway assignment has the major shortcoming of providing no assistance for the process of adapting an existing street system to a new freeway system, which is the reason the Highway Department had to do the hand tracing.

Figure 13 shows the same network as it was described at the conclusion of the first pass or desire assignment reported in this paper. Compared to Figure 12, it makes clear the effects of not dividing interzonal volumes by means of a diversion curve before assigning them (in that there are even more unrealistically high volumes), but again its general pattern is comparable to the results of the other two methods. This kind of desire assignment has been used for planning purposes in a number of cities, because it combines the advantage of objectivity offered by a computer assignment with the advantage of testing a total major street system as does the hand-trace assignment. However, the results of this assignment, too, can be questioned on the grounds that heavily overassigned and underassigned links make it unavoidable that some of its assigned volumes will vary from actual traffic flows, and this might not be the case if alternate paths could be determined by the assignment technique.

Figure 14 shows the results of the fourth pass of the capacity-restrained assignment mapped in the same way as the others. This is the only assignment discussed here that can meet the objection mentioned. It prescribes a system of seventeen freeway links of six lanes each and eight freeway links of four lanes each. It demonstrates an assigned volume on every link beyond the capacity of a smaller freeway (thus justifying construction), and within the prescribed capacity (thus realistically aiding in design).

It is confidently predicted that, if the network were revised to provide these freeway capacities and to provide additional capacity at the indicated points of access to the freeways, the revised network would pass the assignment test by "converging" quickly to typical speeds and capacities, requiring minimal additional vehicular miles and hours, and carrying longer trips on its freeways. In other words, it would prove to be a "good" network providing sufficient capacity where it is needed. More important, it is probable that the revised network would carry Flint traffic in 1980 without undue congestion, assuming the validity of the 1980 trip table and a typical distribution of local (intrazonal) traffic.

Comparative Time and Cost

It has been difficult to compare time and cost for the various assignments discussed here. The largest part of the capacity-restraint costs have been for experimental program development that will not need to be repeated. The Highway Department's record of expenditures for the hand trace includes only direct costs, and it is difficult to separate costs of the other assignments from various overhead charges. One reliable comparison is that the path-tracing part of the hand assignment took 44 man-days at the Highway Department and the path-tracing part of one comparable pass took 12 hr on the 650-RAMAC computer at Wayne State University.

This is by no means a reliable estimate of the ratio of total times and costs, however. A computer-hour is more expensive than a man-day, and a network for the

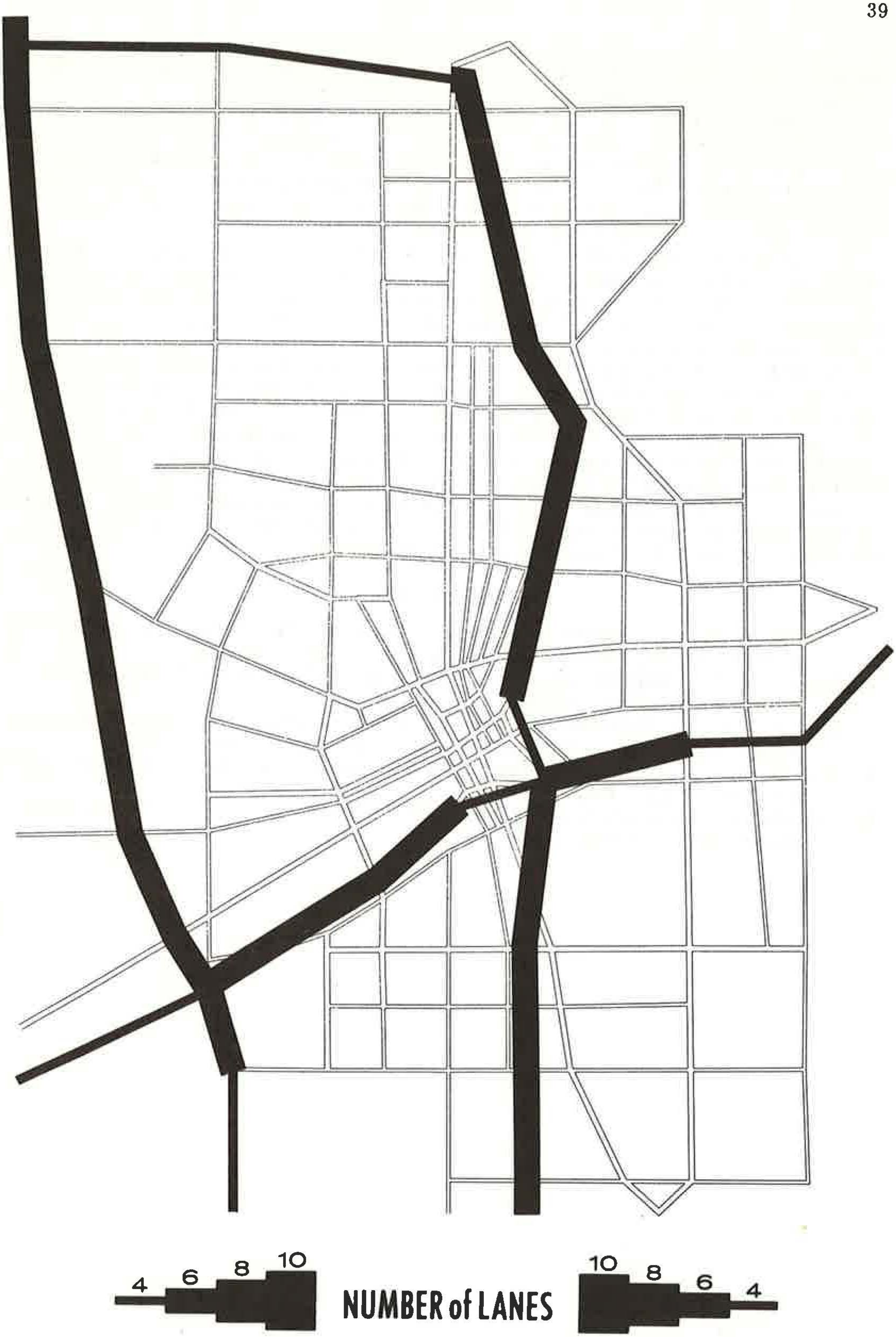


Figure 14. Freeway design, capacity restrained.

computer needs more preparation than a network for hand tracing. The Highway Department reports that it can complete a hand-trace assignment to a network like the present one in about one month at a total direct cost of about $2,000.

It seems likely that, with their own 650 computer, a capacity-restrained assignment would take them about the same amount of time from beginning to end, although much less time would be spent on path-tracing and much more time on network description and output analysis. The computer assignment probably would cost at least twice as much as the hand-trace, although the cost consequences of reducing it to a routine, mass-production operation are hard to estimate.

So long as relatively small computers are the only ones readily available to highway departments and local planning offices, the capacity-restrained assignment must justify itself on the grounds of its superior usefulness and reliability; it cannot be entirely competitive in terms of cost. This paper has tried to demonstrate, however, that they are nevertheless superior enough to justify a greater cost, and now that a 650 program can be made generally available it is to be hoped that they will be more widely used.

In the long run, the solution to the need for such assignments on a routine basis is the wider availability of high-speed computers. The U. S. Bureau of Public Roads is developing programs for the IBM 704 and 7090 which are comparable to the one described here but reduce assignment time to a matter of minutes and reduce costs to a few hundred dollars, and such computers are becoming more readily available around the country for traffic assignment on a routine basis. If this work continues, there is no doubt that it will soon be possible to provide more and better traffic assignments for highway planners than have been available in the past.

REFERENCES

1. Brokke, G. E.,"Transportation Studies for Smaller Urban Areas." U. S. Bureau of Public Roads, Washington, D. C., 14 pp. (1962). Mimeo.
2. Smock, R. B., "Determining the Capacity of Arterial Segments." Detroit Area Traffic Study (1957). Mimeo.
3. Simmons, J., "A Mathematical Model for Traffic Assignment." Unpubl. Masters Essay, Department of Mathematics, Wayne State University (1961).
4. Smock, R. B., "An Iterative Assignment Approach to Capacity Restraint on Arterial Networks." HRB Bull. 347 (1963).
5. Pakkala, P., "A Prediction of 1980 Traffic in Flint." Detroit Area Traffic Study, Wayne State University (1962).

A Capacity-Restraint Algorithm for Assigning Flow to a Transport Network

WALTER W. MOSHER, JR., Assistant Research Engineer, Institute of Transportation and Traffic Engineering, University of California, Los Angeles

The task of assigning units of movement to a complex transport network has, in the past, been accomplished by techniques that are limited by their requiring constant network link costs. In general, these existing transport flow assignment techniques ignore the fact that as the number of units of movement assigned to a network is increased, localized congestion will occur and total network cost will rise disproportionately. Thus, the effect of link capacity on transport flow is not considered, and as a result many links in the transport network become unrealistically overloaded and others receive very little or no assigned flow.

This paper presents a "capacity-restraint" algorithm which permits the evaluation of network performance based on arbitrarily selected network figures of merit. The values of these figures of merit depend on network loading, and are governed by individually determined link performance functions. Each link performance function may be any linear or nonlinear relation of link flow to cost (e.g., distance) for that flow. Once a performance function is established for each network link, the network can be loaded in an optimum manner either by minimizing the figure of merit for the entire system or by equalizing the path figures of merit over appropriate sets of paths. For road networks, the latter is the one most likely to occur.

• THE QUESTION of how a complex transport system should be designed must be considered in the context of what the ultimate goals are of the affected socio-economic group. Further, because these ultimate goals are not static but dynamic functions of the progress of society, it is necessary to have planning techniques that allow new technology to influence the design and ultimate realization of the system.

The problem reduces to that of supplying to the populace a system designed to achieve optimum living conditions. This implies maximum interaction and freedom of movement at a minimal cost, and therefore requires that desired living conditions be compromised if the minimum total cost for a particular system cannot be borne by the populace.

To design such a system it is necessary to determine what factors comprise optimum living conditions. A recent survey of 400 recognized leaders of the planning profession established that there are at least 6 broad planning criteria (8). These are resources, land use, economic, transportation, socio-humanistic, and catastrophe. Of course, the tabulation of these broad planning criteria does not imply that the relative importance of them and their associated subcriteria is known. The nature of these criteria is such that they interact with each other to a great extent, thereby making it extremely difficult to determine absolute relations between them and the desired, social-oriented, living conditions.

It is evident from the preceding that the first step in designing a transport system

Paper sponsored by Committee on Origin and Destination.

is to establish, empirically or theoretically, these relations between the various planning criteria. Inasmuch as the determination of the relative importance of the factors constituting "optimum" or "desired" living conditions is subject to individual interpretation, and because they are also subject to change as society advances, it is imperative that these relations provide for dynamic system goals.

After this has been accomplished, an analytical model representing the proposed system as well as the present system must be devised. This entails the determination of mathematical relations for and between the various planning criteria as functions of time and cost (in the general sense; e.g., safety, travel time, distance, and psychological stress). Boundary conditions in consonance with the proposed society goals must also be determined. Whenever possible, these boundary conditions should not be expressed as discrete values, but instead as regions within which system realization would be acceptable. For example, in the case of land allocation to families, rather than state a particular square footage per person (e.g., 15,000 sq ft), the constraint should be given as not less than a fixed amount or as between two fixed limits (e.g., 7,500 to 20,000 sq ft).

When this has been accomplished, analytical techniques may be used to determine the compatibility of the various subsystems comprising the planned system, and to indicate necessary revisions where incompatible situations exist. Further, as the planned system is effectuated, information concerning its operation at any given time may be used to check the correctness of the system model and to indicate changes that will improve the accuracy with which future system configurations can be forecast.

THE TRANSPORT PLANNING PROBLEM

Once a set of defining factors for a proposed system has been established, including its associated equations and parameters, it is necessary to determine the type and extent of facilities required to accommodate it. In the transportation sense, this task amounts to the design of roads, streets, freeways, and mass transit facilities such that the desired system interaction and freedom of movement is provided. For the support functions of the system (e.g., communication, power, sewage, and flood control), the design task is that of providing adequate capacity to handle the expected loads. For both transportation and support, it is desirable to build toward the ultimate required capacity in parallel with total system realization and in such a way that the partially realized system is operating in its optimum manner at each moment of time.

It is apparent that many alternate designs for accommodating the various transport fluxes are possible. Route selection, capacity, configuration, accessibility, etc., of each network link will affect the overall system realization, because the cost of constructing each link as well as the cost to society for using that link will vary as these factors are changed. It is, therefore, necessary to consider many alternate plans when attempting to find the best configuration for a given system, and to balance the cost of constructing each configuration with the benefits derived therefrom.

Because there is no assurance that any particular design will attain a solution in consonance with the defined system goals, each proposed design must be tested and, if necessary, modified until an allowable solution is achieved. These considerations indicate that an iterative procedure is required that will allow the evaluation of how well the system goals are satisfied after each iterative pass, and which will indicate changes that should or must be effectuated to achieve these goals better.

The heart of such a procedure will be a technique for the determination of patterns of movement within a proposed system over its proposed transport links. This entails the determination of the number of units of movement, for each transport flux, between every pair of origins and destinations, and then the assignment of these units of movement to appropriate transport links. Because the number of units of movement between zones is a function of the cost to accomplish these movements, which is, in turn, dependent on the total network loading, an iterative procedure is again suggested. In this instance, the first iteration cycle begins by determining the number of units that will move between each zone pair for an estimated set of link use costs. These costs might be estimated from the link cost function by arbitrarily setting all link flow values to zero. Next, the network is loaded, and the resulting link costs determined. From these new costs, the interzonal movements are redetermined, thus beginning the second iteration cycle. This procedure is repeated until the interzonal movements do not change significantly from one iteration cycle to the next.

From the preceding discussion, three distinct phases to the transport planning problem become evident. They are (a) design of the transport networks, (b) determination of interzonal movements for each transport flux, and (c) assignment of the interzonal movements to the transport networks.

Figure 1 shows the sequence of these phases of the transport planning problem within the framework of the total socio-economic system planning task. As shown, a mathematical representation of the "desired" or "optimum" living conditions is established by relating the various planning criteria to each other and to the socio-economic group. These relationships form the framework for a mathematical model of the transport plans, and are used to determine the extent to which any proposed transport plan is compatible with the "desired" living conditions. From the mathematical model, the transport networks are designed and loaded. Network loading is done in consonance with the mathematical relationships of the model that govern interzonal movement. Using the loaded networks as a basis, the cost (e.g., freedom of movement, esthetics, safety, as well as dollars) to the socio-economic group for using the proposed system is determined and compared to their desired goals.

The next section of this paper is a brief review of existing techniques for the forecasting phase (phase 2) and the assignment phase (phase 3) of the transport planning task.

EXISTING NETWORK TRANSPORT ASSIGNMENT AND PERFORMANCE EVALUATION TECHNIQUES

During recent years, the transport forecasting and transport assignment phases of the transport planning task have received considerable attention from the planning profession. The end result is a multiplicity of techniques for studying transport network performance. Transport forecasting involves the prediction of future travel patterns

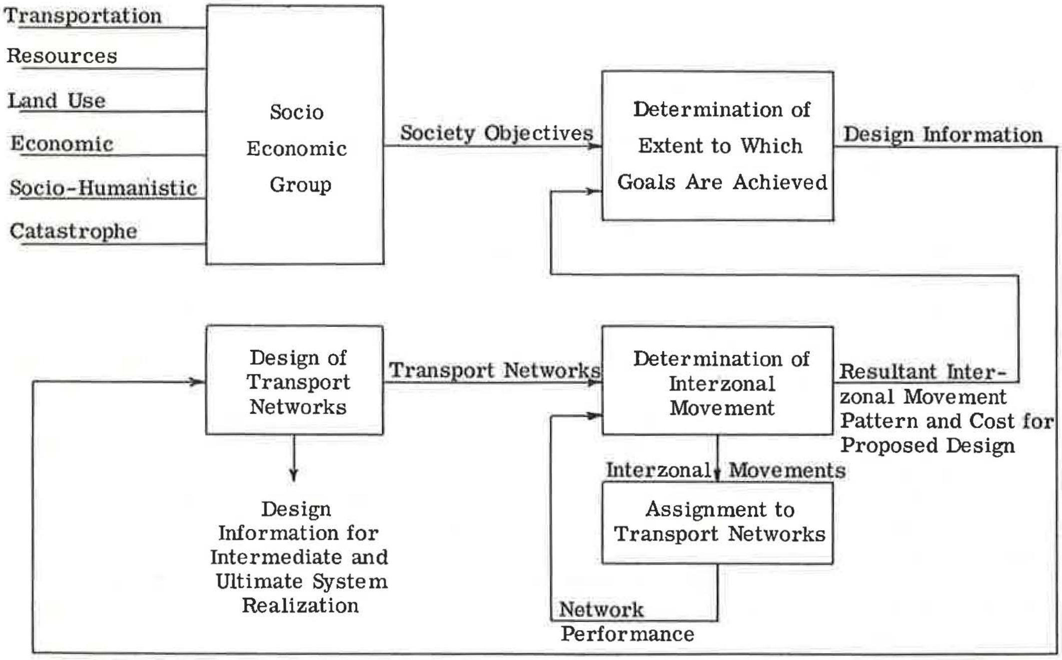

Transportation
Resources
Land Use
Economic
Socio-Humanistic
Catastrophe
Socio Economic Group
Society Objectives
Determination of Extent to Which Goals Are Achieved
Design Information
Design of Transport Networks
Transport Networks
Determination of Interzonal Movement
Resultant Interzonal Movement Pattern and Cost for Proposed Design
Interzonal Movements
Assignment to Transport Networks
Network Performance
Design Information for Intermediate and Ultimate System Realization

Figure 1. Socio-economic system planning.

and requirements for a transport system. Transport assignment, on the other hand, entails the determination of routes, within the proposed or existing transportation network, over which the forecast transport flows will move.

Transport Forecasting

Numerous theories and procedures have been developed to assist the transport planner in the task of predicting future transport requirements. The common denominator of these techniques is a recognition that future transport requirements depend on the ultimate configuration of the land use plan. In this respect, the type, intensity, and rate of development of the land use for each zone of a planned system greatly influences the evolution of every transport network within the system. If a region is to expand and grow according to some master plan, it is imperative that techniques be made available with which the system planner can determine how to provide transport facilities at a rate commensurate with total region expansion. To this end, transport forecasting and transport assignment work hand in hand; however, the best transport assignment technique cannot reliably determine network performance if the transport forecast is inaccurate.

The multitude of transport forecasting techniques fall within four broad classifications. These are growth factor models, gravity models, opportunity models, and multiple regression models.

Until recently, the most popular techniques were based on the growth factor model. These techniques accomplish the forecasting by multiplying the existing transport interzonal flows by predicted growth ratios. These ratios were determined by dividing the expected future transport demand by present transport demand. The original techniques did this, based on a system-wide ratio of expected growth, and thus the resulting forecasts were not accurate in zones where extreme differential growth occurred. This problem was very serious, particularly in metropolitan areas where urban development was replacing agricultural land use. In these situations, the demand for transport service at the time of a study could be less than one-thousandth of the expected demand required at the forecast time, yet the average growth factors would predict only small increases in transport demand.

To overcome these difficulties, methods were designed that provide for differential growth in the various study zones. The first such technique was developed by T. J. Fratar in 1952 in conjunction with a traffic survey in Cleveland, Ohio. Subsequently, modified versions of this technique have been proposed and used. Of these, the more well known are the Detroit method (5) and the average factor method.

Growth factor techniques have several severe handicaps. The most significant is that they ignore the distance between zones as well as the cost, congestion, speed, etc., of the routes over which the interzonal flows must travel. The end result is that it is as attractive for units of movement to move between two zones that are adjacent and directly connected as it is for them to move between zones at opposite extremities of the system which can be reached only by slow, costly, and lengthy paths. For some simple and geographically small systems, this characteristic will not be of major significance, but most transport systems do not fall within this category.

A second major handicap is that no provision has been made to consider bidirectional flow. Interzonal flows are estimated as a total flow without regard to direction. This makes it impossible to study the effect of peak system loads on the network because, for most transport systems, the directed peak loads are not in balance. Quite frequently, in complex transport systems, the relative magnitudes of the directed interzonal flows will differ by several orders.

Finally, the computer requirement for processing a moderate-size system (500 zones) is in the microsecond clock time, 32,000-word storage class. Even with machines such as the IBM 7090, the processing time will be approximately 10 minutes for one such forecast. To forecast large networks, such as those that would be required for metropolitan areas, approximately 10 to 20 times as many zones must be considered. Because the amount of high-speed storage required and the computing time used will increase as the square of the number of zones in the system, it is apparent that large systems cannot be processed by these techniques.

At present, the most widely used forecasting techniques are extensions of the basic gravity model. Gravity model techniques overcome the major objections of the growth factor model techniques first, by including friction factors between zone centroids to discourage undesirable trips, and second, by providing a mathematical structure that treats that interzonal flow as two separate directed flows.

Initially, the gravity model postulated that the transport flow desire between any two zones in a transport system is directly proportional to the product of the population in each zone and inversely proportional to some power of the distance between the centroids of the zones. In use, however, modifications were made on this basic model so that it would more closely approximate observed, real world, network performance. Friction factors were incorporated into the model whereby the flow between zones is governed, in part, by empirically determined functions of time, distance, etc.

The primary weakness of gravity model techniques, which is incidently a weakness of all existing forecasting techniques, is that the effect of network capacity on the forecasted flows is not considered. This problem has been partially overcome in several instances by the inclusion of an assignment procedure in the forecasting model. These techniques begin by assigning the forecasted interzonal flows to the transport network. Then, based on the resulting network congestion in the loaded network, new interzonal friction factors are determined. These friction factors are fed back to the forecasting model and a new forecast of interzonal flow is obtained. This cycle is repeated until stable interzonal forecasts are achieved.

A second weakness is that the effect of socio-economic factors on the system's transport requirements is not included in the gravity equations. This is not an inherent fault of the model, but stems from lack of information as to how these factors influence the transport requirement for each zone and socio-economic group. If these relationships could be determined, they can be incorporated into the friction factors of the gravity model equation, and will thereby greatly increase the reliability and accuracy of the forecasted interzonal movements.

A final fault is that the forecasted interzonal movements do not necessarily yield the required number of trip ends at each destination zone. This indicates that the interzonal friction factors were not in consonance with the actual impedance to flow between those zones. To overcome this problem, an iterative approach is used wherein the friction factors are modified between successive iteration passes. Convergence is generally achieved after two or three passes.

The opportunity model, sometimes called the Chicago method, was developed by the Chicago Area Transportation Study group (CATS) and is based on the following premise: Total travel time from each origin zone to all destination zones is minimized, subject to the condition that every destination zone has a stated probability of being accepted as a trip end if it is considered. To simplify the model, these probabilities are assumed to be constant.

The premise has been restated as follows: A trip prefers to remain as short as possible, but its behavior at any destination node in the network is determined by the probability of it ending at that node. If the nearest destination node is an unacceptable trip end, the trip must consider the next nearest destination node, and if that is unacceptable, consider the next nearest, and so on.

This technique has essentially the same limitations and disadvantages as gravity model techniques have. In use by CATS, however, it has achieved good success. When compared with seven different gravity model forecasts in a test forecast study, the Chicago method did a better job of forecasting known interzonal movements.

The multiple regression model, developed by the California Division of Highways, approaches the forecasting task in a manner considerably different than the approaches used by the previously mentioned models. This model consists of a system of equations, each of the same form, with coefficients determined from empirical transport performance data. After the coefficients have been determined, a forecast of current interzonal flows is made and compared to actual observed flows. If they do not agree, the form of the equations is changed, either by adding more or higher order terms, until the current forecast agrees with the observed current interzonal flows.

The most serious limitations of the multiple regression forecasting model are computation time and reliability of equation form. Because of the complexity of the regression technique, its computation time will be between 10 and 20 times as great as that required for any of the previously mentioned techniques. Because the time required for computing the regression coefficients increases as the square of the number of zones in the network, the upper limit to network size with existing computer techniques is in the neighborhood of 1,000 zones.

The question of validity of the equation form results from the nature of the technique. Equations that fit the survey year data may predict lower or in some cases negative interzonal flows for the forecast year. This can be caused by negative regression coefficients for the "growth" characteristics, and may occur, even when all characteristics are positive, for any of the following reasons: The variables associated with the various characteristics are highly correlated, the zone variables are not descriptive of the zone, or an incorrect equation form has been used.

On the positive side, in several different network studies, comparison forecasts with the gravity and growth factor techniques were made, and the multiple regression technique yielded the best forecasts in each instance.

In summary, of the existing techniques, the gravity and the opportunity models offer the most promise for future research in the transport forecasting field. These techniques entail relatively simple computational procedures and provide for considerable flexibility in choice of equation form. Where research could do the most good, is the area of developing relations between the many variables affecting the desirability or probability of a trip transfer occurring between pairs of zones.

Transport Assignment

The transport assignment phase of the transport planning task has received considerably more attention than has the transport forecasting phase. In general, transport assignment consists of loading a proposed network and then determining how the loaded network performs. Most assignment techniques begin with the allocation of forecasted interzonal flows to the interconnecting links of a proposed transport network. When this is complete, system performance is checked with the planned goals of the system. Measures such as average travel time per unit of distance, cost per unit of movement, and density of flow per network link can be evaluated and used to locate regions within the system where network design improvements are required if the proposed system goals are to be achieved.

In accomplishing the preceding, the usual assignment criterion is to assign interzonal flow such that total network operating cost will be minimized. Unfortunately, most existing techniques fall far short of accomplishing this goal. The underlying problem is that all techniques employ some form of "shortest route" algorithm wherein the interzonal flows are assigned to a supposed shortest or best route interconnecting each zone pair.

It is evident that the determination of which routes are the "shortest" or "best" within a complex network is not a simple task. The best path between two zones when the network is moderately loaded does not necessarily remain the best path as network loading increases. Assignments based on network link costs for any predetermined value of link flow will, in general, be incorrect for other values of link flow.

To date, the mathematical models proposed for the assignment task are very limited in number. Most techniques are digital in nature and can be grouped into two groups: shortest route tree techniques and matrix techniques. All are based on the doubtful assumption that link loading does not affect the selection of route for a given interzonal flow. Only three investigators have attempted to modify these basic techniques to circumvent this problem.

Within the shortest route tree techniques, there are several basic digital algorithms as well as a couple of analog methods. The analog methods however, are not applicable to transport planning and are not discussed. In general, shortest route tree techniques determine the "shortest" or "best" route tree (distance, time, cost, or other arbitrary parameter or combinations thereof may be used as the "shortest" criterion) from any single node to all other nodes in a network. The shortest route tree is defined as the set of links and nodes, connected such that no loops occur, that constitute the least cost paths from the origin or starting node to all of the other nodes in the network. In networks where more than one path between a pair of nodes has the same least cost, only one of these paths is included in the shortest route tree.

The first algorithm for determination of the shortest route tree, suggested by Dantzig (6), is based on the simplex method, and is very slow and cumbersome to apply. It is generally used only for simple hand solutions. The techniques of Minty (24) and Moore (22) are considerably more efficient, and are suitable for computer solution. Both Minty and Moore have formulated their algorithms so that the links of the networks are directional. Because very few practical transport networks are symmetrical, and because each link in these networks can usually be traversed only in one direction, directional links are an essential part of any transport network. Of the two, Moore's algorithm is less time consuming and easier to implement, a fact that makes it attractive for large complex networks.

The problem of simultaneously finding the shortest route between every pair of nodes in a network was first solved by Shimbel (29). Subsequently, Bellman (2) formulated a matrix method algorithm for obtaining solutions according to Shimbel's technique. Although Bellman's method is applicable to the solution of large transport network assignment problems, it is not feasible to use, because the amount of computer memory and computation time required for the determination of all minimum paths in a network is too great. To date, those who have developed computer programs for the transport assignment task have used techniques based on the repeated application of Moore's algorithm.

To present a complete discussion of each transport assignment program presently in use would be too voluminous for this paper. Instead, Table 1 gives the relative capabilities and limitations of the various techniques. Because some of the programs listed include transport forecasting techniques as a part of the transport assignment package, they are so identified. Detailed discussions of each listed program are found in a report by Mertz (7).

Most of the existing techniques are limited as to size of network that can be studied (Table 1). Even more disconcerting is the fact that only three of the programs are published, and of these only two are programed for computers available in the United States. Of the two, only the Washington, D. C., program is of adequate size to study complex networks.

The most serious limitation of these existing programs is treatment of the capacity-restraint problem. Only three of the systems consider the problem at all, and then only in a limited sense.

The Detroit Arterial system, after the completion of the assignment cycle, determines all paths that are loaded in excess of their maximum capacity. It then redistributes the loads for these paths to several of the next best paths in proportion to their ability to receive them.

The Chicago system treats capacity restraint on a zone-by-zone basis. After the trips from one zone are distributed to their destinations in accordance with the appropriate minimum path tree, new link impedances are computed from the partially loaded network. From these new link impedances, the minimum path tree for another zone is determined, and the trips from this zone are assigned accordingly. This cycle is repeated until all trips have been assigned.

Traffic Research Corporation, Ltd., uses a relaxation technique whereby trips are assigned to as many as four different paths between each pair of zones, and are assigned in a manner such that the loaded path costs are equal for all paths of each zone pair. To accomplish this, the first step is to load the minimum path trees based on their zero flow link impedances. Next, new link impedances are computed from the link flow volumes of the loaded network, and new minimum path trees are computed therefrom. Finally, the network is reloaded by assigning each interzonal flow to its original and new minimum path in inverse proportion to its respective path costs. This process is repeated until the system stabilizes with up to four paths allowed for each pair of zones.

TABLE 1

TRANSPORT ASSIGNMENT PROGRAMS

System or Organization	Computer	Number of Zone Centroids	Number of Nodes	Number of Links	Number of Links from Any Zone	Capacity Restraint	Forecasting Option	Published	Computer Time for Typical Job (hr)	Corresponding Typical Job
Chicago system	32K IBM 704	700	4,095	14,000	Unlimited	Optional	Opportunity	No	5	650 centroids, 4,000 nodes, 13,000 links, no capacity restraint
Washington, D. C., and Minnesota system	IBM 704[a]	Any node a centroid	4,000		4	No	Growth factor	Yes	5	500 centroids, 3,500 nodes
Detroit Expressway	IBM 650 with 2 tape units and 1 Ramac	400	75	425		No		No	20	30 nodes, 20,000 interzonal movements
Detroit Arterial	IBM 704	Any node a centroid	999			Yes		No	4	Maximum network size
Service Bureau Corp.	8K IBM 704	Up to 300 nodes a centroid	1,350		7	No	Gravity	No	6	250 centroids, 1,300 nodes
Traffic Research Corp., Ltd.	IBM 650[b] with tapes	100	1,800			Yes	Gravity	No	60	100 centroids, 250 nodes
State of Connecticut	Remington Rand File Model 1	Up to 145 nodes a centroid	10,000			No	Gravity	No	39	120 centroids, 1,400 nodes, 2,500 links
California	IBM 650[c]	Any node a centroid	699	1,000		No		No	125	699 nodes, 210 of which centroids
Missouri	IBM 650	200	791			No		Yes	120	270 centroids, 1,300 nodes
Road Research Lab.	Ferranti Pegasus	44		255		No		Yes	1	100 links

[a] Reprogramed to run on IBM 7090 with use of IBM 704-7090 compatibility program.
[b] Being reprogramed for IBM 7070.
[c] Being reprogramed for IBM 8K 704.

Of these three systems, only the Traffic Research Corporation system approaches a realistic solution. Granted that the other systems are better than ignoring capacity altogether, they leave much to be desired. Even the TRC system has several drawbacks, the most significant of which is the number of times the minimum path trees must be constructed before system stability is achieved. Tests of convergence made by Irwin et al. (14), indicate that approximately ten assignment cycles are required before the system stabilizes. For large networks, this could be very time consuming. Another questionable aspect of the technique is the limitation to four paths for any interzonal flow. In many transport networks it is entirely possible that ten, twenty or even a hundred alternate routes of near equal cost might exist. This is particularly true for flows between network zones that are geographically far apart. That all feasible alternate routes should be considered is self evident, yet they cannot be, for their inclusion would certainly increase, to an impractical extent, the number of assignment cycles required to achieve system stability.

In spite of these limitations, the TRC system is far more sophisticated than any of the others in Table 1. Some other interesting features of this system are forecasting feedback, land use feedback, and consideration of transport mode capacity restraints. The forecasting feedback option allows the forecast interzonal flows to be changed as a consequence of network congestion. In this manner, the forecasted flows will be consistent with the ultimate path costs of the loaded network. Similarly, units of movement may switch from one transport mode to another as a function of the loaded network cost for each transport mode. The land use feedback option allows the planner to determine the compatibility of the proposed transport plans and the land use plan. If ultimate land use realization is known as a function of freedom of movement within the system, changes in the land use plan, due to network performance, can be estimated and new assignments made.

In the next section of this paper a new approach to the transport assignment task is offered. It is an analytical matrix technique, and is designed to overcome the network assignment and analysis difficulties, inherent in existing assignment techniques, that are due to inadequate consideration of network link capacity-flow relations.

THE CAPACITY-RESTRAINT ALGORITHM

The primary feature of the capacity-restraint algorithm is its provision for arbitrary linear or nonlinear network link cost functions. In addition, it allows an entire transport network to be treated at once, with any number of permissible paths for each interzonal movement. Finally, because the technique is analytic, it will yield the solution for all paths of each interzonal movement, simultaneously. Provision has been made for limiting the permissible paths for any origin-destination pair to those meeting various desirability criteria, such as complexity, cost, and distance. Also, if network capacity is inadequate to accommodate any given interzonal movement, the saturated network links, responsible for each flow constriction, are determined and identified.

There are two basic options available with this assignment algorithm: (a) loading of the network to obtain optimum total system figure of merit, and (b) loading of the network so that for each pair of centroids the corresponding path figures of merit are equalized. These figures of merit can be any desired measure of network and path performance. The first option is appropriate for networks in which the route assignment for each unit of movement can be enforced. In networks in which each unit of movement is a free agent, the ultimate network assignment will be more realistically given by the second option. This latter option is applicable to such networks as those for highway transportation.

The following are several phases applicable to both assignment options:

1. Construction of the "connection matrix."
2. Expansion of the matrix as a set of modified determinants.
3. Determination of the "link cost" functions.
4. Determination of the "path cost" functions, elimination of undesirable paths, and making of preliminary checks for inadequate link capacity.

5. Establishment of the system of equations describing the network.
6. Elimination of path flow variables that must be zero to satisfy the boundary conditions; determination and identification of any links that are supersaturated.
7. If no links are supersaturated, solution of the system of equations.

All seven phases of this algorithm are extensions of existing mathematical techniques. Before the detailed procedures for each phase are presented, a brief description of the underlying theory and purpose of each phase is given. Perhaps the most interesting aspects of the algorithm are those of phases 1 and 2. The procedure used here was first described by Yoeli (30). To the author's knowledge, no concise, and simple explanation of this procedure has been published. Briefly, the technique determines all possible paths between each pair of nodes in a complex network in a manner such that all redundant paths are eliminated, and only paths that do not double back on themselves remain.

Link cost functions are established during phase 3. For each link, these are generally nonlinear functions of link loading, and represent the cost, time, safety, distance, etc., or combinations thereof, of using that link. Linear functions and constants, however, should be used whenever possible because their use reduces the overall complexity of the system of equations to be solved during phase 7.

Phase 4 of the algorithm serves three purposes: first, to determine the path cost functions from the link cost functions; then, to find links with inadequate link capacity; and last, to provide to the investigator the opportunity to eliminate paths from the network that are not satisfactory to him. The elimination of paths at this stage in the algorithm will greatly reduce the complexity of the equations to be solved, and thereby substantially reduce the required computer time.

The fifth phase consists of setting up a system of network describing nonliner equations. This set of equations will have more unknowns that equations, and will be solved in phase 7 either by the Lagrange method of undetermined multipliers (if total network cost is to be minimized), or, after writing equal path cost equations, as simultaneous nonlinear equations (if origin-destination flow path costs are to be equalized).

Phase 6 produces a preliminary interzonal flow assignment to determine the existence of any regions of the network wherein the links are saturated so that some of the network's interzonal flows are not accommodated. In such instances, the saturated links are tabulated and the analysis is terminated until the network is modified to provide the required additional capacity. In addition, if the first assignment option is selected, the zero flow and saturation flow boundary conditions are introduced during this phase. It is necessary to eliminate any variable from the system of equations for which the first partial derivative of the total network cost equation with respect to that variable does not vanish for some value within that variable's allowed range of variation.

The final phase of the algorithm consists of the solution of the system of equations. Because of the complexity of the equation, this cannot be done by analytical techniques. The method of solution proposed is that of successive evaluation and minimization of error, a technique that has been programed for the IBM 7090 and has been used successfully for problems of this type. A user's manual is available (33).

Following is a detailed description of the procedure for each phase of the capacity-restraint algorithm. These descriptions are intended to provide a base from which computer programs can be written.

Phase 1—Construction of Connection Matrix

To construct the connection matrix, the network must be represented as a lattice-weighted, directed, linear graph. This is accomplished by splitting the area served by the network into subareas of arbitrary size and configuration. Each subarea, referred to as a zone, is assigned a mnemonic which is associated with its mathematical centroid. These mathematical centroids, each of which is located at its corresponding zone's center of gravity are determined from the zone characteristics. After all zone centroids have been located, the network graph is constructed by interconnecting these centroids with the transport links of the proposed network. Finally, each link is assigned a mnemonic for each direction of allowed traversal. During phase 3, these mnemonics are replaced by mathematical functions of link flow.

Figure 2 is an example of a lattice-weighted, directed, linear graph for a transport network of 10 zones, interconnected by 21 directed links. The arrows on the graph indicate the allowed directions of movement over each link, and the associated mnemonics identify each directed link's use function. The network nodes are identified by mnemonics A through J.

The connection matrix for a network of N nodes will be a square matrix of order N, and will be constructed with the assistance of the lattice-weighted linear graph in the following manner. First, the rows and columns of the matrix are labeled with the network's node mnemonics. The rows of the matrix are arbitrarily chosen to represent origin nodes, and the columns represent destination nodes. Because each network zone can be both an origin and a destination zone, it is evident that the network node mnemonics of each zone centroid must appear as the label of one row and one column in the matrix. Next, the values for each cell of the matrix are entered directly from the lattice graph. This is accomplished by recording the link mnemonic for each origin-destination pair in a cell of the matrix located at the intersection of the link's origin row and destination column. If a particular destination cannot be reached directly from a given origin (because no direct link exists), the matrix cell corresponding to that origin-destination pair receives a zero entry. Matrix cells for which the origin and destination zones are the same receive an entry of one.

Using this procedure, the connection matrix for Figure 2 is shown in Figure 3. In the situation where each link is bidirectional and has the same use function in each direction, the entries $f_{i,j}$ and $f_{j,i}$ will be identical, and the matrix will be symmetrical. For most complex networks, however, this matrix will be nonsymmetrical.

Phase 2—Expansion of Matrix as a Set of Modified Determinants

To determine the routes of all paths from every origin to every destination in a network, the connection matrix is split into N^2 submatrices. When each submatrix is expanded as a determinant, an expression results which, after changing all arithmetic signs to a plus and simplifying by Boolean techniques, indicates all possible paths between a given origin and a given destination.[1] The submatrix for a particular origin-destination pair is obtained from the connection matrix by eliminating the matrix column corresponding to the given origin and the matrix row corresponding to the given destination. The elimination of the column corresponding to the origin centroid pre-

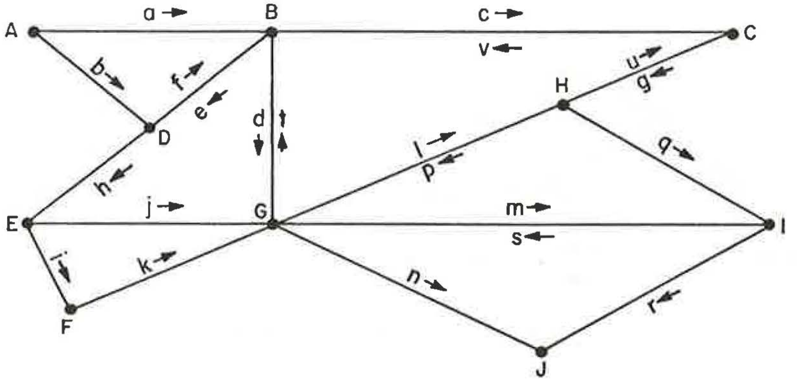


Figure 2. Lattice-weighted, directed, linear graph.

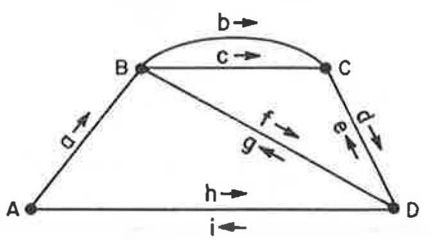
[1] In Boolean algebra techniques, the expression a + b is read a "or" b; thus if two points in a network are connected by two paths, for example a and b, the admissible paths between these two points are given by the expression a + b.

	A	B	C	D	E	F	G	H	I	J
A	1	a	0	b	0	0	0	0	0	0
B	0	1	c	e	0	0	d	0	0	0
C	0	v	1	0	0	0	0	g	0	0
D	0	f	0	1	h	0	0	0	0	0
E	0	0	0	0	1	i	j	0	0	0
F	0	0	0	0	0	1	k	0	0	0
G	0	t	0	0	0	0	1	ℓ	m	n
H	0	0	u	0	0	0	p	1	q	0
I	0	0	0	0	0	0	s	0	1	r
J	0	0	0	0	0	0	0	0	0	1

Figure 3. Connection matrix.

cludes that centroid from being used in any path as an intermediate destination. Likewide, the elimination of the row corresponding to the destination centroid prevents that centroid from occurring in any path as an intermediate origin.

After the Boolean expression for a submatrix has been simplified, it is rewritten as a first-order expression.[2] In this expression, each concatenation of mnemonics represents one possible path for the given origin-destination zone pair. Each such path can be traced on the network lattice graph by marking each link whose mnemonic appears in that concatenation. Because of the techniques used in expanding the submatrix and in the subsequent Boolean simplification, the sequence of the mnemonics in each concatenation does not necessarily correspond to the sequence in which the corresponding links are encountered when moving between the given origin and destination. The terms are, however, all-inclusive and nonredundant. That is, all links of a path are represented once and only once in that path's concatenation. Likewise, the simplified Boolean submatrix expansion is all-inclusive and nonredundant. Any redundant concatenations occurring as a result of the determinant expansion (i.e., concatenations that do not include the mnemonics for enough links to form a contiguous path from the



	A	B	C	D
A	1	a	0	h
B	0	1	b+c	f
C	0	0	1	d
D	i	g	e	1

Figure 4. Simplified network.

[2] A first-order expression is one that contains only concatenated link mnemonic terms separated by the Boolean "or." Each such term represents a path between the origin and destination zones of the submatrix from which it was obtained. For example, if the path between two zones consists of hypothetical links c, d, and e, the term representing this path could be written as (c)(d)(e), c•d•e, or simply cde. In any case, the concatenation indicates that the path consists of links c "and" d "and" e. It should be noted that the sequence of the link mnemonics within the path concatenation is not important.

origin zone to the destination zone, and concatenations that include some mnemonics more than once) are eliminated by the Boolean simplification procedure.

To illustrate this procedure, the simplified network shown with its connection matrix in Figure 4 is given. For example, to determine all paths from origin B to destination C, column B and row C are eliminated from the connection matrix. The resulting submatrix is expanded and simplified as shown in Figure 5.

As Figure 5 shows three paths exist from origin B to destination C: (a) the path consisting of link b; (b) the path consisting of link c; and (c) the path consisting of links f and e. Terms hbi and hci, both of which represent noncontiguous and therefore improper paths, were eliminated by the Boolean simplification of the expression. The Boolean algebra rule used to eliminate these terms was (1 + X) = 1. In the preceding case, the hbi term is eliminated by the b term, whereas the hci term is eliminated by the c term; for example, b + hbi = b(1 + hi) = b.

In a like manner, all paths for each remaining origin-destination zone pair can be found. Table 2 gives all paths between all origins and all destinations for the network shown in Figure 4.

Phase 3—Determination of Link Cost Functions

Because the link cost functions should include all factors influencing the performance of each network, the form of these functions may vary considerably for different transport fluxes. For some, the functions may be linear relations, whereas for others, highly complex mathematical relations might be required. Considerable research by economists, sociologists, engineers, and land use planners is necessary before accurate link cost functions can be formulated. For certain transport fluxes, however, these functions can be approximated to a degree sufficient for useful network performance studies. Because it is beyond the scope of this paper to determine or argue

$$\begin{array}{cccc} & A & C & D \\ A & 1 & 0 & h \\ B & 0 & b+c & f \\ D & i & e & 1 \end{array} \qquad \begin{aligned} P_{BC} &= b + c + fe + h(b+c)i \\ &= b + c + fe + hbi + hci \\ &= b + c + fe \end{aligned}$$

Figure 5. Submatrix and resulting paths from B to C.

TABLE 2

ALL PATHS BETWEEN ALL ORIGIN-DESTINATION PAIRS FOR NETWORK OF FIGURE 4

Origin	Destination	Path
A	B	P_{AB} = a + hg
A	C	P_{AC} = ab + ac + afc + he + hgb + hgc
A	D	P_{AD} = abd + acd + af + h
B	A	P_{BA} = bdi + cdi + fi
B	C	P_{BC} = b + c + fe
B	D	P_{BD} = bd + cd + f
C	A	P_{CA} = di
C	B	P_{CB} = dg + adi
C	D	P_{CD} = d
D	A	P_{DA} = i
D	B	P_{DB} = g + ai
D	C	P_{DC} = e + bg + cg + aib + aic

the merits of various cost function formulations, the following discussion is limited to illustrating two or three proposed functions and their relationship to the capacity-restraint algorithm.

The most widely used link cost function is the constant. As mentioned previously, existing network analysis techniques with few exceptions, use constant link cost functions for all links. The few exceptions use "step" link cost functions, whereby the cost functions change, either at fixed points in the assignment process, or between successive iterative assignments to the network. Step link cost functions are suitable, for some links, in networks for certain of the "service" transport fluxes such as communications, water, and power where cost increases occur at discrete levels of flow.

Another function, more suitable for general use, is the "linear" link cost function. This function allows cost per unit of flow to increase linearly with flow, but it does not prevent inordinately high link loadings from occurring; therefore, link loadings in excess of link capacity can occur in densely loaded networks.

Because neither of the preceding functions is able to prevent flows in excess of link capacity from being assigned to the links, they are not suitable for the general transport link cost function. They can, however, be used to great advantage in the capacity restraint model whenever they closely approximate a particular network link cost function throughout its range of conceivable assigned usage. The advantage in these instances is that the network equations, to be solved at later stages in the algorithm, become simpler as the number of constant or linear link cost functions used is increased.

Two functions that have been suggested as reasonable candidates for the general transport link cost function are the hyperbolic and logarithmic functions. The characteristics of these functions are such that the change in "cost per unit of flow" as flow increases is small for low flow values, but large as saturation flow is approached. For both functions, both incremental "cost per unit of flow" and "total link cost" approach infinity as link saturation flow is approached.

The logarithmic link cost function is given by

$$f_i(m_i) = \log(M_i) - \log(M_i - m_i) + T_i \qquad (1)$$

in which

$f_i(m_i)$ = cost per unit of flow for m_i units of flow;
m_i = number of units of flow using link i;
M_i = saturation flow for link i; and
T_i = zero flow cost for link i.

From Eq. 1, the cost as the limits of possible flow are approached is described by the following:

$$\text{As } m_i \to M_i, \qquad f_i(m_i) \to \infty$$

$$\text{As } m_i \to 0, \qquad f_i(m_i) \to T_i.$$

The logarithmic link cost function is shown in Figure 6.

The hyperbolic link cost function is

$$f_i(m_i) = \tau_i - \frac{M_i(T_i - \tau_i)}{m_i - M_i} \text{ and } \tau_i < T_i \qquad (2)$$

in which

$f_i(m_i)$ = cost per unit flow for m_i units of flow;
m_i = number of units of flow using link i;
M_i = saturation flow for link i;
τ_i = function's cost asymptote; and
T_i = zero flow cost for link i.

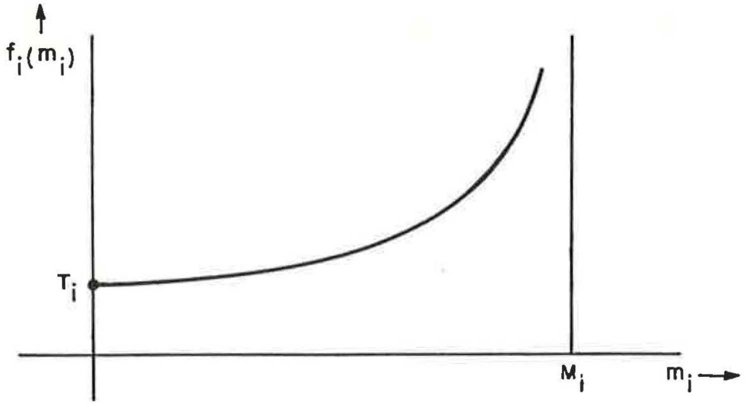


Figure 6. Logarithmic link cost function.

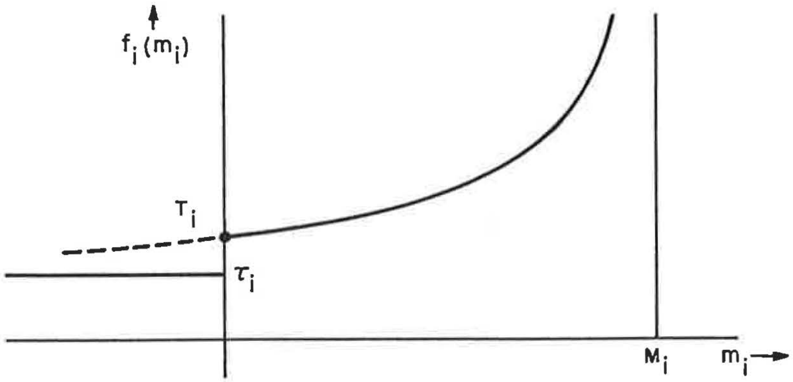


Figure 7. Hyperbolic link cost function.

The function cost asymptote, τ_i, controls the flatness of the function at low values of m_i. As τ_i decreases, the rate of change of $f_i(m_i)$ for low values of m_i increases, and for increasing values of τ_i, the rate of change of $f_i(m_i)$ decreases. The smallest possible rate of change for $f_i(m_i)$ occurs at the upper limit of τ_i, namely T_i. From the hyperbolic link cost function, the link costs as the limits of possible link flow are approached are given by the following:

$$\text{As } m_i \rightarrow M_i, \qquad f_i(m_i) \rightarrow \infty$$

$$\text{As } m_i \rightarrow 0, \qquad f_i(m_i) \rightarrow T_i$$

Figure 7 shows the hyperbolic link cost function. This function is used as the link cost function in the discussion of the remaining phases of the capacity-restraint algorithm. Zero flow cost for any link is assumed to be analogous to the zero flow travel time required to traverse that link and, for simplicity, it is assumed that the cost asymptote for all link cost functions is zero.

In Figure 4, an assignment of hypothetical forecast flows from centroid A to centroids B and D, and from centroid D to centroid B will be accomplished. These three flows will be assumed to be

$$M_{AB} = 25 \tag{3}$$

$$M_{AD} = 100 \tag{4}$$

$$M_{DB} = 450 \tag{5}$$

For the purpose of this example all other interzonal flows will be assumed zero. From Table 2, the paths for the three non-zero flows are

$$P_{AB} = a + hg \tag{6}$$

$$P_{AD} = abd + acd + af + h \tag{7}$$

$$P_{BD} = g + ai \tag{8}$$

From Eqs. 6 through 8, it is observed that link cost functions are required for links a, b, c, d, f, g, h, and i.

Table 3 gives, for each of these links, the assumed values of their zero flow travel time (T_i) and their saturation link flow (M_i). Also given for each link are the hyperbolic link cost function [$f_i(m_i)$], the unit flow cost [$f_i(1)$], and the maximum allowable number of units (max_i) that may be assigned to that link. The maximum flow limit, max_i, is one unit of flow less than M_i.

Phase 4—Determination of Path Cost Functions and Test Links

Phase 4 begins by assigning to each path a mnemonic that will identify that path and its subsequent flow. Once this is accomplished, each path cost function is obtained by adding the link cost functions for every link included in that path. Next, these path cost functions are evaluated for unit flow loading, thereby providing a means for comparing the various paths of any particular origin-destination pair at the zero flow boundary. The flow capacity of each path is then obtained by determining the smallest "maximum link flow" of any link in that path.

TABLE 3

LINK COST FUNCTIONS

Link	M_i	T_i (sec)	$f_i(m_i)$	$f_i(1)$	max_i
a	100	10	$\frac{1,000}{100 - m_a}$	10.10	99
b	200	5	$\frac{1,000}{200 - m_b}$	5.03	199
c	25	5	$\frac{125}{25 - m_c}$	5.21	24
d	200	10	$\frac{2,000}{200 - md}$	10.05	199
f	300	15	$\frac{4,500}{300 - m_f}$	15.05	299
g	500	10	$\frac{5,000}{500 - m_g}$	10.20	499
h	50	10	$\frac{500}{50 - m_h}$	10.20	49
i	500	4	$\frac{2,000}{500 - m_i}$	4.01	499

At this point in the algorithm, tests and conditions are applied to the paths which, if they are not satisfied, will eliminate those paths from further consideration. The number and type of such tests change for each type of network under consideration, and may vary in complexity to whatever extent the investigator is willing to apply logic to the system under study. The elimination of paths at this time will greatly simplify the balance of the analysis, and thereby decrease the required computational time.

The tests and conditions discussed next are not intended to be exhaustive. Many other considerations will become obvious as particular networks are studied.

In communication networks, a reasonable limitation to the number of links in a path is a realistic restriction. The number of links in any path is easily determined from the path equations obtained during phase 2. All paths consisting of more than a predetermined number of links might be eliminated from the list of allowable paths.

From the path use function evaluation, it is possible to limit the assignment to the best N paths between each pair of zones. These best paths are determined by comparing the path unit flow costs. Caution must be exercised in applying this type of restriction because it is entirely possible that when the network is loaded with all interzonal flows, the individual costs for some or all of these best paths will greatly exceed that of the best path that was discarded.

Paths may also be eliminated for reasons such as total cost is too great, distance is too long, or travel time is too great. Such restrictions require detailed knowledge of the link characteristics, but considerable information of this type is already available, because it is required in phase 3 to determine the link use functions. Excessive cost and travel time paths can be rejected at this point in the algorithm only on the basis of unit flow loading, or on approximate equal path cost loading.

As undesirable paths are eliminated from the list of allowable paths for any origin-destination pair, the total capacity of the balance of the paths for that origin-destination pair must be compared to the required interzonal flow over them. If these remaining paths do not have adequate excess capacity, the analysis must stop, and the network must be redesigned. The question of how much excess capacity there should be for a given interzonal flow cannot be determined analytically; after some experience has been gained with the use of this algorithm, however, it should be possible to estimate appropriate values.

Continuing the example of Figure 4, each path was assigned a mnemonic. The path cost functions were determined and evaluated for unit flow, and maximum path flows were obtained (Table 4).

Phase 5—Establishment of System of Equations

The system of equations to be solved during phase 7 is obtained from the network link use functions, the interzonal flow requirements, the network total cost equation, and a set of relations between path and link flow. Many of these equations are linear relations, and can be used to eliminate variables from the system of equations through simple substitution techniques. For solution by conventional manual techniques, much

TABLE 4

PATH COST FUNCTIONS

O-D Pair	Path	Path Mnemonic	Path Cost Function	Unit Flow Cost	Path Flow Capacity
A-D	abd	1	$\frac{1,000}{100 - m_a} + \frac{1,000}{200 - m_b} + \frac{2,000}{200 - m_d}$	25.18	99
	acd	2	$\frac{1,000}{100 - m_a} + \frac{125}{25 - m_c} + \frac{2,000}{200 - m_d}$	25.36	24
	af	3	$\frac{1,000}{100 - m_a} + \frac{4,500}{200 - m_f}$	25.15	99
	h	4	$\frac{500}{50 - m_h}$	10.20	49
A-B	a	5	$\frac{1,000}{100 - m_a}$	10.10	99
	hg	6	$\frac{500}{50 - m_h} + \frac{5,000}{500 - m_g}$	20.40	49
D-B	g	7	$\frac{5,000}{500 - m_g}$	10.20	499
	ai	8	$\frac{1,000}{100 - m_a} + \frac{2,000}{500 - m_i}$	14.11	99

simplification is achieved if this is done. Substitution is not advisable if computer techniques are to be employed for the algorithm, because the required computer program logic for accomplishing the substitution would be very complex, and the elimination of these variables would not greatly reduce the computer solution time.

The system of equations consists of three basic equation types. First, the relationship between the link cost functions and total network use cost yields

$$C_T = \sum_i m_i f_i(m_i) \tag{9}$$

in which

C_T = total network use cost;
i = index of summation (taking on mnemonic for each link in analysis);
m_i = number of units of movement assigned to link i; and
$f_i(m_i)$ = link cost function for link i evaluated for m_i units of movement.

An alternate form of this equation, found by summing the individual origin-destination cost equations, is

$$C_T = \sum_{p,q} C_{p,q} = \sum_{p,q} \sum_j N_j g_j \tag{10}$$

in which

C_T = total network use cost;
p, q = index of summation (taking on mnemonics for each origin-destination pair in analysis);
$C_{p,q}$ = total cost of flow from origin p to destination q by all paths interconnecting them;
j = index of summation (taking on mnemonics for each path of a given origin-destination pair);
N_j = number of units of movement using path j; and
g_j = path cost function for path j found by summing link cost functions of all links comprising that path.

Because in most network studies it is important to individually evaluate the total cost ($C_{p,q}$) for each interzonal flow of the network, this second form of the total cost equation is more convenient.

The next group of equations are restraining equations, and are formulated by setting each required interzonal flow equal to the sum of all of the path flows which could accommodate it. There will be as many equations of this type as there are interzonal flows in the network. These equations are of the form:

$$M_{p,q} = \sum_j N_j \tag{11}$$

in which

$M_{p,q}$ = required interzonal flow from origin p to destination q;
j = index of summation (taking on mnemonic for each path of a given origin-destination pair); and
N_j = number of units of movement using path j.

The last group of equations are also constraining equations, and are formulated by setting the link flow variable of each link equal to the sum of all path flows that traverse that link. There will be one equation of this type for every network link. These equations are of the form:

$$m_i = \sum_k N_k \tag{12}$$

in which

m_i = number of units of movement assigned to link i;
k = index of summation (taking on mnemonic for each path using link i); and
N_k = number of units of movement using path k.

Returning to Figure 4, Eq. 9 becomes

$$C_T = m_a f_a(m_a) + m_b f_b(m_b) + m_c f_c(m_c) + m_d f_d(m_d) +$$

$$m_f f_f(m_f) + m_g f_g(m_g) + m_h f_h(m_h) + m_i f_i(m_i) \tag{13}$$

Eq. 10 and its individual total interzonal cost functions are written, with the help of Table 4, as follows:

$$C_T = C_{AB} + C_{AD} + C_{DB} \tag{14}$$

$$C_{AB} = N_5\left[f_a(m_a)\right] + N_6\left[f_h(m_h) + f_g(m_g)\right] \tag{15}$$

$$\begin{aligned} C_{AD} = {} & N_1\left[f_a(m_a) + f_b(m_b) + f_d(m_d)\right] + \\ & N_2\left[f_a(m_a) + f_c(m_c) + f_d(m_d)\right] + \\ & N_3\left[f_a(m_a) + f_f(m_f)\right] + N_4\left[f_h(m_h)\right] \end{aligned} \tag{16}$$

$$C_{DB} = N_7\left[f_g(m_g)\right] + N_8\left[f_i(m_i) + f_a(m_a)\right] \tag{17}$$

Next, from Eq. 11, using Table 4 and Eqs. 3, 4, and 5, three constraining equations are

$M_{AB} = N_5 + N_6$ (18) $\quad$ $M_{AD} = N_1 + N_2 + N_3 + N_4$ (19) $\quad$ $M_{DB} = N_7 + N_8$ (20)

Last, from Eq. 12, using Table 4, eight more constraining equations are

$$m_a = N_1 + N_2 + N_3 + N_5 + N_8 \tag{21}$$

$$m_b = N_1 \tag{22}$$

$$m_c = N_2 \tag{23}$$

$$m_d = N_1 + N_2 \tag{24}$$

$$m_f = N_3 \tag{25}$$

$$m_g = N_6 + N_7 \tag{26}$$

$$m_h = N_4 + N_6 \tag{27}$$

$$m_i = N_8 \tag{28}$$

Eqs. 13 through 28 now define the entire network, but inasmuch as there are more unknowns than equations, an explicit solution of them is not possible. If the first assignment option is chosen, this situation will be handled by the Lagrange method of undetermined multipliers, and a solution will be found that minimizes the total network cost function, C_T. If the second assignment option is chosen, additional equations will be written from the path cost functions (g_j) of Eqs. 15 through 17. In this case, each path cost function in each of these equations (the bracketed expressions in each equation are the path cost functions) will be set equal to an undetermined constant. These constants will be the same for all path cost functions of a given interzonal cost equation, but different for those of different interzonal cost equations. When this is accomplished, there will be as many unknowns as equations and an explicit solution for them will be possible. This solution will load the network so that, for each pair of network zones, the path costs per unit of flow of each of its paths will be equalized.

Before these equations can be solved, boundary conditions must be applied and the network must be tested for supersaturation.

Phase 6—Testing for Supersaturated Links and Applying Boundary Conditions

At this point in the algorithm, the network is loaded by assigning all flow between each pair of origin and destination zones to the best paths (least cost for unit flow paths) between them. From the required interzonal flows, the list of permissible paths (determined in phase 3), and the list of unit flow path costs (determined in phase 4), the network assignment is accomplished as follows:

1. All interzonal flows for origin-destination pairs having only one admissible path are assigned.
2. Interzonal flows for which there is more than one admissible path, are assigned in sequence to increasing number of admissible paths.
3. The interzonal flow for each multi-path origin-destination pair is assigned to its best zero flow path until that path is saturated. Any remaining flow is assigned to the next best paths, in order of their desirability, until either the entire flow is assigned or all allowable paths for that flow are saturated.
4. After each interzonal flow has been assigned, the cumulative link loadings must be determined. Whenever a link becomes saturated, all paths, for interzonal flows not yet assigned, which use that link are temporarily eliminated from the list of admissible paths. If any interzonal flow cannot be accommodated because all of its paths contain at least one saturated link, these saturated links and that interzonal flow are noted, and the additional capacity required to accommodate the flow is determined.

When the network capacity is inadequate to satisfy all interzonal flows, the analysis terminates at the completion of the preceding assignment. The tabulation of saturated links and unsatisfied interzonal flows can be used either to redesign the transport network, or to indicate where changes must be made in the land use plan.

If the assignment is successful, and the second assignment option has been selected, phase 6 is complete. The next step is to obtain the solution of the system of equations by the procedure in phase 7.

When the first assignment option is chosen, phase 6 continues. The next step is to modify the system of equations for the network such that the conditions imposed by the Lagrange technique are satisfied. Because it is desired to find variable values that will minimize total network cost, and because the necessary condition for the existence of such an extremum for a differentiable function is the vanishing of the first partial derivatives of the function with respect to its independent variables (29), it is necessary to determine the behavior of each first partial derivative of the cost function throughout the admissible range for its corresponding variable. Each first partial derivative of the cost function must vanish for an admissible value of its corresponding variable.

The differentiation and solution of complex functions is a time-consuming and difficult task to accomplish with a digital computer, especially when the form of the equation can change from application to application. Because the nature of this algorithm is such that a digital computer must be used, and because each network studied will undoubtedly have equations of a form peculiar to itself, it is not feasible to use classical methods to obtain the first partial derivatives of the network cost equation. Instead, a procedure that requires only the repeated evaluation of the cost equation and a comparison of the resulting values has been devised.

As background, the total network cost equation consists of the arithmetic sum of the products of link flow and the corresponding link cost functions. The link cost functions are monotonic increasing functions of the link flow variables. Each link flow variable is a dependent variable, and is dependent on the path flow variables of all paths using that link. All path flow variables that received assigned values during the preliminary assignment are also dependent variables, and are dependent on the network's interzonal flows and link saturation values. The path flow variables that received no assigned flow during the preliminary assignment are the independent variables of the system of equations, and are, therefore, the variables for which the first partial derivatives of the cost equation must vanish. Any such variable, for which the corresponding partial derivative does not vanish for an admissible value of that variable, must be set equal to zero and eliminated from the system of equations. This implies that the minimum total network cost is achieved when these variables are zero, and therefore when no flow is assigned to their corresponding paths. All link flow variables that are functions only of eliminated path flow variables must also be dropped from the equations. Finally, because the link cost functions are monotonic increasing functions of flow, their evaluation as saturation link flow is approached yields successively higher values of cost per unit of flow.

Keeping the preceding facts in mind, the path flow variable test procedure is as follows:

1. The total network cost function is evaluated for the assigned flows that were determined by the preliminary assignment of this phase of the algorithm.
2. These assigned path flows are decremented one at a time, by one unit of flow, and a new total cost is computed for each case.
3. These new total costs are grouped by origin-destination pairs.
4. For each zone pair group, the path flow variable which when decremented yields the lowest total network cost is determined. These path flow variables are used as reference variables for the balance of this test procedure, and are used at their decremented value whenever they are used as reference.
5. By zone pair group, the path flow variables that received no initial assignment are incremented one at a time by one unit of flow. Using the appropriate reference variable, the new total network cost is computed. (At this point in the procedure, situations may occur where some link loadings are incremented to their saturation boundary limit. Whenever this happens, one of the other path flow variables using that link must be decremented by one unit, and a corresponding path flow variable not using that link must be incremented by one unit of flow. This must be done in a manner such that the decremented variable produces the maximum decrease in total network cost, and the incremented variable produces the minimum increase in total network cost.)
6. The total network costs computed in step 5 are now compared to the total network cost computed in step 1. If the new cost is higher than the original cost, the corresponding path flow variable will be zero and it is eliminated from the system of equations. If the new cost is less than the old cost, the corresponding path flow variable remains in the system of equations. Variables passing this test, and thereby remaining in the analysis, will not necessarily be assigned non-zero values in the final assignment. Because units of movement are integral entities, the final path flows will be rounded off to their nearest integral values, thus it is entirely possible that some paths will be assigned zero flow.

The following is offered as proof that the preceding procedure eliminates all variables for which the corresponding first partial derivatives of total network cost do not vanish for an admissible value of those variables, and furthermore that it eliminates no other variables. As outlined, the procedure determines the differential of the total cost function at each independent variable's zero flow boundary. If the increment used in determining this differential cost is allowed to approach zero, the true derivative at the boundary is obtained. Now, if the partial derivative of the total cost function is negative at a variable's zero flow boundary, it is only necessary to show that this derivative becomes positive somewhere in the admissible range of that variable in order to establish the existence of a minimum extremal value of the function for that variable. Because any partial derivative of the function is always positive at the saturation flow boundary of its corresponding variable (because total cost approaches infinity as a variable approaches its saturation value), it follows that any variable that has a negative derivative at the zero flow boundary has an admissible value for which the corresponding first partial derivative will vanish.

To complete this proof, it is necessary to show that if, at any variable's zero flow boundary, the partial derivative of the total cost equation is positive, it will remain positive throughout that variable's admissible range. Now, because the path use functions are monotonic increasing functions of link loading, and because the derivatives of the individual link use functions are also monotonic increasing, it follows that the derivatives of the path's use functions are monotonic increasing. Because the total cost equation is given by the sum, for all network paths, of the products of path flow and the corresponding path use function, it is apparent that the partial derivative of this equation with respect to any path flow variable will be monotonic and will be arithmetically smallest at the zero flow boundary of that variable. Therefore, if the differential of total cost with respect to a particular variable is positive at that variable's zero flow boundary, it will remain positive throughout that variable's admissible range.

To illustrate this phase of the algorithm, the network of Figure 4 is considered. From the required interzonal flows of Eqs. 3, 4, and 5, and the unit flow path costs of Table 4, a preliminary assignment is made in accordance with the rules for part 1 of this phase. The resulting flows are as follows:

$$\begin{array}{ll} N_1 = 0 & N_5 = 25 \\ N_2 = 0 & N_6 = 0 \\ N_3 = 51 & N_7 = 450 \\ N_4 = 49 & N_8 = 0 \end{array}$$

From these path flows, and Eqs. 21 through 28, the individual link flows are as follows:

$$\begin{array}{ll} m_a = 76 & m_f = 51 \\ m_b = 0 & m_g = 450 \\ m_c = 0 & m_h = 49 \\ m_d = 0 & m_i = 0 \end{array}$$

Because all interzonal flow requirements are satisfied, the analysis may advance to the test procedure of part 2 of this phase. For step 1 of the test procedure, the total network cost function is evaluated with the help of the path cost functions given in Table 4, using Eq. 9 as follows:

$$\begin{aligned} C_T &= \sum_i m_i f_i(m_i) \\ &= m_a\left[\frac{1,000}{100-m_a}\right] + m_b\left[\frac{1,000}{200-m_b}\right] + m_c\left[\frac{125}{25-m_c}\right] + \\ &\quad m_d\left[\frac{2,000}{200-m_d}\right] + m_f\left[\frac{4,500}{300-m_f}\right] + m_g\left[\frac{5,000}{500-m_g}\right] + \\ &\quad m_h\left[\frac{500}{50-m_h}\right] + m_i\left[\frac{2,000}{500-m_i}\right] \\ &= 3,167 + 0 + 0 + 0 + 922 + 45,000 + 24,500 + 0 \\ &= 73,589 \end{aligned} \qquad (29)$$

TABLE 5

UNIT DECREMENTED VARIABLE VS RESULTING NETWORK COST

Zone Pair	Decremented Variable	Resulting Network Cost
A-D	N_3	73,400
	N_4	61,089
A-B	N_5	73,422
D-B	N_7	72,609

The results of the next three steps of the test procedure are given in Table 5.

The results of step 5 of the test procedure are given in Table 6. In the case of path 6, the special situation where a link flow cannot be incremented occurred. From Eq. 27 it was noted that $m_h = N_4 + N_6$, and because N_4 already used all available capacity for link h, it was necessary to decrement N4 by one unit before N_6 could be incremented. As a result of this, it was necessary to increase one of the alternate path flows for zone pair A-D by one unit in order to maintain the required A to D interzonal flow. Because there are three such alternate paths to choose from (namely, paths 1, 2, and 3), all were individually incremented and the respective total network costs were computed. These costs are 61,104, 61,104, and 61,111. Because the minimum of these total costs occurred for both path 1 and path 2, the path 1 flow was arbitrarily chosen to be incremented. The total network cost shown in Table 6 for incremented N_6 was therefore computed with N_4 decremented by 1 and N_1 incremented by 1.

Comparing the total network costs of Table 6 to the original total network cost of Eq. 29 shows that variable N_6 must be set equal to zero and eliminated from the system of equations. Variables N_1, N_2, and N_8 must remain in the system of equations.

Phase 7—Solution of System of Equations

The equation solution procedure of this phase of the algorithm is entirely different for the two assignment options, and is therefore presented as two distinct processes. For the first assignment option, which requires total network cost to be minimized, the Lagrange technique of undetermined multipliers is used. The second assignment option, which requires the path costs for each origin-destination group of paths to be equalized, employs an iterative technique to solve a consistent set of equations for the flow variables.

The procedure for the first option begins by eliminating from the system of equations of phase 5 all variables so destined by phase 6. When this is complete, the Lagrange technique is employed to create a set of consistent equations. Finally, these equations are solved by the method of successive evaluation and minimization of error. The detailed steps are as follows:

1. Appropriate variables are eliminated from the equations of phase 5, and the constraining equations rewritten setting them equal to zero.
2. The differentials of the total network cost equation and all of the constraining equations of step 1 are determined, and each differential is set equal to zero.

TABLE 6

UNIT INCREMENTED VARIABLE VS RESULTING NETWORK COST

Zone Pair	Incremented Variable	Reference Decremented Variable	Resulting Network Cost
A-D	N_1	N_4	61,284
	N_2	N_4	61,284
A-B	N_6	N_5	74,624
D-B	N_8	N_7	72,829

3. Each differentiated constraining equation is multiplied by an undetermined function, λ, where λ is a different function for each equation. These functions are the Lagrangian multipliers.

4. All equations of step 3 are added to the differentiated total cost equation of step 2, and terms collected according to derivative.

5. Each such collection of terms is set equal to zero. These equations, together with the constraining equations of step 1, constitute a consistent set of equations, and may be solved by conventional techniques.

6. The system of equations is simplified and solved by the method of successive evaluation and minimization of error.

Concluding the example problem of Figure 4, the preceding procedure is implemented as follows: During phase 6 it was determined that N_6 must be eliminated from the system of equations; therefore, from Eqs. 18 through 28 the new constraining equations are

$$0 = M_{AB} - N_5 \quad (30)$$

$$0 = M_{AD} - N_1 - N_2 - N_3 - N_4 \quad (31)$$

$$0 = M_{DB} - N_7 - N_8 \quad (32)$$

$$0 = m_a - N_1 - N_2 - N_3 - N_5 - N_8 \quad (33)$$

$$0 = m_b - N_1 \quad (34)$$

$$0 = m_c - N_2 \quad (35)$$

$$0 = m_d - N_1 - N_2 \quad (36)$$

$$0 = m_f - N_3 \quad (37)$$

$$0 = m_g - N_7 \quad (38)$$

$$0 = m_h - N_4 \quad (39)$$

$$0 = m_i - N_8 \quad (40)$$

Next, the total network cost equations (Eq. 9) and the constraining equations (Eqs. 30 through 40) are differentiated. Because each term of Eq. 9 is of the same form, given in general by Eq. 2, the differential of the equation is computed for the general case:

$$dC_T = 0 = \sum_i \frac{\partial\left[m_i f_i(m_i)\right]}{\partial m_i} \, dm_i$$

$$= \sum_i \left\{ m_i \left[\frac{M_i(T_i - \tau_i)}{(m_i - M_i)^2}\right] + \left[\tau_i - \frac{M_i(T_i - \tau_i)}{m_i - M_i}\right] \right\} dm_i \quad (41)$$

in which i = a, b, c, d, f, g, h, i.

If, as for this example, $\tau_i = 0$, Eq. 41 becomes

$$dC_T = \sum_i \frac{M_i^2 \, T_i}{(m_i - M_i)^2} \, dm_i \quad (42)$$

Now as a result of eliminating path 6, N_5 becomes a constant and can be eliminated from the system of equations by substituting its value into each equation in which it occurs. This is done because it will reduce the number of equations and thereby simplify their solution. From Eqs. 30 through 40, using $N_5 = M_{AB}$, the results of step 2 and step 3 are

$$0 = \lambda_1(-dN_1 - dN_2 - dN_3 - dN_4) \tag{43}$$

$$0 = \lambda_2(dm_a - dN_1 - dN_2 - dN_3 - dN_8) \tag{44}$$

$$0 = \lambda_3(dm_c - dN_1) \tag{45}$$

$$0 = \lambda_4(dm_d - dN_2) \tag{46}$$

$$0 = \lambda_5(dm_d - dN_1 - dN_2) \tag{47}$$

$$0 = \lambda_6(dm_f - dN_3) \tag{48}$$

$$0 = \lambda_7(dm_g - dN_7) \tag{49}$$

$$0 = \lambda_8(dm_h - dN_4) \tag{50}$$

$$0 = \lambda_9(dm_i - dN_8) \tag{51}$$

Adding Eqs. 43 through 51 to Eq. 42, then collecting terms by corresponding differentials, and finally setting each collection of terms equal to zero (steps 4 and 5) yields the following system of equations. The values T_i and M_i used in Eq. 42 are given in Table 3.

$$\lambda_1 + \lambda_2 + \lambda_3 + \lambda_5 = 0 \tag{52}$$

$$\lambda_1 + \lambda_2 + \lambda_4 + \lambda_5 = 0 \tag{53}$$

$$\lambda_1 + \lambda_2 + \lambda_6 = 0 \tag{54}$$

$$\lambda_1 + \lambda_8 = 0 \tag{55}$$

$$\lambda_7 + \lambda_{11} = 0 \tag{56}$$

$$\lambda_2 + \lambda_9 + \lambda_{11} = 0 \tag{57}$$

$$\lambda_2 + \frac{100,000}{(m_a - 100)^2} = 0 \tag{58}$$

$$\lambda_3 + \frac{20,000}{(m_b - 200)^2} = 0 \tag{59}$$

$$\lambda_4 + \frac{3,125}{(m_c - 25)^2} = 0 \tag{60}$$

$$\lambda_5 + \frac{400,000}{(m_d - 200)^2} = 0 \tag{61}$$

$$\lambda_6 + \frac{1,350,000}{(m_f - 300)^2} = 0 \tag{62}$$

$$\lambda_7 + \frac{2,500,000}{(m_g - 500)^2} = 0 \tag{63}$$

$$\lambda_8 + \frac{25,000}{(m_h - 50)^2} = 0 \tag{64}$$

$$\lambda_9 + \frac{1,000,000}{(m_i - 500)^2} = 0 \tag{65}$$

These equations, together with the constraining equations (Eqs. 30 through 40), can now be solved. Because the preceding equations are relatively simple, it is not difficult to combine them so that all the variables are eliminated. This simplification will always be easy to accomplish, regardless of network or link cost function complexity.

In this example, Eqs. 55, 56, and 58 through 65 are substituted into Eqs. 52 through 54 and 57. The resulting equations, together with the equations obtained by substituting Eqs. 3, 4 and 5 into Eqs. 30 through 40 are the system of equations to be solved.

$$\frac{25,000}{(m_h - 50)^2} - \frac{100,000}{(m_a - 100)^2} - \frac{200,000}{(m_b - 200)^2} - \frac{400,000}{(m_d - 200)^2} = 0 \tag{66}$$

$$\frac{25,000}{(m_h - 50)^2} - \frac{100,000}{(m_a - 100)^2} - \frac{3,125}{(m_c - 25)^2} - \frac{400,000}{(m_d - 200)^2} = 0 \tag{67}$$

$$\frac{25,000}{(m_h - 50)^2} - \frac{100,000}{(m_a - 100)^2} - \frac{1,000,000}{(m_i - 500)^2} = 0 \tag{68}$$

$$\frac{2,500,000}{(m_g - 500)^2} - \frac{100,000}{(m_a - 100)^2} - \frac{1,000,000}{(m_i - 500)^2} = 0 \tag{69}$$

$$25 - N_5 = 0 \tag{70}$$

$$100 - N_1 - N_2 - N_3 - N_4 = 0 \tag{71}$$

$$450 - N_7 - N_8 = 0 \tag{72}$$

$$m_a - N_1 - N_2 - N_3 - N_5 - N_8 = 0 \tag{73}$$

$$m_b - N_1 = 0 \tag{74}$$

$$m_c - N_2 = 0 \tag{75}$$

$$m_d - N_1 - N_2 = 0 \tag{76}$$

$$m_f - N_3 = 0 \tag{77}$$

$$m_g - N_7 = 0 \tag{78}$$

$$m_h - N_4 = 0 \tag{79}$$

$$m_i - N_8 = 0 \tag{80}$$

The solution of these equations yielded the values given in Table 7. Because fractional units of flow are not admissible, the nearest integral value for each flow is also given on the table.

From Eq. 10, the total cost for each interzonal flow and the total network cost for the loaded network are computed. Table 8 shows these results for both the computed path flow values and the nearest integral path flow values.

If the second assignment option is chosen, the solution procedure begins by writing the equal path cost equations described in the discussion of phase 5. These equations, together with the constraining equations (Eqs. 18 through 28) are then solved for the path flow variables. The following are the steps in detail:

1. The equal path cost per unit of flow equations are written. These are obtained from the path composition information of Eqs. 6, 7, and 8 and the link use functions given in Table 3, or from the path use functions given in Table 4, by setting each path use function equal to an appropriate unknown path cost constant.
2. A system of equations is set up consisting of the network constraining equations and the equations of step 1.
3. This system of equations is solved by the method of successive evaluation and minimization of error. Because this method of solution adjusts the values of the path flow variables between each equation evaluation cycle, it is possible that, as convergence to the solution is approached, some path flow variables will be reduced to zero. If this occurs, and the corresponding equation's zero flow path cost is higher than the loaded cost of every other path of the same zone pair, then that equation is temporarily eliminated from the system of equations. Equations so eliminated are reinserted at later equation evaluation cycles, only if the path costs for the loaded paths of that equation's zone pair become larger than the eliminated path's zero flow value.

TABLE 7

VALUES OF PATH FLOW VARIABLES FOR FIRST ASSIGNMENT OPTION

Path		Nearest Integral Flow
N_1	20.21	20
N_2	2.55	3
N_3	32.95	33
N_4	44.29	44
N_5	25.00	25
N_6	0	0
N_7	442.27	442
N_8	7.73	8

TABLE 8

TOTAL INTERZONAL FLOW COST AND TOTAL NETWORK COST FOR FIRST ASSIGNMENT OPTION

Zone Pair	Computed Flow Total Cost	Nearest Integral Flow Total Cost
A-B	2,163	2,273
A-D	9,637	9,702
D-B	39,005	38,865
Total	50,805	50,840

For this assignment option, the example problem is concluded, first, by writing the equal path cost per unit of flow equations:

Path 1

$$\frac{1,000}{100 - m_a} + \frac{1,000}{200 - m_b} + \frac{2,000}{200 - m_d} = K_{AD} \tag{81}$$

Path 2

$$\frac{1,000}{100 - m_a} + \frac{125}{25 - m_c} + \frac{2,000}{200 - m_d} = K_{AD} \tag{82}$$

Path 3

$$\frac{1,000}{100 - m_a} + \frac{4,500}{300 - m_f} = K_{AD} \tag{83}$$

Path 4

$$\frac{500}{50 - m_h} = K_{AD} \tag{84}$$

Path 5

$$\frac{1,000}{100 - m_a} = K_{AB} \tag{85}$$

Path 6

$$\frac{500}{50 - m_h} + \frac{5,000}{500 - m_g} = K_{AB} \tag{86}$$

Path 7

$$\frac{5,000}{500 - m_g} = K_{DB} \tag{87}$$

Path 8

$$\frac{1,000}{100 - m_a} + \frac{2,000}{500 - m_i} = K_{DB} \tag{88}$$

The network's constraining equations were determined during phase 5 and are given by Eqs. 18 through 28. The required network interzonal flows, given by Eqs. 2, 4, and 5, are substituted into these equations to complete the system of equations.

The solution of this system of equations yielded the path flow assignment given in Table 9. Because the values of the path cost constants for the loaded network are determined in the process of solving these equations, each path's total cost may be computed by multiplying its assigned flow by its corresponding path cost constant. These total path costs are also given in the table.

Also from Table 9, the path cost constants sometimes differ considerably for paths between the same zone pair. This can occur for either of two reasons. First, because only integral flow assignments are permissible, and because cost of a path can change by a considerable amount as saturation loading is approached, it follows that identical path costs are impossible. Second, the unit flow cost for some paths will be greater than the loaded cost of other paths, even when these other paths carry the entire interzonal flow. Because such paths will receive no assignment, their costs are unimportant.

TABLE 9

PATH FLOW ASSIGNMENT AND COST FOR SECOND ASSIGNMENT OPTION

Zone Pair	Path	Flow	Cost Per Unit Flow	Total Path Flow Cost
A-D	N_1	18	100.0	1,800
	N_2	3	100.1	300
	N_3	34	100.2	3,407
	N_4	45	100.0	4,500
A-B	N_5	25	83.3	2,083
	N_6	0	186.2	0
D-B	N_7	442	86.2	38,101
	N_8	8	87.4	699

TABLE 10

TOTAL INTERZONAL FLOW COST AND TOTAL NETWORK COST FOR SECOND ASSIGNMENT OPTION

Zone Pair	Cost
A-B	2,083
A-D	10,007
D-B	38,800
Total	50,890

Table 10 gives the total interzonal flow costs and the total network cost for the second assignment option. For this example assignment, the network assignment and total network costs were not substantially different for the two assignment options.

ACKNOWLEDGMENTS

Robert Brenner was of considerable assistance in formulating the techniques for applying the boundary conditions to the system of equations of the capacity-restraint algorithm. B. Bussell provided valuable information concerning the use and construction of connection matrices. Frank Haight was instrumental in furnishing insight into the applicability of various types of mathematical functions to the representation of transport flow costs.

REFERENCES

1. Bock, F. C., and Cameron, S. H., "Investigation of Methods for the Assignment of Trip Demand to a Road Network. Phase 2, A Family of Special Purpose Computing Machines for High Speed Solution of Assignment Problem." Armour Research Foundation of Illinois, Institute of Technology ARF Project E091 (Oct. 1957).
2. Bellman, R., "On a Routing Problem." Quart. Appl. Math., 16:87-90 (April 1958).
3. Calland, W. B., "Traffic Forecasting and Origin and Destination Trip Assignment Techniques." 11th Annual Convention, Western Section, Institute of Traffic Engineers, Proc., paper B (June 1958).
4. Carroll, J. D., Jr., "A Method of Traffic Assignment to an Urban Network." HRB Bull. 224, 64-71 (1959).
5. "Detroit Metropolitan Area Traffic Study. Part II, Future Traffic and a Long Range Expressway Plan, March 1956." Speaker - Hines and Thomas, Lansing, Mich. (1956).
6. Dantzig, G. B., "Discrete-Variable Extremum Problems." Operations Res. Soc. of America, Jour. 5:266-277 (1957).
7. Edwards and Kelcey. "Electronic Computer Program for Traffic Assignment (BPR Program No. T-4)." U. S. Bureau of Public Roads, Office of Operations, Division of Development.
8. Case, H. W., Brenner, R., and Campbell, B., "An Examination of Criteria for Judging Regional Master Plans." UCLA, Dept. of Eng., Special Report (1962).
9. Ford, L. E., Jr., and Fulkenson, D. R., "Maximal Flow Through a Network." Canad. Jour. Math., 8:399-404 (1956).
10. Fratar, T. J., "Forecasting Distribution of Interzonal Vehicular Trips by Approximations." HRB Proc., 33:376-384 (1954).
11. Fratar, T. J., "Vehicular Trip Distribution by Successive Approximations." Traffic Quart., 8:53-65 (Jan. 1954).

12. Garrison, W. L., and Marble, D. F., "Analysis of Highway Networks: A Linear Programming Formulation." HRB Proc., 37:1-17 (1958).
13. Hoffman, W., and Pavley, R., "Applications of Digital Computers to Problems in the Study of Vehicular Traffic." Western Joint Computer Conf., Proc., Los Angeles, pp. 159-161 (May 1958).
14. Irwin, N. A., Dodd, N., and von Cube, H. G., "Capacity Restraint in Assignment Programs." HRB Bull. 297, 109-127 (Oct. 1961).
15. Kalaba, R. E., and Juncosa, M. L., "Communications Networks: I. Optimal Design and Utilization." RAND Corporation, Res. Memo. RM-1687 (April 1956).
16. Lynch, J. T., Brokke, G. E., Voorhees, A. M., and Schneider, M., "Panel Discussion on Inter-Area Travel Formulas." HRB Bull. 253, 128-138 (Jan. 1960).
17. Mertz, W. L., "Review and Evaluation of Electronic Computer Traffic Assignment Programs." HRB Bull. 297, 94-105 (Oct. 1961).
18. Mertz, W. L., "The Use of Electronic Computers." Traffic Eng., 30:23-27+ (May 1960).
19. Mertz, W. L., "Traffic Assignment to Street and Freeway Systems." Traffic Eng., 29:27-33+ (July 1959).
20. Michle, W., "Link Length Minimization in Networks." Operations Res. Soc. of America, Jour., 6:234-243 (1958).
21. Minty, G. J., "A Comment on the Shortest-Route Problem." Operations Res. Soc. of America, Jour., 5:724 (1957).
22. Moore, E. F., "The Shortest Path Through a Maze." International Symposium on the Theory of Switching, Proc., Part II, pp. 285-292 (April 2-5, 1957). Annals, Computation Lab., Harvard Univ., Vol. 30 (1959).
23. Osofsky, S., "The Multiple Regression Method of Forecasting Traffic Volumes." Traffic Quart., 13:423-445 (July 1959).
24. Pollack, M., and Weibenson, W., "Solutions of the Shortest-Route Problem—A Review." Operations Res., 8:224-230 (March-April 1960).
25. Prager, W., "Problems of Traffic and Transportation." Proc. Symp. on Operations Res. in Business and Industry, Kansas City, Mo., pp. 105-113 (1954).
26. Prager, W., "On the Role of Congestion in Transportation Problems." Zeitschr. augewande Math. u. Mech., 35:264-268 (June-July 1955).
27. Rapaport, H., and Abramson, P., "An Analog Computer for Finding an Optimum Route Through a Communication Network." IRE Trans., Professional Group on Communications Systems, CS-7:37-42 (May 1959).
28. Shimbel, A., "Structure in Communication Nets." Proc. Symp. on Information Networks, Polytechnic Institute of Brooklyn, 3:199-203 (April 12-14, 1954).
29. Sokolnikoff, I. S., and Sokolnikoff, E. S., "Higher Mathematics for Engineers and Physicists." McGraw-Hill pp. 158-167 (1941).
30. Yoeli, M., "The Theory of Switching Nets." IRE Trans., Professional Group on Circuit Theory, CT-6:152-157 (May 1959).
31. "Traffic Assignment by Mechanical Methods." HRB Bull. 130 (Jan. 1956).
32. "Traffic Assignment." HRB Bull. 61 (Jan. 1952).
33. Reichenbach, H. G., "Routine TBU." University of California, Los Angeles, Institute of Transportation and Traffic Engineering, Report 62-21 (May 1960).

Appendix

GLOSSARY

Capacity Restraint.—The process whereby assigned volumes are related to the capacity of the highway facilities in such a manner that overloaded routes become less attractive as minimum path candidates.

Centroid.—The center of gravity of a zone, located at the mathematical center of the zone as determined by the cost functions of the zone.

Connection Matrix. —A matrix representation of all links interconnecting the nodes of a network.

Converted Flow. —A component of the normal flow pattern which has made a change in its usual mode of transport.

Diverted Flow. —A component of flow which has changed from its previous path of travel to another route without a change in origin, destination, or mode of transport.

Generated Flow. —Transport flow that exists because of a particular land use.

Generator. —An area that, due to its particular kind of land use, creates transport demand.

Induced Flow. —The added component of flow which did not previously exist in any form but which results when new or improved transport facilities are provided.

Land Use Feedback. —When the location, size, costs, etc., of a proposed transport network have been fixed by virtue of transport assignment and other considerations, it is recognized that the building of the proposed system will in itself alter land use patterns which in turn affect the forecasts of interzonal movements. This definition implies that the whole transport planning process is a continuing iterative system.

Link. —The one-way portion of the transport network connecting two nodes.

Link Flow. —The total interzonal flow assigned to a link in the network. These are sometimes referred to as leg volumes.

Link Impedance. —A value assigned to each link in the network. This impedance may be some average value of travel time, it may be distance if minimum distance routes are desired, it may be cost for use of the link, or it may be any other parameter or combination of parameters so desired.

Link Cost Function. —The mathematical relation describing the cost per unit of flow expended by using a given link.

Loading the Network. —The process of assigning the interzonal flows to the network.

Network Description. —The transport network under consideration described in tabular form as nodes, directional links, link impedances, link distances, turn restrictions, etc.

Node. —A point of intersection in a transport network.

Path. —The aggregate of all links in a route between any two nodes in a transport network.

Path Cost Function. —The mathematical relation describing the cost per unit of flow expended in using a given path.

Potential Flow. —The total flow that would in all probability move between two zones (on a given route), assuming ideal transmission facilities.

Routing or Trace. —A part of a tree. In the Moore algorithms, it is the minimum path through the network from one node to another.

Transit. —The movement of people on mass public media, such as buses, streetcars, and subways, and thus a subclassification of "transportation."

Transport. —Used in the most general sense to include the movement not only of people and goods, but also of other fluxes such as energy, information, water, and sewage.

Transportation. —The movement of people and goods, and therefore, a subclassification of the more general "transport."

Transport Network. —A network of links and nodes describing the possible flow paths for interzonal or intercentroidal movements for any single transport flux.

Tree. —The aggregate of all the minimum path routings from a node to all other nodes in the network.

Tree Building. —The use of an algorithm for computing minimum paths.

Zone. —An area of a system throughout which its describing parameters can be considered constant.

A Direct Approach to Traffic Assignment

MORTON SCHNEIDER, Systems Research Chief, Chicago Area Transportation Study

• THE MOST straightforward way of obtaining the distribution of traffic in a road network is simply to distribute it: to resolve the traffic into trips, each with an origin and a destination, then find the best path through the network for each trip and note the aggregate appearances on every network member of trips following these paths. This technique is known as assignment, and is the conventional approach to the distribution problem. Though the convention is hardly an antique one, partial and minced assignments have been attempted for some time; however, the first full-scale assignment to a fairly complete metropolitan street system was performed only a few years ago. This was done by the Chicago Area Transportation Study, in the summer of 1958, on an IBM 704 computer. The computer program was then rewritten a year later to allow simpler coding and mapping procedures, and to incorporate sensitivity to congestion of streets, as well as other features. It was modified early in 1961 for the IBM 7090.

The apparently reasonable procedure of assignment has, however, several flaws, both fundamental and practical, which give rise to rather bizarre results here and there. (Also to variations on the basic method. These variations sometimes have a tenuous plausibility, are very elaborate computationally, can be adjusted to give any desired answers, and are called models.) In the first place, very little more is understood about trip-end distribution than about traffic distribution itself. Besides, what constitutes the best path (it is usually taken to be the path of least travel time), and do all trips have the same idea of best? What happens as trips interfere with each other? This latter point is widely held to be most crucial, and one ambition of nearly every neo-assignment is to account for it.

There is also the matter of zones, less widely held to be crucial. Computers can handle only so much detail, and the assignment method is, in a sense, too unsophisticated to give the computer any help. Among other reductions, the study area must be divided into arbitrary zones, and all trips emanating from a zone must be construed to have their origin at a single point representing that zone. Thus, the zonal interchange (the number of trips moving from some one zone to some other zone) replaces the trip as the elementary unit. At the same time, the synthetic and warped problem of predicting these zone-to-zone movements becomes dominant, blunting what little insight there is into trip behavior; even a perfect theory of trip distribution could correctly calculate interchanges only with a difficult explicit treatment of zone geometry. More than that, though, this condensation into points instantly shatters the locational precision of the system. Just how much does an area of some hundreds of points, all bursting with trips, spaced among thousands of street intersections resemble a metropolitan region? The errors introduced by zoning are surely considerable, possibly greater than all others. Perhaps the simulation refinements added to assignments to correct for congestion, diversion from best path, etc., are really, in their operation, nothing more than mechanical devices that spread out zone-connected errors, which is not at all the same as eliminating those errors.

Although assignments, once their peccadilloes are understood and forgiven, can be made fairly useful in a rough or over-all way, they tend to be lumpy and irresponsible in detail. Yet there is a growing demand for minor information--the effect of building a bridge here, changing the location of a ramp there. Even if assignments were very good at this sort of thing, they are too expensive and arduous to undertake for ordinary details. Certainly a small piece of area cannot be dissociated from the large region in which it resides, but that does not mean a tolerable estimate cannot be made without a

Paper sponsored by Committee on Origin and Destination.

ponderous total solution. The traffic load on a particular street section is generally not much influenced by anything more than a few miles away.

There are other odds and ends of difficulties with the assignment technique, notably boundary problems, but this paper really has nothing personal against assignment. Some if its best friends are assignments (for example, the programs mentioned earlier). The object here is to suggest a new, less stiff research posture.

A highly simplified situation is used as an example. There is a square region of unlimited size, and the rate of trip origins per day per square mile is the same everywhere in this region. Three road networks of distinctly different quality overlay the region; each network is a uniform grid. The first, highest quality network--it might even be called expressway--connects with the second, or arterial, network at every intersection of the two, but does not connect at all with the third, or local. The arterial system in turn connects with the local at every intersection of the two. All trips originate and terminate on the local system. The region is so large that perturbances due to its borders may always be ignored. Then the total length of road in each network is $2L^2/z_i$, in which L is the side of the region, and z_i is the grid interval of the particular network (the spacing between parallel members). The total number of vehicle miles per day driven in the region is by definition $\rho L^2\bar{r}$; ρ is the density of trip origins, and $\bar{r}$ is the average over-the-road trip length.

It is assumed that the daily volume of traffic is the same on every link of a particular network. The total vehicle-miles driven on each network is, therefore, just the network volume times the miles of road in the network:

$$2L^2(V_1/z_1 + V_2/z_2 + V_3/z_3) = \rho L^2\bar{r} \tag{1}$$

or, eliminating the size of the region,

$$V_1/z_1 + V_2/z_2 + V_3/z_3 = \rho\bar{r}/2 \tag{2}$$

V_i is the volume found on the i network.

For the volume to be the same on every link of a network, it is necessary (though not sufficient) that the trips leaving a network at any intersection with another network be equal to those entering at that intersection. If p_{ij} is the probability of transferring from the i to the j network,

$$p_{12}V_1 = p_{21}V_2 \text{ and } p_{23}V_2 = p_{32}V_3 \tag{3}$$

or

$$V_2 = \frac{p_{12}}{p_{21}} V_1 \text{ and } V_3 = \frac{p_{23}}{p_{32}} V_2 \tag{4}$$

Setting

$$\frac{p_{ij}}{p_{ji}} = U_{ij} \tag{5}$$

Eq. 2 may be rewritten

$$V_1\left(1/z_1 + U_{12}/z_2 + U_{12}U_{23}/z_3\right) = \rho\bar{r}/2 \tag{6}$$

It should be remembered throughout the argument that, inasmuch as the situation is set up to be entirely uniform, there is no reason for corresponding quantities at different places in the region to be anything but the same.

Here a little definition is in order. Among other things, the phrase "networks of distinctly different quality" is taken to mean that a trip that once moves from a higher to a lower quality network will never again return to the higher network—it has used the higher street as much as it profitably can and is on the way down to its destination. This is a more important construction than it may seem, because it permits a particularly simple line of attack on the p_{ij}'s.

With this delimitation, the total number of trips using a network for any part of their journey must exactly equal the number that turn from that network to a lower one, except, of course, for the local network. Once a trip enters a network, it must--whatever else it does—turn down, and it can do that only once. So the number of trips that use expressways is the number that leave an expressway at each arterial intersection, times the number of intersections; that is, $p_{12}V_1R/z_2$. Here, R is the total length of expressways. But the number of trips using expressways may also be measured as the vehicle-miles on expressways divided by the average distance traveled on expressways by those trips that use expressways at all. Thus,

$$p_{12}V_1R/z_2 = RV_1/\bar{r}_1 \tag{7}$$

Solving gives

$$p_{12} = z_2/\bar{r}_1, \text{ and similarly for arterials, } p_{23} = z_3/\bar{r}_2 \tag{8}$$

The expressway average, $\bar{r}_1$, may be gaged by choosing a convenient approximation. Although trips do not appear to be strictly governed by a distance distribution, especially in their detailed behavior, the force of distance is quite strong. Certainly, data from several regions show a consistent decay in number of vehicle trips as trip distance increases and suggest, moreover, that this decay is exponential. Probably no simple-minded distribution concept fits the data better than

$$dn/dr = ae^{-br} \tag{9}$$

The constants, a and b, are established from the natural conditions of distributions,

$$\int dn = 1 \text{ and } \int rdn = \bar{r} \tag{10}$$

yielding

$$\frac{dn}{dr} = \frac{1}{\bar{r}} e^{-r/\bar{r}} \tag{11}$$

Now this function has an amiable property: the average remaining length of a trip which has already traveled any distance, C, is just the over-all average length plus the minimum remaining length, m. From Eq. 11, the number of trips longer than C + m is

$$\int_{C+m}^{\infty} \frac{1}{\bar{r}} e^{-r/\bar{r}} dr = e^{-(C+m)/\bar{r}} \tag{12}$$

and the sum of the remaining lengths of these trips is

$$\int_{C+m}^{\infty} \frac{1}{\bar{r}} (r - C) e^{-r/\bar{r}} dr = (\bar{r} + m)e^{-(C+m)/\bar{r}} \tag{13}$$

Therefore, the average (Eq. 13 divided by Eq. 12) emerges as $\bar{r} + m$. A trip just entering an expressway may be regarded as having already traveled its total approach distance, both at origin and destination, whereas its minimum remaining length is merely the distance to the next exit, z_2. The average expressway length of expressway trips is

$$\bar{r}_1 = \bar{r} + z_2 \tag{14}$$

There is one difficulty, however. Although C, the distance already traveled, need not be constant, it must not be a function of total trip length or the calculation of the average breaks down. That is to say, expressway trips must use the expressway network as much as they can, using lower networks only as approaches to the highest, rather than as alternate routes. This further augments the definition of quality. (Further, the larger the expressway spacing, the more C tends to become related to the total trip length.)

Simple-mindedly adopting the preceding distribution gives a working evaluation of $\bar{r}_1$. But the arterial average, $\bar{r}_2$, is a different matter, because trips clearly do not use arterials as much as they might. If an arterial link is defined as a piece of arterial between two expressways, and then if every trip that used an arterial passed, say, the midpoint of a link once and only once, the total number of trips using arterials would be the arterial volume times the number of links; i.e., V_2R/z_1. (Earlier R was the total length of expressways; here it is the total length of arterials.) Actually, this quantity includes some trips counted at more than one link and neglects some trips that pass no link midpoint. Therefore, the volume times the number of links equals the total trips plus a residual; referring to Eq. 7, this may be written

$$V_2R/z_1 = V_2R/\bar{r}_2 + (D - E)R/z_1 \tag{15}$$

or

$$1/z_1 = 1/\bar{r}_2 + (D - E)/z_1V_2 \tag{16}$$

D is the number of double-counted trips at each link midpoint, and E is the number of trips per link that are never counted. It is natural to hope that D and E cancel each other out, or at least that (D - E) is small compared to V_2. It also seems reasonable when expressway spacing is in the range usually discussed. At any rate, a first approximation is better than nothing; therefore, letting the rightmost term in Eq. 16 drop out,

$$\bar{r}_2 = z_1 \tag{17}$$

This is really more an assertion than a derivation. But Eq. 16 gives some clue to the order of error.

Looking at Eq. 8, the probabilities of turning from the highest network to the intermediate, and from that to the lowest are determined, more or less, by

$$p_{12} = z_2/(\bar{r} + z_2) \text{ and } p_{23} = z_3/z_1 \tag{18}$$

One obstacle is left—the probabilities of turning from the lowest network to the intermediate, and then to the highest, p_{32} and p_{21}. This will be hurdled by the assumption that the probability of a trip entering the i network from the j network is the same as the probability that a trip already on the i network will continue on it. For example, a trip approaching an expressway on an arterial is assumed to have just the same grounds for entering the expressway as a trip on the expressway has for staying there. In notation,

$$p_{ji} = 1 - p_{ij} \tag{19}$$

Putting this, together with Eq. 18, back into Eq. 5 gives

$$U_{12} = \frac{z_2}{\bar{r} + z_2} \Big/ \left(1 - \frac{z_2}{\bar{r} + z_2}\right) = \frac{z_2}{\bar{r}}$$

and

$$U_{23} = \frac{z_3}{z_1} \Big/ \left(1 - \frac{z_3}{z_1}\right) = \frac{z_3}{z_1 - z_3} \tag{20}$$

which fairly well completes the solution. (The U_{ij}'s can be greater than 1, although the p_{ij}'s, of course, cannot.) All that remains is to rewrite Eq. 6,

$$V_1 = \rho\bar{r}/2\left(\frac{1}{z_1} + \frac{1}{\bar{r}} + \frac{z_2}{\bar{r}\,(z_1 - z_3)}\right) \tag{21}$$

$$V_2 = \frac{z_2}{\bar{r}} V_1 \tag{22}$$

$$V_3 = \frac{z_3}{z_1 - z_3} V_2 \tag{23}$$

The average trip length, $\bar{r}$, must simply be taken at its empiric value—about six miles. A quick numeric example would be a region with a trip density of 10,000 vehicles per day per square mile, expressway spacing of 5 miles, arterial spacing of 1 mile, and local spacing of 1/10 mile. The daily volumes past a point will then be expressways, about 75,000 vehicles; arterials, about 12,500 vehicles; and local streets, about 250 vehicles; which are not at all bad orders of magnitude.

The extension of these equations to more than three networks is trivial—the entire argument could have been carried out, unaltered, in generalized subscripts and open summations. The extension to more complicated connections among networks is less trivial, but more tiresome. Each connection introduces an equation of the form of Eq. 4, adds a degree of freedom in equations analogous to Eq. 8, and complicates network averages. The most elaborate situation that has been solved is one of five networks, each with as many as two connections to lower networks and two connections to higher networks. The final expressions are ugly nuisances. That kind of generalization probably drives the whole line of argument too far; the assumptions and suppositions—notably the assumption of uniform volumes on all parts of a network--begin to look doubtful even in hypothetical idealizations.

To a great extent, of course, all this really begs the question. Road systems cannot be decomposed into distinctly different networks, whatever that means; network averages are not really calculable; the assumptions of uniform volumes and complementary probabilities do not hold exactly. It is probably important, though, to carry elementary considerations to some sort of provisional conclusion, creating a platform from which enlargements can be launched.

There is no reason for this general point of view to be limited to whole networks interacting with each other. The turning probabilities could be associated with each intersection, irrespective of the kinds of networks involved. These probabilities are most interesting quantities. If they could be determined by some bold stroke of insight, they would implicitly define the entire realm of street quality and trip distribution. If these were understood then, conversely, the probabilities would be determined. A treatment of this sort might very well yield, without half-trying, the sensitivity of trip distribution to the transportation system. The deeper matter of trip generation is, at the moment, a little harder to fit into the grand vision.

In the meantime, on a more pedestrian level, some attempt ought to be made to relate the probabilities explicitly to a street quality measure. This could easily lead to expressing the p_{ij}'s as functions of the traffic volumes themselves, providing a truly elegant treatment of problems associated with trips interfering with each other. Further, trip distribution concepts must be sharpened, both to strengthen the base of the first-step uniform case and to probe the extremely important generalization to non-uniform densities.

Surprisingly, the simple expressions, as they are set down here and in somewhat modified form, are proving rather useful. There is no intention to document this as a realistic theory, but the results are realistic in a soft, fuzzy way — just as about as realistic, in fact, as assignment results. Non-uniformities must be largely ignored; and the network averages stated here — especially the arterial average — are valid only for a restricted family of spacings, which leads sometimes to patently spurious computations, most noticeable in the local network volumes. But the uses of traffic volume information are, right now, not too fine; orders of magnitude usually are enough or, at least, all that is expected. By abstracting the real street system and taking average local densities, minor questions can be answered roughly, and quickly. Certainly it would be wrong to pick this method up, as it stands, and run with it. Yet these simple equations have, perhaps, some value and some charm.