Capacity Restraint in Assignment Programs

N. A. IRWIN, NORMAN DODD, and H. G. VON CUBE; respectively, Senior Project Engineer, Traffic Research Corp., Division of KCS Limited; Transportation Research Planner, Metropolitan Toronto Planning Board; and Project Director, Traffic Research Corp., Division of KCS Limited, Toronto, Canada

Traffic prediction methods have been developed primarily to help shape and justify road planning decisions. It is an unfortunate fact that the traffic systems to which such methods must be applied are almost always inadequate. Some degree of peak-hour congestion exists in most cities today and will continue to exist in the future. To build a road network on which no congestion would occur would probably be prohibitively expensive. Any realistic method of traffic prediction must therefore recognize the presence of traffic congestion and its effect on traffic patterns.

An assignment program is essentially a means of estimating traffic flows on the basis of trip demand and traffic network data. Hitherto, designers of such programs have paid scant attention to the effect on traffic patterns, whether real or simulated, of road capacity limitations. Such limitations are, of course, the cause of traffic congestion. It is well known that delay caused by congestion will cause travelers to seek other less congested routes and will persuade some travelers to cancel their trips entirely; conversely, it is known that the existence of an uncrowded direct route between two areas will quickly draw traffic from other more crowded routes and will even create new trip demand between the two areas. These effects are implicit in the traffic pattern of any city. Assignment programs which do not take them into account in an explicit manner can make no allowance for changes in the pattern of congestion which will inevitably occur as the city evolves. In fact, such programs tend to freeze the congestion pattern to whatever shape it had when the last O-D survey was made. This can cause considerable error, especially if there are significant differences between the land-use pattern and/or traffic network proposed for the future and that in existence when the survey was made.

Under contract with the Metropolitan Toronto Planning Board, the Traffic Research Corporation has developed a systematic model for predicting vehicular traffic flow using high-speed electronic computing techniques. This model contains a direct feedback mechanism by which capacity restraints and the resultant congestion are allowed to affect route generation, trip distribution and vehicle assignment in successive program blocks. It has been tested in full-scale studies of the Toronto area and has been found to give promising results. This paper describes briefly how the simulation model works and shows how some of the results obtained from it compare with observed data.

Description of Traffic Simulation Model

Input data for the model falls into three basic categories; (a) land use (that is, where people live, work, shop, play, etc.); (b) traffic facilities (that is, what expressways, roads, etc., are available); and (c) trip characteristics (that is, on what basis people choose their destinations, mode of travel, route followed, etc.). This information is assembled in the form of tables and formulas.

Traffic simulation is then carried out by means of five types of program blocks, as follows: (a) trip generation, (b) route generation, (c) trip distribution, (d) vehicle assignment, and (d) travel time calculation. The order of occurrence of these blocks is shown in Figure 1. Briefly, the sequence is as follows:

1. A trip generation is carried out for the land-use pattern and the time period (hour of day) under study. Once completed, this block is never repeated, unless it is
desired to start another run based on a different land-use pattern and/or time period.
2. A first set of routes is determined, one between each pair of zones. Each route is the shortest possible in terms of travel time. Because at this stage there is no traffic moving on the road network, these travel times are called "ideal travel times" and the resultant routes are called the "ideal routes".
3. Trips generated in each zone are distributed to all other zones on the basis of the attractiveness of each zone and the ideal travel times as determined in step 2.
4. Trip interchanges from step 3 are translated into terms of vehicles and assigned to the pertinent ideal routes. The volume of vehicles per hour using each road section is then determined.
5. On the basis of volumes from step 4 a new travel time is calculated for each road section, or link.
6. A second set of routes is now generated. These are, in general, different from the ideal routes, because they are based on the new link times found in step 5.
7. A second assignment is carried out. The vehicles proceeding from an origin O, to a destination, D, are divided proportionately between the routes now available, on the basis of the route times.
8. A second travel time calculation is performed for each link and each route on the basis of link volumes determined in step 7.
9. The sequence of steps 6, 7, 8 is iterated twice again to produce up to four routes between each O-D pair.
10. A new distribution is performed to determine the trip interchange between each O-D pair which results when the congested travel time is used instead of the ideal time.
11. The sequence of steps 7, 8 is now iterated several times until changes in link volumes from one iteration to the next remain less than a predetermined value. In practice, the sequence of steps 10 and 11 may be repeated again if necessary. The resulting link volumes and travel times comprise the traffic pattern predicted by this simulation model for the specified time period, land-use pattern, road network and trip behavior.

Figure 1. Flow diagram for traffic simulation model.
It can be seen from the foregoing that travel time is the variable which ties all blocks of the model together. It is, therefore, important that the formula by which it is calculated be clearly understood. This leads to a more detailed discussion of the means by which capacity restraints are used to calculate the effect of traffic volume on travel time.

CAPACITY FUNCTION

The term, "capacity function," as used in this paper, means the mathematical formula developed to describe the relationship between traffic volume and travel time for a given road section. Flow of vehicles along a road is a very complex phenomenon depending on many factors. From observations it is well known that larger volumes of traffic can be moved on multi-lane roads than on single lane roads, although the volume does not increase arithmetically as the number of lanes increases. It is also a known fact that as traffic volume increases the delay due to congestion and stoppage increases. The form of this time-volume relationship is obviously affected by many things: road width, lane markings, intersections, traffic signals, transit vehicles, transit stops, trucks, turning movements and pedestrian traffic, to name a few. Each road section is probably unique in its combination of these variables and should, for precise simulation, have a unique capacity function to describe it. Moreover, it should have a different capacity function for different time periods and different weather conditions, etc. Perhaps computer capacity and empirical data will be sufficient someday for this degree of precision. For present purposes, however, it has been approximated as follows:

The road network in the study area is represented schematically by a grid of nodes and links. Because the schematic grid can rarely be as detailed as the actual road network, each link usually represents several parallel adjacent roads having similar traffic flow properties. Each link then is classified into one of ten categories, according to the speed limit and number of signalized intersections per mile on the road sections which it represents. Link properties for each category are given in Table 1. The basic capacity function formula, which applies to all categories, is

\[
T_v = \left[ t_c + \frac{d(F_v - F_c)}{F_c} \right] L
\]

in which

- \( F_v \) = link volume as calculated by assignment block (cars per hour per lane);
- \( F_c \) = critical volume = practical capacity of link above which flow becomes unstable and travel time rises rapidly (cars per hour per lane);
- \( T_v \) = link travel time at volume \( F_v \) (minutes);
- \( t_c \) = link travel time per mile at critical volume \( F_c \) (minutes per mile);
- \( L \) = link length (miles); and
- \( d \) = delay parameter (minutes per mile); \( d_1 \) for \( F_v \leq F_c \); \( d_2 \) for \( F_v > F_c \).

The variables \( F_c \) and \( t_c \), when substituted in Eq. 1, define a unique volume-time relationship for each of the ten link categories. The delay parameters, \( d_1 \) and \( d_2 \), could be made unique for each category, but present observations can be fitted adequately by using the same values for all categories; that is, \( d_1 = 0.5 \) and \( d_2 = 10.0 \).

The ten capacity functions resulting from these values are shown in Figure 2.

Data on which these curves are based come from a wide variety of sources. Most prolific of these has been the system of radar detectors mounted at the approaches to eight intersections in Toronto as part of the Computer Automated Traffic Control project. Input from these detectors is analyzed every two seconds by the computer in such a way that delay at each intersection approach can be calculated as a function of approach volume. Results from one such location are shown in Figure 3, in which the observations have been correlated by two straight line segments. The volume at which the slope
changes is known as the critical volume, $F_c$, which could also be called the free-flow capacity of the link. Above this volume, flow becomes unstable and time delay tends to increase rapidly with volume.

Curves of the type shown in Figure 3, when reduced to a volume-per-lane basis, adjusted to a standard cycle and phase time, and grouped according to category characteristics, resulted in the ten functions of Figure 2. Some of these curves are based on sketchy evidence; experiments are presently being carried out in Toronto by the Metropolitan Toronto Traffic Department, the City of Toronto Traffic Department, the Metropolitan Toronto Planning Board and the Ontario Department of Highways to fill some of the empirical gaps.

Curves of the shape shown in Figure 3 were predicted theoretically on the basis of elementary queuing theory before the observations were made. Different assumptions of arrival and departure distribution could lead also to curvilinear functions. It was felt, however, that straight line segments were the best simple functions to fit present data.

DESCRiPTION OF PROGRAM BLOCKS

Trip Generation

The study area is broken down into zones, each of which is represented schematically in the model by a centroid point or node possessing all properties of the zone except size. Land-use data for the area provide significant information such as population, employment, car registration, number of dwelling units, etc., for each zone. Relationships between this information and the number of trips actually generated in a zone were determined as follows:

The Metropolitan Toronto Planning Board carried out a home-interview survey in 1956. This indicated how many trips for each purpose and mode of travel were generated in each zone for all time periods during the day. Regression analyses were carried out to correlate the data for each time period, trip purpose and travel mode. Workable relationships were found between automobile trips generated and three land-use parameters: population, dwelling units and car registration. Multiple correlation coefficients for the time periods studied exceeded 0.95.

Using these relationships, the following categories of trips emanating from each zone are calculated:

1. Primary trips; that is, trips having one terminus at the place of residence. These are broken down into three purpose categories: (a) work, (b) business-commercial, and (c) social-recreational.

2. Secondary trips; that is, trips having neither terminus at place of residence. These are subdivided into the same three categories of trip purpose. The distinction between primary and secondary trips is necessary because different generating relationships are required for each type.

<table>
<thead>
<tr>
<th>TABLE 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CAPACITY FACTORS</strong></td>
</tr>
<tr>
<td><strong>Description of Link</strong></td>
</tr>
<tr>
<td><strong>Speed Limit (mph)</strong></td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>50</td>
</tr>
<tr>
<td>60</td>
</tr>
</tbody>
</table>
Figure 2. Capacity function curves.
EGLINTON AVE. AND AVENUE RD. - EASTERN APPROACH
CYCLE TIME = 70 SEC. (GREEN 30, AMBER 5, RED 35)
NUMBER OF WESTBOUND LANES = 2
NO BUS OR STREETCAR STOPS AT THIS APPROACH
NO PARKING ON APPROACH DURING RUSH HOUR

Figure 3. Volume-delay properties of a signalized intersection.
As input for the trip distribution block later in the program it is also necessary to determine how many trips of each purpose will be attracted to each zone during the period under study. Again the home-interview survey was used to determine relationships between land-use parameters and trip attractor figures for the three purposes. It was found that the work attractors (that is, the number of work trips arriving at each zone) could best be correlated by the variable total employment; business-commercial attractors were correlated by retail employment, and social-recreational attractors by population. In dealing with vehicle trips the availability of parking in a zone was also a factor.

The three attractors are thus calculated for each zone on the basis of employment and population within the zone. They are then adjusted on a pro rata basis so that the sum of the attractors over all zones for a given trip purpose is equal to the total trips generated in all zones for that purpose. This matching of totals is necessary to insure reasonable closure of the simulation model.

When 1956 land-use data are used, these calculations yield values of trips generated and attracted for each zone in good agreement with the survey figures, as would be expected. Relationships established for 1956 will not necessarily remain valid for future years, and a study is being carried on to determine what trends exist, if any. For present purposes, however, the established relationships must suffice. Quantities such as population, employment, car registration, etc., can be estimated for the future on the basis of past trends, zoning restrictions and economic forecasts. If desired, ranges of such quantities can be studied. Having made these estimates, the trips generated in and attracted to each zone can be calculated for the future time period under study. This is the function of the Trip Generation program block.

Route Generation

As previously mentioned, the road network under study is represented by a grid of nodes and links. Each link is fully described for purposes of traffic simulation by three variables: one indicating which capacity function is applicable, a second indicating its number of lanes and a third indicating its length in tenths of a mile. Turning delays and restrictions can also be included, if desired.

![Figure 4. A typical set of routes emanating from a selected node.](image)
Given the volume of cars per hour using a link at any time during the simulation procedure, it is possible, using the appropriate capacity function, to calculate the travel time required to traverse that link. This information is required by the Route Generation block, which determines by a rapid method of trial and error the shortest route in terms of travel time between every pair of centroid nodes in the grid.

The algorithm used is based on that developed by E. Moore and G. Dantzig but has been modified to minimize the number of times a given node must be queried. This means that routes going from an origin to nodes close by are minimized before longer routes are built onto them. In this way revisions of the longer routes are avoided with a resultant saving in machine time. A typical set of routes emanating from one node is shown in Figure 4.

As the simulation commences, link volumes are all zero, so that link travel times are all ideal. The set of routes found on the basis of these link times is known as the set of ideal routes. The grid shown in Figure 4 contains 250 nodes, of which 99 are centroid nodes. The number of routes in the ideal set is therefore 99 x 98 = 9702. (Note that the direct route A to B and the reverse route B to A may follow quite different paths.) As the simulation proceeds, other sets of routes are generated on the basis of whatever link times apply at the time. A given set of routes tends, therefore, to avoid areas which are congested at the time it is generated. All ideal routes are stored in the computer memory for later use. Most routes generated in subsequent iterations are also stored; as many as four routes can at present be retained for any O-D pair. The effect of this procedure is to allow travelers from an O to a D a reasonable choice of routes to follow. This is illustrated in Figure 5 where four such routes are shown: the first route follows a roundabout course to use 60-mph expressways for most of its length; route 2 also makes use of an expressway, whereas routes 3 and 4 are forced by expressway congestion to use slower arterial roads.

The generation of a new route is initiated only if the ratio of O-D travel time after the last updating of link times to the ideal O-D travel time exceeds a parameter which can be pre-set at any desired level or which can be changed by the program itself during the run. To avoid duplication a new route is recorded only if it differs in at least one link from any of the previously recorded routes.

Figure 5. Example of four routes between an O-D pair.
Trip Distribution

The distribution problem consists of determining what number of trips will be made from each origin to each destination. It is assumed that this number will depend, for each trip purpose, on the total number of trips available for distribution at the origin in question, the total number of trips attracted to the destination in question, and the ease of getting from O to D, as measured by the travel time. This leads to the basic Gravity Model formula:

\[ J_{ij} \propto G_i A_j F_{ij} \]  \hspace{1cm} (2)

in which

- \( J_{ij} \) = number of trips going from origin \( i \) to destination \( j \) for the purpose in question;
- \( G_i \) = trips generated for this purpose at origin \( i \);
- \( A_j \) = trips attracted for this purpose at destination \( j \);
- \( F_{ij} = (T_{ij})^{-a} \) = time factor for the trip between origin \( i \) and destination \( j \); and
- \( T_{ij} \) = average travel time from origin \( i \) to destination \( j \).

To insure that the number of departures from an origin actually equals the trips generated there (that is, that \( \sum_j J_{ij} = G_i \)) the basic formula is put into proportional form:

\[ J_{ij} = \frac{A_j F_{ij}}{\sum_j A_j F_{ij}} G_i \]  \hspace{1cm} (3)

However, this formula does not insure the correct number of arrivals at each destination; (that is, there is no guarantee that \( \sum_i J_{ij} = A_j \). In practice with 99 centroid nodes and the pattern of \( G_i \) and \( A_j \) as found in Toronto, it is found that a significant number of nodes have a factor \( A_j / \sum_i J_{ij} \) less than 0.5 or greater than 2. To lessen this effect, a second iteration of the distribution calculation is carried out using adjusted attractors:

\[ A'_j = A_j^{2} / \sum_i J_{ij} \]  \hspace{1cm} (4)

in which

- \( A'_j \) = adjusted attractor at destination \( j \), for use in second iteration of trip distribution calculation.

One such iteration within the trip distribution block brings the attractors and arrivals into balance to within 5 percent in most cases.

A separate distribution calculation must be carried out for each of the three trip purposes, because each purpose has its own set of generators and attractors. It has been found also that different time factor indices are necessary for different purposes. Those currently being used in the vehicular simulation are, as follows:

(a) for work trips: \( F_{ij} = T_{ij}^{-1} \)  \hspace{1cm} (5)
(b) for business-commercial and social recreational trips:

\[ F_{ij} = T_{ij}^{-2} \]  \hspace{1cm} (6)

Investigations presently being carried out indicate that these indices should be reduced.

The modified gravity formula produces results which compare well with observed data. It takes into account the effect of travel time and, as modified, insures adequate
balancing of arrivals with attractors. This form of distribution calculation therefore allows the effects of capacity restraints and congestion to make themselves felt on inter-zonal trip interchange in a reasonable way. It has been adopted as the best empirical formula available and one that is capable of further refinement as experience grows.

**Vehicle Assignment**

In the assignment block, vehicles going from each O to each D are loaded onto the links comprising the route or routes available between each O-D pair.

Where more than one route is available the vehicles are proportioned among them according to the following formula:

\[
J_{rij} = \frac{(T_{rij})^{-b} J_{ij}}{\sum_{r=1}^{n} (T_{rij})^{-b}} \tag{7}
\]

in which

- \(J_{rij}\) = number of trips going from origin \(i\) to destination \(j\) via \(r\)th available route \((1 \leq r \leq n)\);
- \(n\) = number of routes available between \(i\) and \(j\) \((1 \leq n \leq 4\) in present program);
- \(T_{rij}\) = travel time from \(i\) to \(j\) via \(r\)th available route;
- \(J_{ij}\) = total number of trips going from \(i\) to \(j\) via all \(n\) available routes; and
- \(b\) = index which can be adjusted empirically.

At present this proportional assignment is carried out with \(b\) set to 1. Further investigation is proceeding to determine a more representative value if necessary.

It can be seen that the proportional assignment is another means by which capacity restraints are taken into account in this simulation model. Although the ideal route from an O to a D will have the shortest travel time under non-loaded conditions, its popularity may lead to traffic congestion which increases its travel time to a higher value than that for the other available routes. The proportional assignment allows this effect to be simulated in a reasonable manner.

**Travel Time Calculation**

Having obtained link volumes from the assignment block, this block calculates link travel times using the capacity function formula. The new link travel times are stored for possible use in subsequent route generations. They are also used to update all route times, should these be necessary for an assignment or a new trip distribution. For the latter the O-D time is calculated as the average of all route times between the O and D, weighted by the number of O-D vehicles using each route:

\[
\bar{T}_{ij} = \frac{\sum_{r=1}^{n} (J_{rij} T_{rij})}{\sum_{r=1}^{n} (J_{rij})} \tag{8}
\]

in which

- \(\bar{T}_{ij}\) = average travel time between origin \(i\) and destination \(j\).
EFFECTS OF CAPACITY RESTRAINTS—SUMMARY

The effects of capacity restraints on travel time make themselves felt at three points in this simulation model:

1. In finding of routes. Route generations are carried out under differing conditions of congestion to provide several reasonable routes from every O to every D.

2. In choice of destinations. Trip distributions are carried out under conditions of practical as well as ideal traffic patterns to simulate the effect of congestion on travelers choice of destination.

3. In choice of route. Confronted with several possible routes from an O to a D, simulated travelers are allowed to choose among them so that more take the shortest route than take the longest.

It should be emphasized that this model is still in the development stage. Much more work is required to verify and refine many of the functions and parameters used. Nevertheless, it is capable of producing meaningful results even in its present state, as results of a current production run will show.

RESULTS OF TEST RUN

Under contract with the Metropolitan Toronto Department of Roads, the Ontario Department of Highways and the Metropolitan Toronto Planning Board, the Traffic Research Corporation is presently applying this traffic simulation model to a study of the Toronto area. The area of emphasis is the Toronto by-pass expressway, Highway 401, and the time period is an evening rush hour in the year 1980.

Before being applied for a future time period the model is always applied to a recent historical time period for the city in question. Observed volume counts across cordon lines and screenlines can then be compared with simulated volumes to check the accuracy of parameters and functions used. On the successful completion of the test run, the model is applied to the future time period using the same program sequence.

The preliminary test run for this study was an evening rush hour in 1958. For checking purposes, comparisons were made between simulated and observed values for that time period. Some of these comparisons are shown in Figures 6 through 10.

Figures 6, 7 and 8 are mainly a check on the validity of the simulated trip distribution. Observed data for these graphs were taken from the O-D Survey of 1956, expanded to 1958 on the basis of land-use changes. As it can be seen in Figures 6 and 7, tolerance limits must be set wide because of wide variability in the survey data. It is evident, however, that the trend of trip interchange volumes and O-D travel times as simulated is close to that shown by the observed data.

Figures 9 and 10 show a comparison of volumes along Highway 401 and across a typical screenline several miles south of the highway. Observed counts were taken during the period 1957 to 1959. Again, simulated volumes are seen to follow the observed trend quite closely.

A convenient yardstick for comparing transportation costs in a system under various conditions is the total time spent by all vehicles on the road during the period in question. This "system cost" is usually expressed in terms of vehicle minutes. Mathematically it is the volume of vehicles using each link multiplied by the travel time for that link, summed for all links in the system:

\[
\text{System cost} = \sum_{\text{all links}} (F_V \times T_V)
\]  \hfill (9)

This parameter is also useful as an indication of the degree of program settlement as successive iterations are carried out during the simulation. Figure 11 shows how the system cost varied during the 1958 run.

The value at point D1 (first distribution) indicates system cost under ideal conditions (zero congestion on all links) with one route (the ideal route) available between each O-D pair. System cost between D1 and A4 (fourth assignment) has little absolute
Figure 6. O-D trip interchange comparison: simulated vs observed.
Figure 7. O-D travel time comparison: simulated vs observed.
Figure 8. Trip frequency—travel time comparison: simulated vs observed.
Figure 9. 1958 traffic volumes on Highway 401—Westbound, P.M. weekday rush.

- HWY 27
- DIXON RD.
- ISLINGTON AVE.
- WESTON RD.
- HWY 400
- JANE ST.
- KEELE ST.
- DUFFERIN ST.
- SPADINA EXP.
- BATHURST ST.
- AVENUE RD.
- YONGE BLVD.
- YONGE ST.
- BAYVIEW AVE.
- LESLIE ST.
- WOODBINE AVE.
- VICTORIA PK. RD.
- WARDEN AVE.
- KENNEDY RD.
- MARKHAM RD.
- LITTLES RD.

Cars per hour graph.
Figure 10. Traffic volumes crossing screenline just north of Bloor Street.
meaning because arbitrary percentages of distributed trips were assigned in order to find alternate routes as quickly as possible. The value at A4 represents the system cost again under ideal conditions, but now with up to four routes available between each O-D pair. This is higher than the cost at D1, as would be expected, because the second, third and fourth routes found could not, under ideal conditions, be as short as the ideal set. The system cost at A10 (tenth assignment) is that under actual equilibrium conditions as found by the simulation model. As would be expected it is higher than either of the ideal system costs.
When early trials were made of the simulation model, it was found that individual link volumes and times (and hence the system cost also) had a tendency to fluctuate violently from iteration to iteration. A degree of link time dampening was introduced to reduce these oscillations and speed convergence: whenever a link time is calculated it is modified by a weighted function of the previous time for that link. The effect of this dampening can be seen as the system cost rises from A4 to A10. Subsequent runs with less dampening than was used here have approached the point of critical dampening, with even faster convergence as a result.

CONCLUSIONS

The use of capacity restraints and the resultant feedback of congestion effects by means of travel time is felt to be essential in traffic simulation programs. The model described in this paper incorporates such effects and has been shown to give realistic results.

Further research into the method is progressing. Similar techniques are being applied to the simulation of combined transit and vehicular traffic, with promising results.

BIBLIOGRAPHY