

HIGHWAY RESEARCH RECORD

Number 297

Improvements
in the
Transportation
Planning Process

6 Reports

Subject Area

55 Traffic Measurements
84 Urban Transportation Systems

HIGHWAY RESEARCH BOARD

DIVISION OF ENGINEERING NATIONAL RESEARCH COUNCIL
NATIONAL ACADEMY OF SCIENCES—NATIONAL ACADEMY OF ENGINEERING

Washington, D. C., 1969

Publication 1675

Price: \$2.40

Available from

Highway Research Board
National Academy of Sciences
2101 Constitution Avenue
Washington, D.C. 20418

Department of Traffic and Operations

Harold L. Michael, Chairman
Purdue University, Lafayette, Indiana

HIGHWAY RESEARCH BOARD STAFF

E. A. Mueller

COMMITTEE ON ORIGIN AND DESTINATION

(As of December 31, 1968)

Alan M. Voorhees, Chairman
Alan M. Voorhees and Associates, Inc., McLean, Virginia

Norman Mueller, Secretary
Bureau of Public Roads, Washington, D.C.

Mark M. Akin
Willard H. Armstrong
Herman Basmaciyan
Austin E. Brant, Jr.
Glenn E. Brokke
Henry W. Bruck
Nathan Cherniack
Donald E. Cleveland
Francis E. Coleman
James H. Cox
Roger L. Creighton
Lewis W. Crump
John A. Dearing
Thomas B. Deen
Harold Deutschman
John W. Dickey
George A. Ferguson
Michael G. Ferreri
Devere M. Foxworth
W. L. Grecco
Douglas F. Haist

Harold W. Hansen
C. S. Harmon
Philip Hazen
Kevin E. Heanue
Donald M. Hill
Frank E. Horton
Thomas F. Humphrey
G. H. Johnston
Louis E. Keefer
Norman Kennedy
Dewey Lonsberry
Dana E. Low
Ted Luke
Frank J. Mammano
Brian V. Martin
James J. McDonnell
W. L. Mertz
John K. Mladinov
John D. Orzeske
Robbie W. Parker
William S. Pollard, Jr.

Lloyd A. Rivard
James J. Schuster
Arthur Schwartz
Billy J. Sexton
Paul W. Shuldiner
Jacob Silver
Bob L. Smith
Max R. Sproles
Vergil G. Stover
Edwin N. Thomas
Anthony R. Tomazinis
Robert E. Whiteside
George V. Wickstrom
David K. Witheford
Martin Wohl
Robert T. Wood
Richard D. Worrall
Robert H. Wortman
J. E. Wright
F. Houston Wynn

Foreword

Highway officials and planners, as well as others involved in the transportation planning process, will find the papers in this RECORD of interest.

In the first paper, Lund proposes a means of estimating urban travel in order to reduce the time and cost of local transportation planning studies. The model used is a modification of a gravity model developed by Tanner of the British Road Research Laboratory, which generates and distributes trips in one operation and which has not been used for a transportation study in this country. The author concludes that his proposal does in fact offer simplicity and economy of application, together with ease and economy of data collection.

Kassoff and Deutschman made a critical evaluation of the consequences associated with alternate approaches to the trip-generation process. Their research deals with two areas of interest: (a) the use of aggregated data to examine the performance of relationships based on aggregate totals as opposed to aggregate rates, and (b) the use of disaggregated data vs aggregated data. They conclude that the trip-generation equations derived from disaggregated data show the most promise of arriving at relationships that reflect the true nature of correlation between a set of independent variables and travel behavior.

McCarthy offers a report of his examination of the validity of the major assumptions underlying the methodology of multiple-regression trip-generation analysis based on data aggregated to the zonal average. His findings imply that zonal averages are not truly representative of the individual household traits and refute the validity of the commonly accepted assumption of zonal homogeneity.

The fourth paper discusses the calibration of transit networks in medium-size urban areas. Hartgen shows that a transit network of average complexity can be calibrated in two to three assignments, given, of course, previous agreement of network bus speeds with reported system speeds.

Spielberg proposes a diversion-curve modal-split model to convert person trips to vehicle trips on a highway network. Curves relating person trips to several variables were developed, and the model was tested. The author concludes that this type of model is well suited to prediction of vehicle usage.

The last paper, by McCann and Maring, describes a license plate survey method of obtaining travel characteristics of motorists. By testing the method in an actual origin and destination survey, the authors identified some limitations, but also decided that the method is feasible.

Contents

A SIMPLIFIED TRIP-DISTRIBUTION MODEL FOR THE ESTIMATION OF URBAN TRAVEL John W. Lund	1
TRIP GENERATION: A CRITICAL APPRAISAL Harold Kassoff and Harold D. Deutschman	15
MULTIPLE-REGRESSION ANALYSIS OF HOUSEHOLD TRIP GENERATION—A CRITIQUE Gerald M. McCarthy	31
CALIBRATION OF TRANSIT NETWORKS IN MEDIUM-SIZED URBAN AREAS David T. Hartgen	44
AUTOMOBILE OCCUPANCY PROJECTIONS USING A MODAL-SPLIT MODEL Franklin Spielberg	57
LICENSE PLATE TRAFFIC SURVEY Howard McCann and Gary Maring	68

A Simplified Trip-Distribution Model for the Estimation of Urban Travel

JOHN W. LUND, Oregon Technical Institute

A means of estimating urban travel is proposed to reduce the time and cost of local transportation planning studies. A trip-distribution model, which is a modification of a gravity model developed by J. C. Tanner of the British Road Research Laboratory, is studied. The principal variables in this model are a measure of the activity within each zone of the study area and a measure of the resistance to travel between each pair of zones. A combination of resident population and employment is used for the activity measurement, and both straight-line distance and minimum off-peak driving time is used for the resistance measurement. Several different forms of the model are investigated including a uniform resistance function, ring variations of the parameters, and the addition of terminal time to the resistance measurement.

Origin-destination data are used from a 1962 survey in Pueblo, Colorado (population 108,243). The ability of the model to simulate the actual travel patterns is based on root-mean-square error analysis. The minimum root-mean-square error gave the best estimates of the origin-destination data as substantiated by screen-line crossing counts, trip-length frequency distribution, and percentage of root-mean-square error. In all cases, the use of travel resistance based on straight-line distance gave better results than the use of minimum driving time. This was because of land use characteristics of the urban area, such as very centralized shopping, business, and work centers within the study area.

•THE METHOD proposed in this report is an attempt to offer a simplified method of making frequent forecasts of urban travel demands so that portions of this process can be performed with a minimum of data collection and trained personnel. The model presented is a modification of a gravity model developed by J. C. Tanner of the British Road Research Laboratory (1) that generates and distributes trips in one operation, a model not yet used for a transportation study in this country. The principal variables in this model are a measure of the activity within each zone in the study area and a measure of the resistance to travel between each pair of zones. The simplified socio-economic measure of activity used is a combination of resident population and employment, because these are the main trip-generation and attraction characteristics in an urban area. The measure of resistance used is both straight-line distance and actual driving time as determined from a minimum-path calculation. Various combinations of activity and resistance are tested against origin-destination data and two screen-line crossings to determine the best simulation of the origin-destination data.

This report uses data from Pueblo, Colorado (study area population of 108,243), and considers only the internal traffic and vehicle trips rather than person trips. The method in this study does not consider trip purpose, transit trips, and intrazonal trips because they could be included with the use of additional information. The purposes of this report are simplification and an appraisal of the effects of the basic parameters. The inclusion of too many details would hinder this purpose.

STUDY AREA CHARACTERISTICS

The Colorado Department of Highways in cooperation with the U. S. Bureau of Public Roads and the Pueblo City-County Planning Commission began a transportation study of metropolitan Pueblo, Colorado, in 1962. The study is referred to as the Pueblo Area Transportation Study or, more commonly, PATS. The origin-destination data and dwelling unit statistics obtained in the study are used as the basis for this report.

The study area, which includes most of Pueblo County, encompasses approximately 127.4 square miles. The area was initially divided into 536 zones, following the boundaries of census tracts developed in the 1960 Censuses of Population and Housing. The 536 zones were then combined into 200 traffic districts, and finally these were combined into 57 analysis districts. The latter districts do not follow the census tracts in all cases. The analysis districts (referred to as AD's), which are the basis of study for this report, range in size from approximately 0.13 to 13.2 square miles. The study area is shown in Figure 1, and the analysis districts are shown in Figure 2.

The PATS survey began in 1962, and data were gathered by home-interview questionnaires. A minimum of 12½ percent of the dwellings in each of the 536 zones were sampled, and in some zones the sample was as high as 100 percent. The average sample was approximately 13 percent. The data were then expanded and compared with the number of work trips and ground counts taken at two screen-line crossings. Based on these comparisons, the origin-destination data were adjusted to agree with the ground counts. The two screen-line crossings were made at the Fountain Creek and the Arkansas River screen lines shown in Figure 2. Only internal trips were considered; external trips were excluded based on a previous study. Based on a 24-hour count, the origin-destination data were 82.2 percent of the ground count for the Fountain Creek screen line and 83.8 percent for the Arkansas River screen line. This error is close to that experienced by other transportation studies (2).

The origin-destination (O-D) data are based on average weekday traffic (AWDT), which is 6 percent higher than average daily traffic (ADT), giving a more conservative

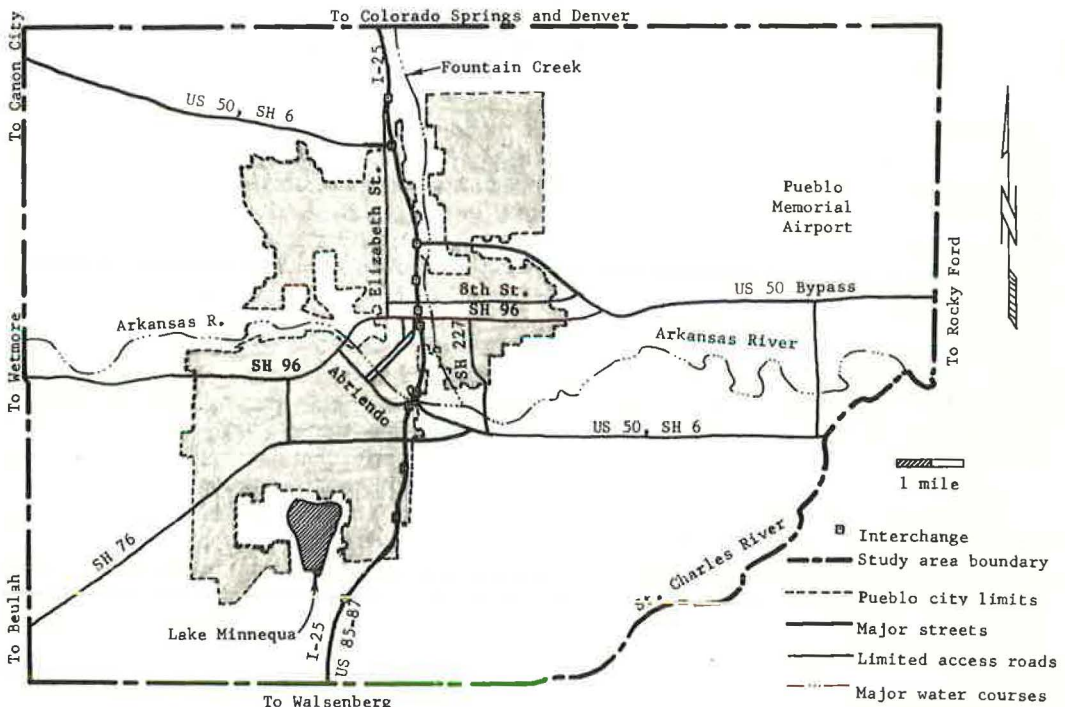


Figure 1. Location of the Pueblo Area Transportation Study area.

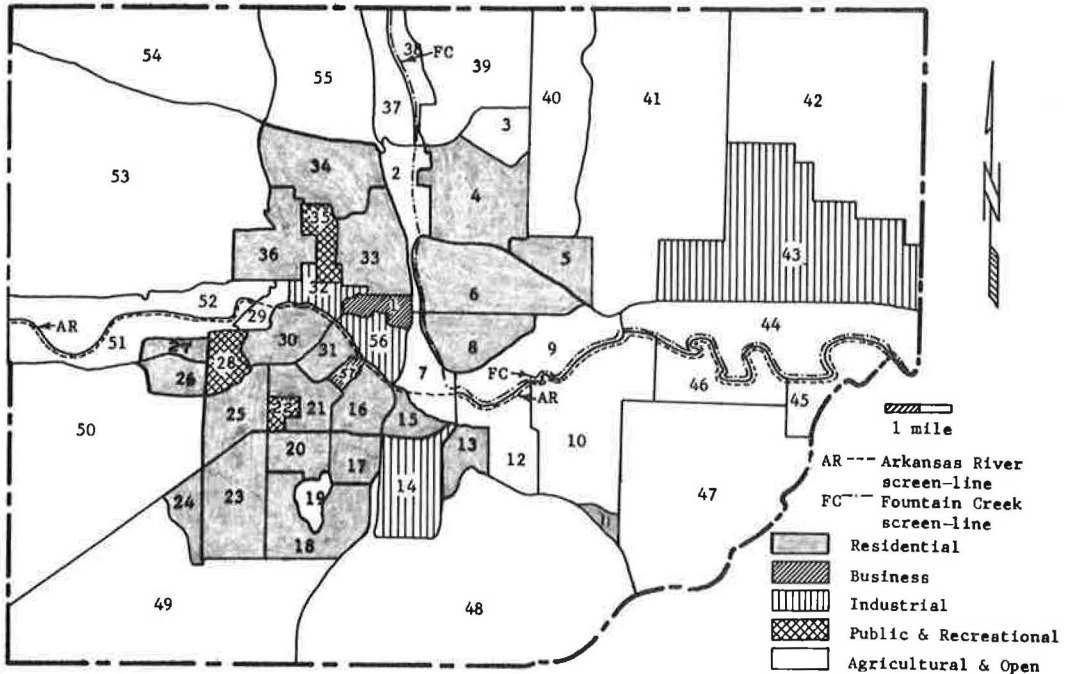


Figure 2. Analysis districts with prevalent land use and screen-line locations.

travel pattern for planning purposes. The AWDT results are used in this report. A summary of the analysis district characteristics is given in Table 1.

TANNER MODEL

Tanner proposes the following model:

$$t_{ij} = \frac{m P_i P_j e^{(-Bd_{ij})}}{(d_{ij})^x} \left[\frac{1}{C_i} + \frac{1}{C_j} \right] \quad (1)$$

where

t_{ij} = the number of trips per day between the zones i and j ;

m = a constant;

B = a constant;

P_i = the population, or other measures of activity of zone i ;

d_{ij} = the distance between the zones, or the time or cost of traveling between them;

C_i = a constant, being defined by $C_i = \sum_k P_k e^{(-Bd_{ik})} d_{ik}/(d_{ik})^x$

$$\left[\text{if } x = 1, C_i = \sum_k P_k e^{(-Bd_{ik})} \right];$$

x = distance exponent; and

$e = 2.718$.

If P equals population, m then becomes the total travel per person per day. Thus, while other gravity-type models are based on the total number of trips generated in each zone, the Tanner model is based on the total travel for all residents of each zone. Tanner also takes into account the total surrounding urban area by the expression in

TABLE 1
ANALYSIS DISTRICT CHARACTERISTICS

Analysis District	Distance From CBD (miles) ^a	Resident Population (persons) ^b	Employment (persons) ^c	Area (acres)	Residential Density (person/acre) ^d
1	0.00	1,102	4,588	214.7	5.12
2	1.89	87	236	604.8	0.14
3	3.16	0	0	436.8	0.00
4	1.91	6,979	382	1,396.8	5.00
5	2.89	1,511	12	241.6	6.25
6	1.47	10,275	726	1,348.8	7.62
7	0.92	921	263	537.6	1.71
8	1.28	5,450	198	724.8	7.52
9	2.72	291	66	1,088.0	0.27
10	3.64	1,134	60	2,110.4	0.54
11	4.61	395	0	116.8	3.38
12	2.73	683	167	777.6	0.88
13	2.39	1,819	55	344.0	5.28
14	2.72	53	7,137	1,020.8	0.05
15	1.80	2,020	284	236.8	8.56
16	1.51	6,587	458	459.2	14.35
17	2.29	4,748	1,162	352.0	13.49
18	3.27	5,672	178	985.6	5.75
19	2.86	0	0	198.4	0.00
20	2.64	5,043	143	396.8	12.72
21	1.87	6,078	434	472.0	12.87
22	2.13	40	276	100.8	0.40
23	3.62	9,317	376	1,068.8	8.72
24	4.49	134	27	339.2	0.40
25	2.60	4,143	248	612.8	6.78
26	3.33	1,377	28	336.0	4.10
27	3.29	602	11	178.2	3.38
28	2.53	22	42	321.4	0.07
29	2.03	63	40	204.0	0.31
30	1.60	4,248	398	512.0	8.30
31	1.09	2,892	100	280.0	10.32
32	0.66	349	70	267.2	1.31
33	0.90	7,009	2,237	847.2	8.26
34	2.08	5,887	254	1,100.8	5.36
35	1.18	152	2,810	307.2	0.49
36	1.87	3,968	49	576.0	6.89
37	3.38	78	4	622.4	0.12
38	3.70	126	0	345.6	0.36
39	3.75	46	0	1,950.4	0.02
40	3.19	0	0	1,889.6	0.00
41	3.71	85	0	5,404.8	0.02
42	7.15	0	0	5,240.0	0.00
43	6.23	84	699	4,056.0	0.02
44	5.85	688	225	2,256.0	0.31
45	6.79	99	0	625.6	0.16
46	5.04	96	0	867.2	0.11
47	5.51	2,071	54	4,142.4	0.50
48	4.55	220	135	8,441.6	0.03
49	5.57	0	14	4,868.8	0.00
50	4.28	133	12	4,673.6	0.03
51	3.67	539	5	859.2	0.63
52	2.62	291	16	1,230.4	0.24
53	3.37	418	0	7,441.6	0.06
54	5.33	224	0	2,968.0	0.08
55	4.13	133	75	1,979.2	0.07
56	0.65	952	2,517	438.1	2.17
57	1.20	894	368	85.6	10.48

^aStraight-line distance between centroids (based on population).

^bDetermined from origin-destination survey.

^cDetermined for April 1963 by district of job location.

^dResident population divided by total area.

the brackets, which provides a measure of competing travel opportunity surrounding each zone. The details of the Tanner derivation are presented elsewhere (1).

The assumption used in this report as applied to Eq. 1 is that t_{ij} represents one-way trips from zones i to j . This departure from the Tanner model was necessary for the use of ring variations as discussed subsequently. Making this assumption for the other variations, summarized at the end of this section, only changes the value of m by a half.

The population or activity measurement that is used in this report is a summation of resident population and employment in the form of

$$P_i = k_r R_i + k_e E_i \quad (2)$$

where

- P_i = activity of zone i ,
- R_i = resident population of zone i ,
- E_i = employment in zone i , and
- k_r, k_e = population coefficients to be determined.

Linear Regression Equation Solution

The process of determining the best solution of the Tanner model requires a determination of the coefficients m and B of Eq. 1. This requires that resistance, d_{ij} , and activities, P_i and P_j , be measured, and k_r and k_e of Eq. 2 and the distance exponent, x , be assumed.

Choosing a value of B , we can determine the corresponding value of m because m is a constant for a particular value of B . As a measure of fit, a least squares calculation is made for each determination of B and m , called the root-mean-square error or RMS. Thus by assigning a value to B and calculating m , we can calculate the root-mean-square error, RMS. If values of B are chosen over the appropriate interval, then the value of RMS will reach a minimum point. This minimum point would then be the optimum value for B and m in the Tanner model based on the assumptions stated earlier. This optimum point can also be determined for various values of x , the resistance exponent, and for k_r and k_e , the population coefficients. In all variations studied, the total number of internal trips (excluding intrazonal trips) was held to 382,800 as determined by the O-D study.

TANNER MODEL VARIATIONS

Uniform Resistance Function

Four different variations of the resistance function were investigated using a uniform resistance to travel over the entire study area. The uniform resistance to travel means that a unit of distance (or time) has the same impeding effect on travel at any point in the study area. The uniform resistance to travel is obviously the simplest and easiest model to consider, especially in the case of a straight-line distance between zones. No terminal times or barrier penalties were considered in these cases.

Straight-Line Distance—The first case studied was that of straight-line distances, where distances were measured between centroids of the analysis districts in miles. The locations of the centroids were based on resident population when there were no employment centers. Various combinations of values for k_r and k_e were chosen so that the ratio of k_e to k_r varies from 0.5 to 3, with x set equal to 1. Values of B used were from +0.40 to -0.40 at increments of 0.05. The results of the least squares analysis are shown in Figure 3. These graphs indicate that the root-mean-square (RMS) reaches a minimum value.

The Arkansas River and Fountain Creek screen-line-crossing values were calculated along with the RMS, and the difference between the calculated and the surveyed trip volumes are plotted on the graphs. In all cases the Arkansas River screen-line zero-error point is closest to the minimum RMS value. Except for k_e/k_r equal to 3, the difference between the screen-line zero-error point and the RMS minimum value is relatively small. The Fountain Creek screen-line zero-error point is associated with large positive values of B and is farther away from the RMS minimum point. An interesting point is that as k_e/k_r increases, the zero-error points of the two screen lines move toward each other, with the Fountain Creek screen-line zero-error point moving at greater increments. The RMS curves in these four graphs are relatively flat, thus the change in RMS by slightly increasing or decreasing B is small. Thus, use of either the Arkansas River screen-line zero-error point for B or the minimum RMS point for B would not seriously affect the accuracy of the Tanner model.

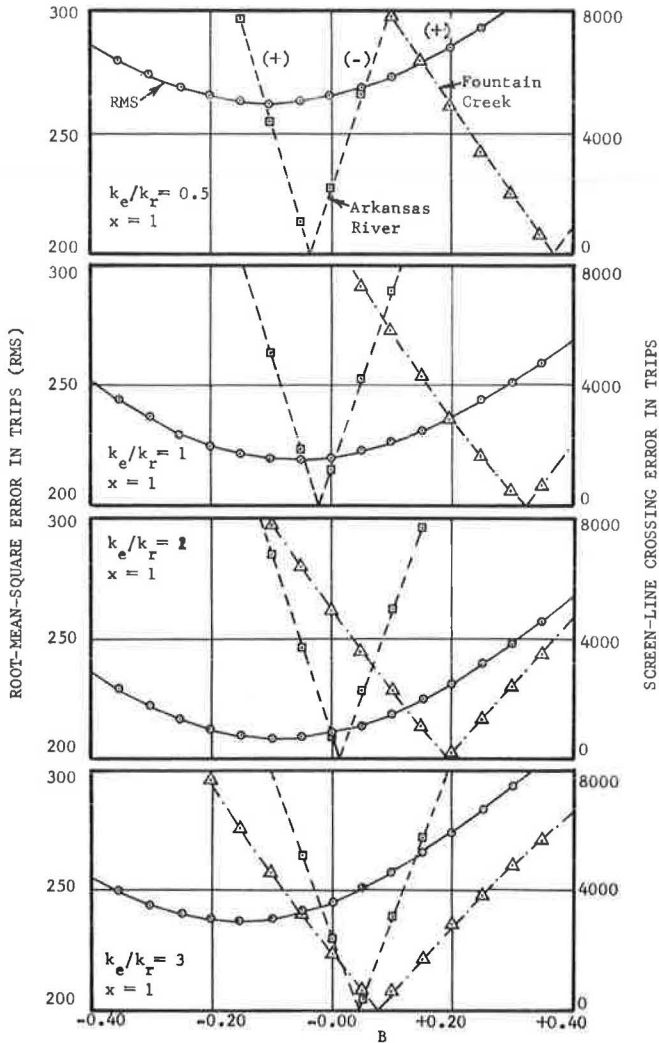


Figure 3. Least squares analysis using straight-line distance.

The various values of k_e/k_r were compared to the minimum RMS and these are plotted in Figure 4. The value of m varied from 1.8 to 3.0 miles. The minimum value of the RMS appears to be somewhere between values of 1 and 2 for k_e/k_r , and closer to 2. A similar investigation was performed for $x = 2$ in the Tanner Model. The results are also illustrated in Figure 4.

Minimum Driving Time—The next logical variation to investigate was the use of driving times between analysis districts. These driving times were obtained from minimum-path times or trees.

A street map with average off-peak driving speeds was available from the Colorado Department of Highways. Based on this map, the principal street system was determined using 100 nodes and connecting all of the 57 analysis district centroids. The driving time of each link was calculated and the minimum path and driving time between each analysis district centroid was determined, based on minimum-path tree solutions. This minimum driving time without terminal times was then used for the travel resistance in the least squares analysis.

The results of the least squares analysis using these minimum driving times for resistance gives curves that are more concave in shape, when compared to the straight-

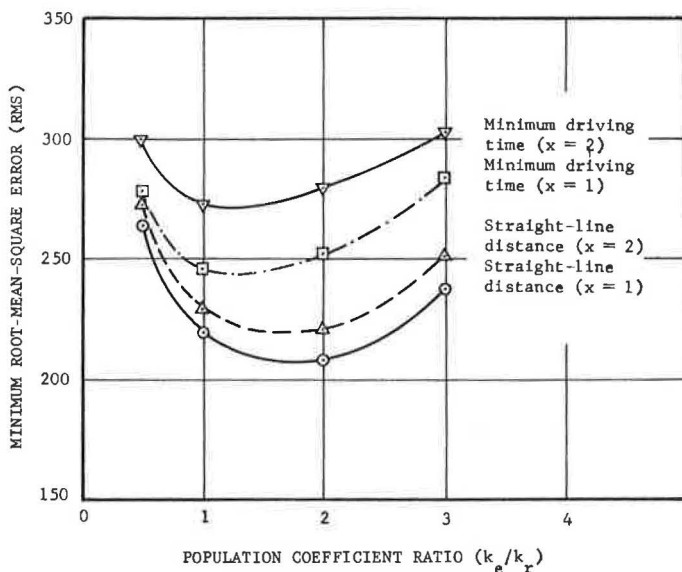


Figure 4. Minimum root-mean-square error vs population coefficient ratio using a uniform resistance function.

line distance curves, and the location of their minimum points are very sensitive to changes in B . The two screen-line-crossing error curves are in the same relative location but are steeper, indicating that they are also more sensitive to changes in B . Figure 4 presents a plot of the minimum RMS for values of k_e/k_r that can be compared with the curves for straight-line distances. The value of m varied from 5.4 to 9.7 minutes. In this case, the best fit appears to be for a value of $k_e/k_r = 1$. The use of $x = 2$ was also investigated, and the results are shown in Figure 4.

Variations of B and m at Ring Intervals From the CBD

A more detailed analysis of the results of the uniform resistance function reveals some interesting trends. If the total trips generated for each analysis district are averaged at 1-mile intervals (rings) from the CBD, the results are shown in Figure 5 for the two best solutions of the RMS for straight-line distance and the best for minimum driving time. It is evident from Figure 5 that there are pronounced differences in the distance ranges of 0 to 1 mile, 1 to 3 miles, and over 3 miles from the CBD. The only question is whether or not the 2- to 3-mile ring should be included with the over 3-mile ring. However, because the over 3-mile ring is prevalently negative or near zero and the 2- to 3-mile ring is positive, it is best grouped with the 1- to 2-mile ring.

A further study of the results of the uniform resistance function was made to determine if there were any unusual variations in trip generation or distribution by sectors from the CBD. The percentage of error in trips generated for each AD was plotted on a map of the study area. The trend that appeared significant was that of ring variations. In the case of $k_e/k_r = 2$ for straight-line distance, of a total of 36 AD's with positive errors, 33 of these fell in the 1- to 3-mile ring, only one (AD 7) in the 0- to 1-mile ring, and two (AD's 38 and 48) in the over 3-mile ring. Similar results were also observed for the best RMS fit using minimum driving time ($k_e/k_r = 1$, and $x = 1$). In this case, 16 of 27 AD's with positive errors fell in the 1- to 3-mile ring and only 1 in the 0- to 1-mile ring. This, then, appears to verify the ring distribution of errors illustrated in Figure 5. No significant geographic distribution of errors was apparent.

Comparison of the trip-generation error to distance from the CBD indicated that the use of this relationship could probably improve the fit (or minimum RMS) of the Tanner model. This requires the use of different B and m values for each of the rings with distances of 0 to 1 mile, 1 to 3 miles, and over 3 miles from the CBD.

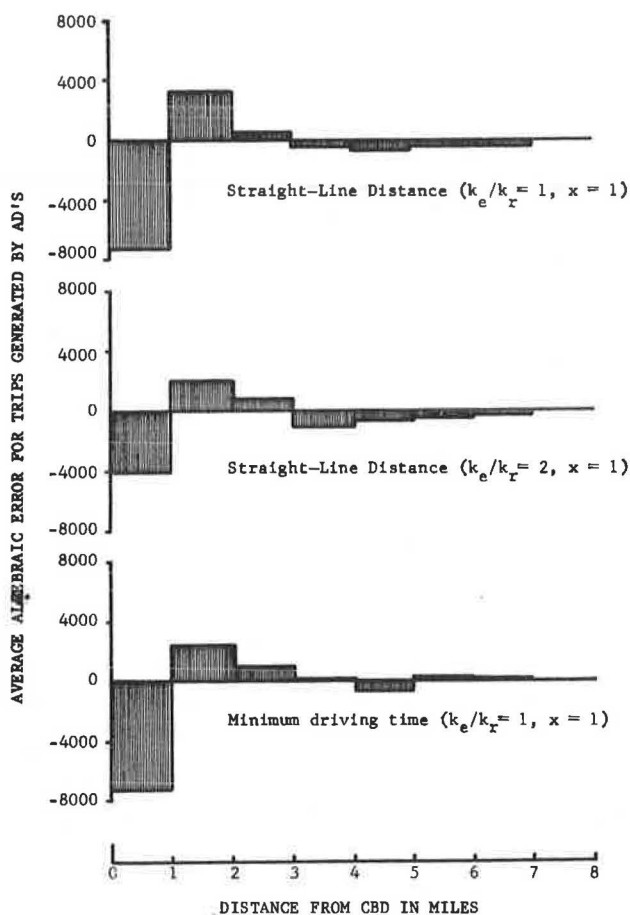


Figure 5. Average algebraic error in trips generated by analysis districts for the uniform resistance function.

variations and those for the uniform resistance function revealed that the former improved the minimum RMS fit for all variations. The minimum point for each curve was improved by an RMS of about 20 trips, or approximately 9 percent, and the minimum point for best case of the two variations changed from an RMS value of 208 to 189.

In this method of analysis by rings, the screen-line-crossing errors could not be immediately observed because there were too many variables involved. Only after the minimum RMS value for each ring was determined, separately and then combined, could the screen-line-crossing errors be calculated for that particular solution. Calculating the screen-line-crossing errors for the minimum RMS for straight-line distance and minimum driving time ($k_e/k_r = 1$, and $x = 1$) gave the following results: For the Fountain Creek screen-line crossing, the error was 13.7 percent and 29.3 percent respectively; and for the Arkansas River screen-line crossing, the error was -0.4 percent and 13.0 percent respectively. This trend of small Arkansas River screen-line error and large Fountain Creek screen-line error is similar to that found in the uniform resistance function investigation.

Terminal Times

In many studies, spatial separation between zones appears to be better approximated by travel time than by driving time. Travel time is the sum of the driving time plus terminal times within the zone of origin and destination.

Again, the root-mean-square error analysis (RMS) as described in the previous section was used to analyze the trips generated in an AD for the minimum RMS by rings at distances of 0 to 1, 1 to 3, and over 3 miles from the CBD. Trips terminating in an AD were given the B and m values from the ring of origin.

Straight-Line Distance—The results of the straight-line distance resistance function are given in Figure 6. This includes straight-line distance and straight-line distance squared, or $x = 1$ and $x = 2$ as used in Eq. 1. In both straight-line distance cases, the minimum point was closest to $k_e/k_r = 1$, with the value of $x = 1$ giving the best fit. The percentage of root-mean-square error (%RMS) was smallest for the CBD ring and largest for the outlying rings.

Minimum Driving Time—Using minimum driving time revealed a trend similar to that of the straight-line distance cases. In this instance the RMS and %RMS were somewhat larger. The results of the minimum driving time analysis, shown in Figure 6, indicate a similar shape but poorer fit compared with the straight-line distance case.

Comparing the results of the minimum RMS values for ring

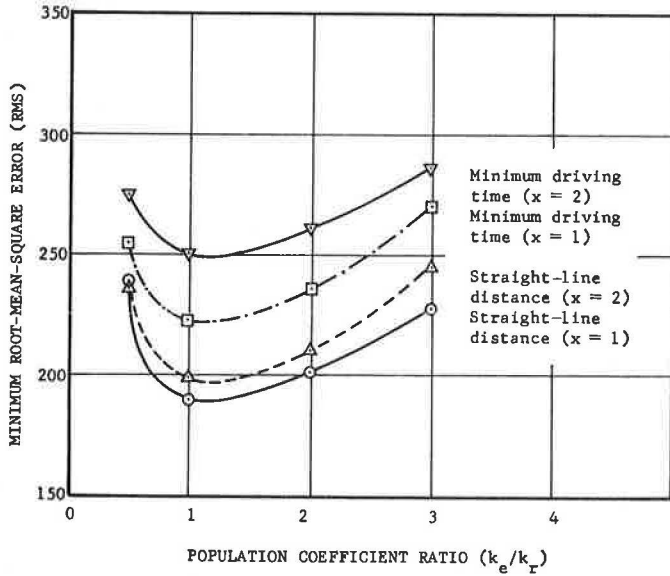


Figure 6. Minimum root-mean-square error vs population coefficient ratio using ring variation of m and B at 1, 3, and 8 miles from the CBD.

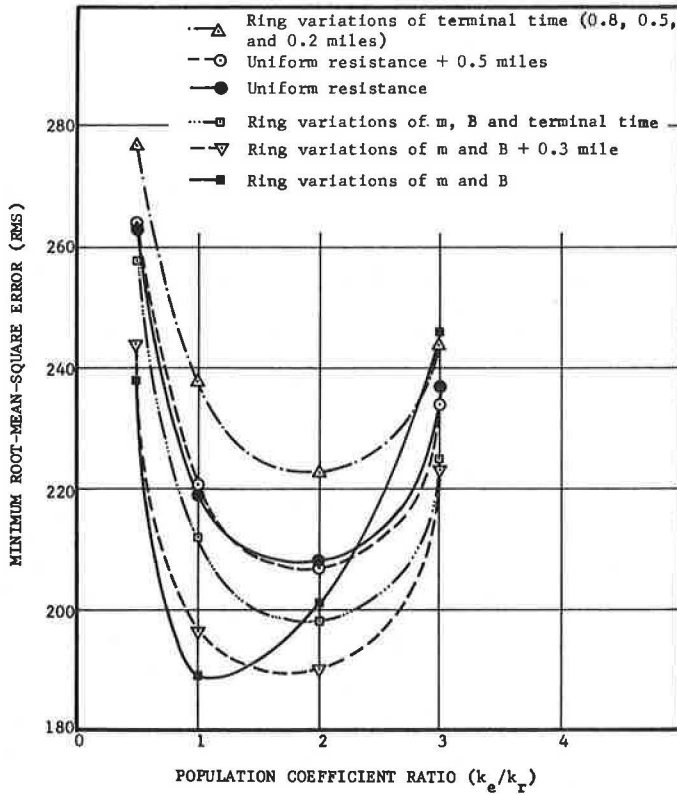


Figure 7. Terminal time added to straight-line distance measurement.

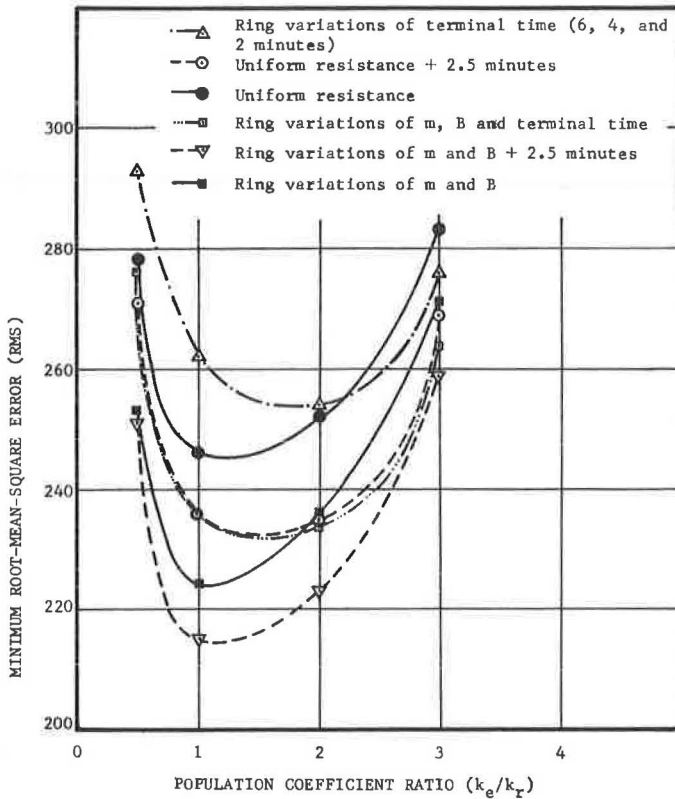


Figure 8. Terminal time added to minimum driving time measurement.

The determination of actual terminal times has been difficult, and various approximations are used such as the terminal time of each zone as determined by the land use characteristics of the zone, or as a function of the distance from the CBD. In one study for a small urban area (4), a terminal time of 5 minutes was used for the CBD, 1 minute for residential zones, and intermediate values for other zones. In the Washington, D.C., study for 1948 and 1955 (9), the estimated terminal times varied from 6 minutes within the central portion of the region to 3 minutes in the outlying suburban residential areas.

There are, of course, many possible combinations of terminal times, resistance measurements, and ring variations that might be investigated. Because it was not practical to study all possible combinations, several of the more significant ones were examined, based on the results observed in the first part of this report. A summary of these variations is briefly outlined as follows:

1. Straight-line distance—(a) uniform terminal time, expressed in miles, with up to 1.0-mile terminal time studied; (b) ring variations of terminal time of 0.8 mile for the central ring, 0.5 mile for the intermediate ring, and 0.2 mile for the outlying areas; (c) ring variations of m and B in the Tanner equation with constant terminal time of 0.3 mile added to each trip end; and (d) ring variations of m , B , and terminal time using the values listed in 1(b). These cases are shown in Figure 7.

2. Minimum driving time—(a) uniform terminal time with up to 5.0 minutes added to each end of a trip studied; (b) ring variations of terminal time of 6 minutes for the central ring, 4 minutes for the intermediate ring, and 2 minutes for the outlying area; (c) ring variations of m and B in the Tanner equation with constant terminal time of 2.5 minutes added to each trip end; and (d) ring variations of m , B , and terminal time using the values listed in 2(b). These cases are shown in Figure 8.

DISCUSSION OF RESULTS

The most unusual finding from the entire study was concerned with the measurement of the resistance parameter used in the Tanner model. This finding is best shown in Figure 9 where, in every case, the straight-line distance form of the resistance measurement gives a better fit (in terms of minimizing the RMS) to the origin-destination data than does the minimum driving time form. This finding was observed without exception throughout the study. In almost all references and transportation study reports (except possibly those of the Detroit Area Transportation Study), minimum driving time or minimum driving time with terminal time is cited as the best indicator of a driver's appraisal of travel resistance. The logic of this argument is obvious and, in most cases, correct. If other factors remain constant, the attraction for travel decreases as resistance to travel between two points increases (decreased driving speed). The trip production is then satisfied by other destinations where the resistance to travel is less (or the opportunity for trip termination is greater). This is especially true for a densely developed downtown area where congestion is common. Outlying shopping centers in medium and large cities, where CBD congestion is the rule, attract a considerable part of the daily volume of shopping trips.

Straight-line distance measurements, on the other hand, do not consider congestion, barriers, or variations in the ease of travel throughout an urban area. This decrease in CBD-oriented travel associated with increase in congestion has been observed in many cities, and has best been accounted for by using actual travel time for the resistance measurement (3).

Pueblo is no exception to problems of congestion and low driving speeds in the downtown area. In addition, several major barriers exist near the CBD, among which are Fountain Creek, the Arkansas River, and the railroad lines. There is also a heavy industrial area in the south-central portion of the study area. All of these factors point toward the use of minimum driving time or total travel time.

However, looking at the study area characteristics in more detail gives a better insight into the reason for the better results occurring with the use of straight-line distance as the resistance measurement. First, there are only three major shopping areas in the Pueblo area, one in the southwest corner of AD 4, one at the intersection of AD's 20, 22, 23, and 25, and one at the west end of AD 1. The two shopping areas outside the CBD (AD 1) offer mainly food and variety store services. Any major shopping or business requirements have to be satisfied by the downtown area; thus, most residents of the area do their shopping in the CBD. Second, transit service was almost nonexistent in 1962, accommodating 3,500 person trips with an average of 8 persons per bus; bus trips represented approximately 0.1 percent of the total vehicle trips. Another major factor is that approximately 25 percent of the available jobs in the study area are accounted for by the steel mills located in AD 14. Another 30 percent of the jobs are located within 1 mile of the CBD. Because a great majority of trips by urban residents are made between home and work, business, or shops, these two areas, AD 1 and AD 14, together with the adjacent business and industrial areas (AD's 32, 56, 57, and the south portion of 33), account for over a fourth of the trips in the study area.

Thus, trips to these areas are not governed as much by accessibility as by necessity. Resistance to travel is not very meaningful in this case, especially minimum driving time that tends to reduce the number of trips to these high-density areas because of low driving speeds and barriers.

It is concluded, therefore, that straight-line distance is a better estimator of resistance to travel in the study area because of the high number of one-location facilities for shopping and business, which probably conditions drivers to think in terms of straight-line distance and not driving time. This conclusion is felt to apply only to the Pueblo study area and is not necessarily valid for other urban areas.

It was found that the minimum RMS in all of the examples investigated was between a k_e/k_r value of 1 and 2, indicating a greater emphasis on employment as a determinant of travel demand. The greater weight given to employment more than likely gave a better estimate of work- or business-oriented trips, as mentioned earlier. Because approximately 40 percent of all person trips and 60 percent of all home-based person

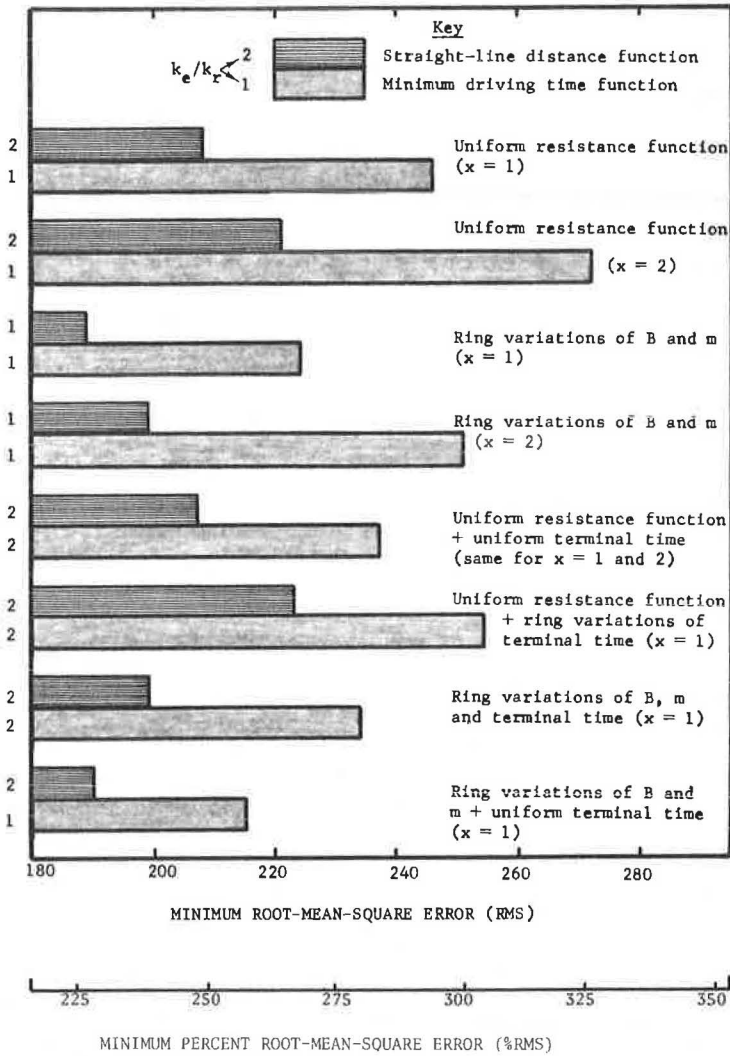


Figure 9. Summary of results from least squares analysis applied to the Tanner trip-distribution model.

trips in an urban area are for the purpose of business, shopping, or work (4), the greater weight given to employment is reasonable (5). However, the emphasis on employment increases at the expense of accuracy for other trip purposes such as social-recreational and school. Thus, there is an optimum balance between the employment factor and the resident-population factor.

A summary of all the variations investigated is shown in Figure 9. In each pair of bar graphs, the top one is the straight-line distance function and the lower one is the minimum travel time function. The best fit is indicated by the shortest bar.

The %RMS is plotted in Figure 10 for the best fit of the uniform resistance function for straight-line distance, minimum driving time, and ring variations of B and m (straight-line distance function). The data of Washington, D.C., gathered in 1955 are plotted for comparison (9). The Tanner model results are based on individual inter-zonal trip interchanges, whereas the Washington, D.C., data are based on assignment to a spider network, and the sample rate lines (3 percent and 13 percent) are based on accumulated trip-trace intersections in 1/4-mile sections of a grid system superimposed over the survey area (3). Thus the Tanner model results are somewhat more stringent

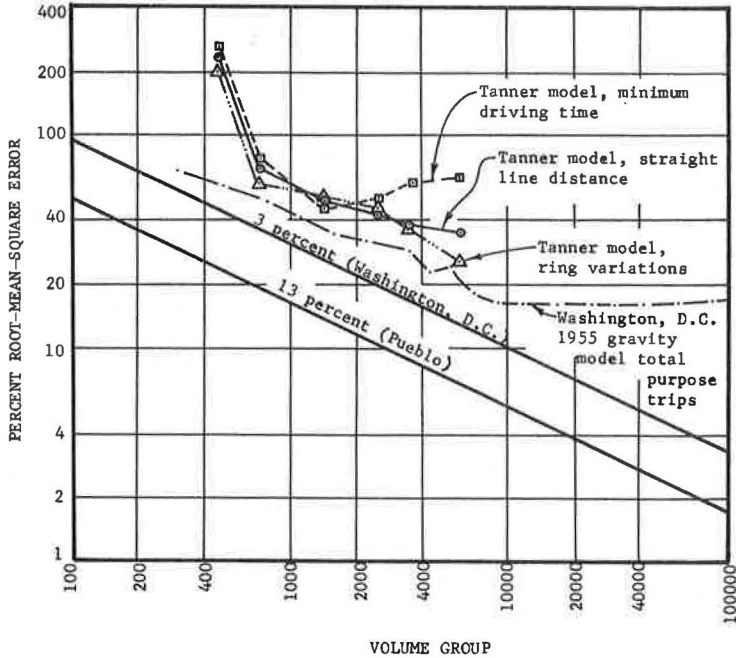


Figure 10. Comparison of percentage of root-mean-square error by volume group.

because there was no chance for geographic or socioeconomic bias to be "averaged out." Use of the spider-network volumes allows positive and negative errors, in trips assigned to a specific link, to cancel each other.

CONCLUSIONS

The main purpose of this study was to develop a method for analyzing urban travel patterns that offered simplicity and economy of application together with ease and economy of data collection. Both of the above requirements are felt to be satisfied by using the Tanner model. The use of straight-line distance between centroids for travel resistance, and resident and employment population for the activity measurement, meets the requirements of ease and economy of data collection. Population is one of the easiest urban characteristics to collect as was proven in the study of another urban area.

The use of the straight-line distance form of the Tanner model, with slight variation in the coefficients obtained for Pueblo, produced favorable results for Boulder, Colorado (population 60,000). No origin-destination data were available, thus the results were checked by several screen lines. The data for the Boulder study were collected in one day, and the program was calculated and adjusted in less than two days.

The method described in this report is not as detailed as many of those developed and used in major transportation studies. However, it is not intended to replace these methods, but instead to supplement major studies and to be used as a planning tool for smaller urban areas. This procedure would be especially valuable for analyzing alternate forecasts of development and growth of urban areas. The use of population is especially valuable because population growth and change can be predicted with a greater degree of reliability and ease than most other urban characteristics. The Tanner model was used for estimating present travel in this report; however, this does not preclude its use for predicting future travel. The method lends itself to forecasting with the use of future population projections.

The main problem encountered in this study was the unusual travel characteristics dictated by the concentration of shopping, business, and work areas in Pueblo. This unfortunately makes the results somewhat difficult to apply to other urban areas that

have highly decentralized commercial areas, because the effect of this on the Tanner parameters is not completely understood.

The form of the Tanner model that gave the best results was with the use of straight-line distance for the resistance function. However, the use of minimum travel times should be investigated for other areas because it may give as good or better results. The use of ring variations instead of the uniform function is questioned because of lack of simplicity and other problems.

Several items of interest were brought out as the study developed, which could offer topics for further research.

1. The application of the Tanner model to U.S. Census data, as information on resident and employment population by place of work becomes available in the 1970 census program (6, 7).

2. The adaptation of the Tanner model to intrazonal trip distribution or to specific subareas such as the CBD, industrial parks, and university districts, along with variations in k_e/k_r values.

3. The use of the Tanner model with resident and employment population for peak-hour trip estimation. Indications are that this method would give good estimates because approximately 80 percent of urban travel during this time is between home and work (8).

4. The use of different k_e/k_r values for rings within the study area or for major traffic generators such as the CBD, industrial parks, and university districts.

ACKNOWLEDGMENTS

This report is a summary of a PhD dissertation prepared under the counsel of Professors Dennis R. Neuzil and Irving Weiss of the University of Colorado. Origin-destination data and other material were obtained from the Planning and Research Department of the Colorado Division of Highways.

REFERENCES

1. Tanner, J. C. Factors Affecting the Amount of Travel. Road Research Laboratory, London, Technical Paper 51, 1961.
2. Oi, W. Y., and Shuldiner, P. W. An Analysis of Urban Travel Demands. Northwestern Univ. Press, Evanston, Ill., 1962.
3. Calibrating and Testing a Gravity Model for Any Size Urban Area. U.S. Govt. Printing Office, Washington, D.C., 1965.
4. Wilbur Smith and Associates. Future Highway and Urban Growth. The Automobile Manufacturers Association, Detroit, 1961.
5. Survey Findings. Chicago Area Transportation Study, Chicago, Vol. I, Dec. 1959.
6. Fisher, R. J., and Sosslau, A. B. Census Data as a Source for Urban Transportation Planning. Highway Research Record 141, 1966, pp. 47-72.
7. Hansen, M. H., and Voight, R. B. Availability of Census Data for Urban Areas. Highway Research Record 194, 1967, pp. 21-31.
8. Martin, B. V., Memmott, F. W., III, and Bone, A. J. Principles and Techniques of Predicting Future Demand for Urban Area Transportation. The Massachusetts Institute of Technology Press, Cambridge, Rept. 3, 1961.
9. Heanue, K. E., and Pyers, C. E. A Comparative Evaluation of Trip Distribution Procedures. Highway Research Record 114, 1966, pp. 20-50.

Trip Generation: A Critical Appraisal

HAROLD KASSOFF, U.S. Bureau of Public Roads; and
HAROLD D. DEUTSCHMAN, Tri-State Transportation Commission

This paper is directed toward a critical evaluation of the consequences associated with alternate approaches to the trip-generation process. The first part of the research examines the use of aggregated data and the performance of relationships based on aggregate totals (such as trips per zone) as opposed to the use of aggregate rates (such as trips per household per zone). The second part concentrates on the implications of using disaggregated data (data not combined and averaged according to predefined areal units) vs aggregated data. Both analyses were performed on the same data base and employed the same set of variables. The analysis tool of multiple linear regression was applied to both phases of the research utilizing two independent sets of data, one for calibrating or fitting the data, and the other for testing the results.

The results indicate that the aggregate total equation has a slight statistical advantage, but the rate equation offers more flexibility and efficiency in analyzing the data, because it is not tied to the data scheme to which it was developed. In statistical tests to measure and evaluate on a common basis the disaggregate trip-generation procedures vs the aggregate procedures, the disaggregate equations produced slightly better results and are the recommended procedure.

•ONE of the distinctive features of current transportation planning is the explicit recognition given to the relationship between travel behavior and the physical, social, and economic state of urban environment. During the past decade, a considerable effort has been made by transportation planners to develop analytical tools that can couch these basic relationships into a quantifiable framework for use in forecasting future travel demands. This has led to the emergence of a set of procedures, under the general heading of trip generation, that can translate information concerning land use, population characteristics, and economic conditions into expressions of travel potential. This paper is directed toward a critical evaluation of the consequences associated with alternate approaches to the trip-generation process.

The bulk of the research effort to date in the area of trip-generation analysis has been oriented toward establishing a set of factors that can be related to trip-making and used in developing travel forecasts. The results have been fruitful in the sense that the basic determinants of urban travel behavior have been fairly well identified and documented, and their use by operational studies in producing forecasts has become more or less standard practice. Less of an effort has been made in developing and refining the specific techniques that have been employed in relating travel behavior to the underlying causal factors. The purpose of this paper is to apply a particular analytical tool, multiple linear regression, in a number of alternate ways, and to evaluate the results in terms of the applicability of each approach to the trip-generation process. Specifically, the paper deals with two areas of interest to the analyst engaged in pre-

paring travel forecasts. The first part examines the use of aggregated data and the performance of relationships based on aggregate totals (such as trips per zone) as opposed to the use of aggregate rates (such as trips per household per zone). The second part deals with the implications of using aggregated data vs data that have not been combined and averaged according to predefined areal units. Both analyses were performed on the same data base and employed the same set of variables.

STUDY APPROACH

The trip-generation process may be viewed in terms of two distinct steps. The first involves the generation of a set of trips on a small-area basis, based on the characteristics of the population of the area under study, with an assignment of a portion of the trips to residential areas. This phase is frequently referred to as residential-trip generation. The second, nonresidential-trip generation, involves the allocation of the nonhome trip ends to nonresidential activities throughout the area. The same general tool of multiple regression is applicable to both phases. Because this research is concerned only with evaluating alternate techniques, it was decided to deal with residential trip generation alone because this phase is generally characterized by relationships that are more accurate and more stable.

The term trips, as used in this paper, refers to unlinked person trips generated by residents (over 5 years of age) of the particular analysis unit under consideration. Walking trips are not included. (Publications of the Bureau of Public Roads on gravity models contain a discussion of the trip linking.)

Perhaps the most critical constraint that must be placed on a research effort aimed at evaluating a number of analytical techniques is that all factors having some bearing on the results of the analysis must be carefully controlled. This means that only the specific items to be tested may be permitted to vary within the overall analysis. To ensure that these conditions were met, the research was conducted on a single data set, employed a single set of variables, and performed evaluations at uniform levels of data aggregation.

Data Source

The source of data for this study was the home-interview survey conducted by the Tri-State Transportation Commission in 1963 and 1964 over a 22-county region encompassing the New York City metropolitan area. A one percent sample of households produced travel data and socioeconomic characteristics for over 50,000 households within the Tri-State cordon area. The 3,600 square-mile area had a population of over 16 million persons in 1963. (It should be noted that none of the techniques discussed in this paper reflects the trip-generation process developed by the Tri-State staff. At Tri-State, an advanced traffic-assignment technique, called the direct traffic estimation method, required trip-destination estimates at a much finer areal level than that dealt with in this paper.)

Selection of Variables

The selection of variables to be used in the analysis represented a significant initial step in the research. The choice was subject to the following criteria:

1. Variables should be highly correlated with trip-making in a statistical sense.
2. Variables should have a strong logical relationship with trip-making in a causal sense.
3. Variables should generally not be difficult to forecast.
4. Variables should have been commonly used by operational studies in trip-generation analysis.
5. Variables should be limited in number so that the analysis is not distorted with a multitude of interrelated factors.
6. Variables must be compatible with all of the techniques to be tested.

TABLE 1
SIMPLE CORRELATION OF VARIABLES

Variable	Variable							
	1	2	3	4	5	6	7	8
1. Trips	—	0.867	0.770	0.705	0.839	0.901	0.873	0.800
2. Persons (over 5 years old)	—	—	0.786	0.925	0.971	0.686	0.896	0.971
3. White-collar labor force	—	—	—	0.547	0.885	0.629	0.946	0.885
4. Blue-collar labor force	—	—	—	—	0.875	0.512	0.713	0.891
5. Labor force	—	—	—	—	—	0.650	0.946	0.992
6. Automobiles	—	—	—	—	—	—	0.743	0.593
7. Income	—	—	—	—	—	—	—	0.919
8. Households	—	—	—	—	—	—	—	—

Note: Variables represent zonal totals.

The sixth criterion eliminates from consideration a number of variables that have been used frequently in trip-generation analysis. A variable such as net residential density, for example, is usually a good measure of trip rates, but it is not applicable in forecasting aggregated totals because it does not reflect size differentials among areal units. A preliminary list of variables was compiled and a simple correlation analysis was performed to evaluate the variables in terms of the first criterion. This was accomplished at the zonal level, and the results are given in Table 1.

The variables most highly correlated with trip-making were total persons (those persons over 5 years of age), labor force, automobiles, and total income. Of these, income was eliminated because it is a difficult variable to forecast, particularly on a small-area basis, and labor force was dropped because it, too, is frequently not readily

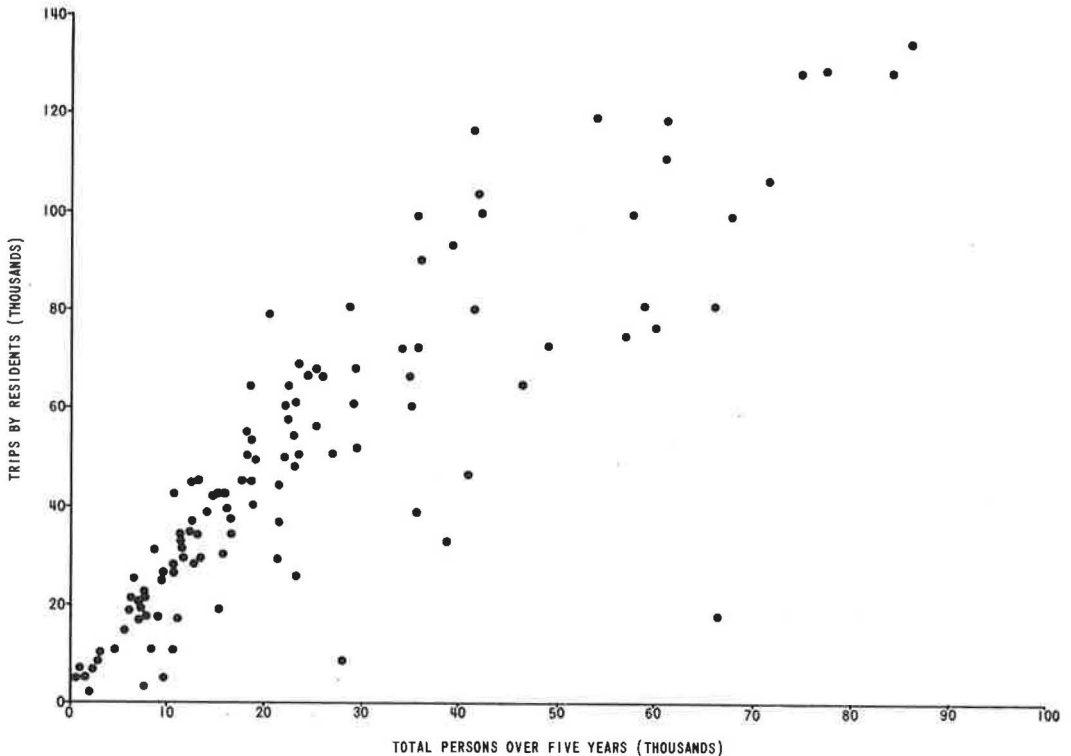


Figure 1. Relationship between trips per zone generated by residents and persons per zone.

available for the forecast year and because it is more specifically oriented to work-trip generation. Persons and automobiles were judged to be the best in terms of the established criteria outlined previously and were adopted for use in this study. A graphic representation of the relationship between these two variables and trip-making is shown in Figures 1 and 2.

Three levels of data aggregation were considered in this research. The Tri-State cordon area was divided into 158 districts, and these were further subdivided into 567 zones. Evaluations were performed using both districts and zones as the basic units of analysis; the third level consisted of individual households as points of observation. In all instances where comparisons were made, the relationships were scaled to the same level of aggregation with statistical indicators appropriately adjusted.

The study was limited by the availability of a single cross section of data. Because the basic goal of the research was to evaluate a set of forecasting techniques, the ideal input would naturally consist of two sets of comparable data for the same area reflecting different points in time. Such information unfortunately was not available. (These data will become more easily obtainable in the near future as many transportation studies complete updates of their surveys as part of the continuing planning process.) It was, however, considered essential to the study that the data on which the different methods were tested be independent of the data to which they were calibrated. Thus, lacking a temporal separation, we devised a geographical separation. This was done by separating the data, aggregated to the largest areal unit used in the study, into two discrete sets, each representative of the cordon-area environment. The relationships were developed on observed data contained in 85 of the 158 districts (referred to as cross section A), and were applied for the purposes of comparative evaluation to data from the 73 remaining districts (called cross section B). This means, for example, that on the zonal level only zones contained in the districts defining cross section A

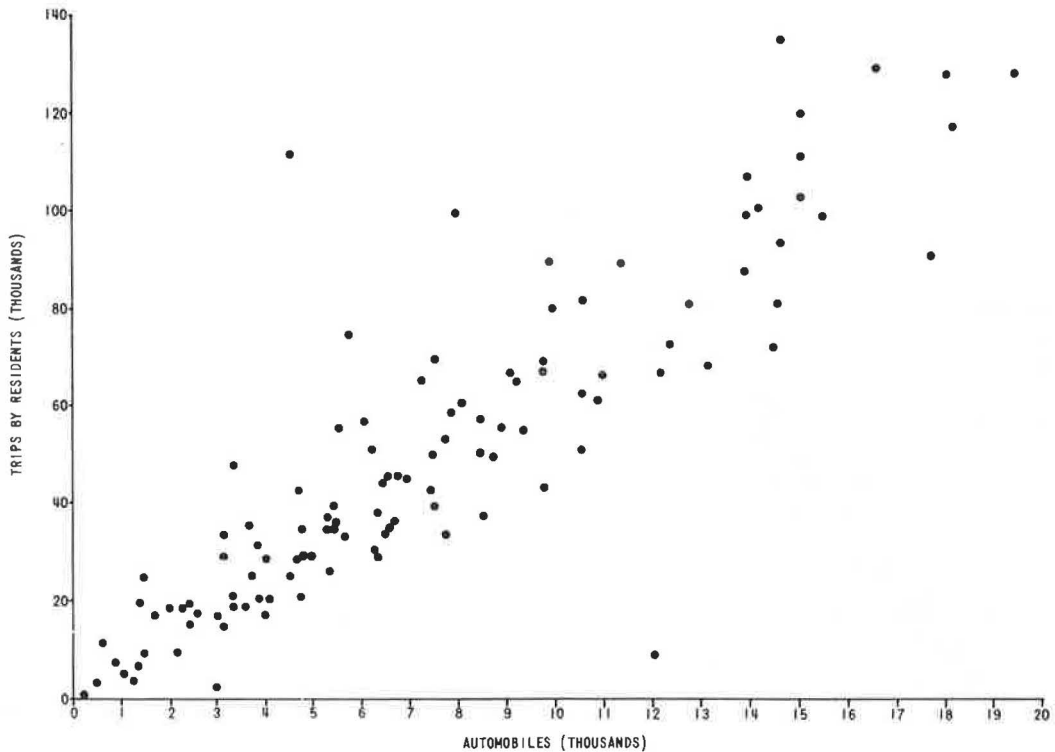


Figure 2. Relationship between trips per zone generated by residents and automobiles per zone.

were used to develop the zonal relationships. These relationships were then applied, for purpose of evaluation, to the zones in cross section B. Figure 3 illustrates the spatial distribution of the two data sections.

AGGREGATE TOTALS VS AGGREGATE RATES

In the course of performing trip-generation analysis, one must decide whether to deal with data in terms of aggregate totals or aggregate rates. (An example of an aggregate total is trips per zone or automobiles per zone, whereas an aggregate rate is average trips per household per zone, average trips per acre per zone, or average automobiles per household per zone.) There have been some studies that have mixed totals and rates within the same relationship, but such a practice is without a logical basis because aggregate totals are a function of the magnitude of the unit of aggregation, and rate variables are independent of size.

As an illustration of the misuse of rate variables in aggregate-total relationships, consider two zones, one twice the size of the other, in terms of trips generated. If both zones had identical automobile ownership rates, expressed in terms of automobiles per household, and this variable were used in an aggregate-total equation, the contribution of this variable in terms of predicted total trips per zone would be the same for each of the two zones, even though the actual volume of trips generated by one might be twice as large as the other.

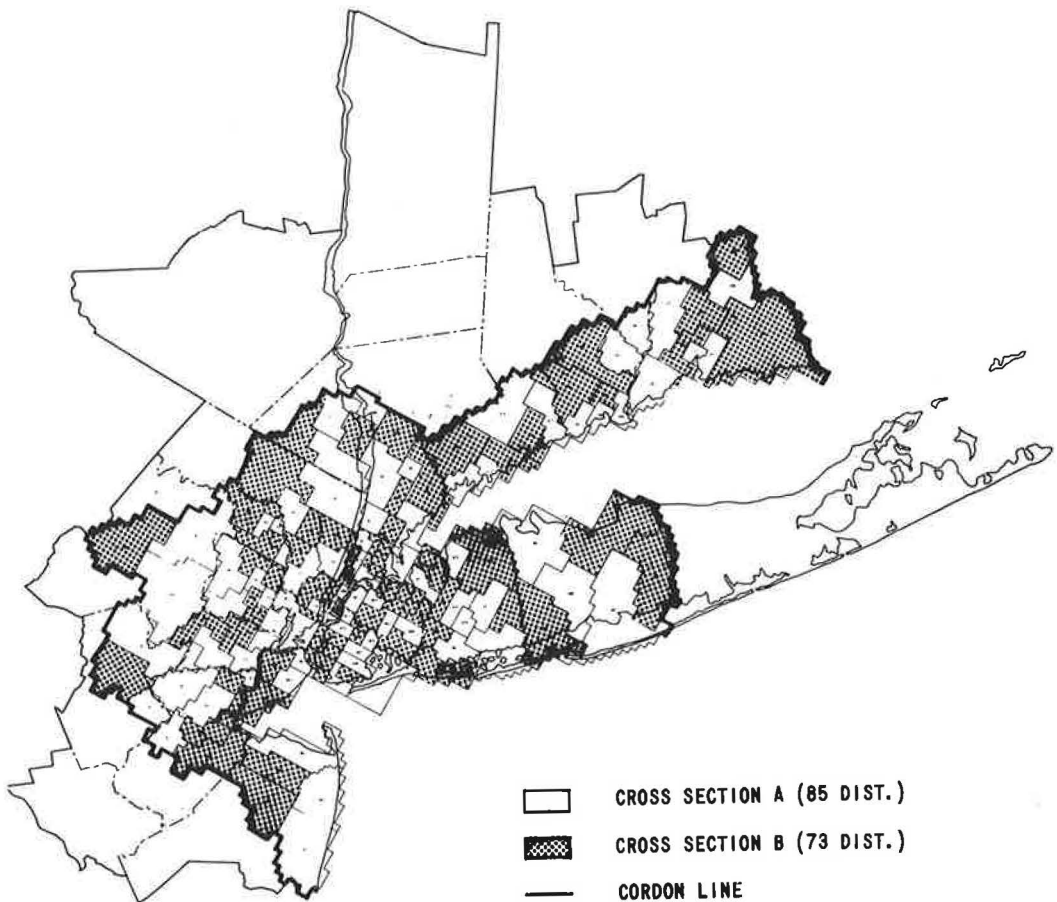


Figure 3. Spatial distribution of the two data sections.

TABLE 2
COMPARISON OF AGGREGATE-RATE AND AGGREGATE-TOTAL TRIP-GENERATION EQUATIONS

Analysis Unit	No. of Observations	Dependent Variable Y	Independent Variables		Equation	Constant/Mean	Standard Error/Mean	R ²
			X ₁	X ₂				
Zone	305	Trips per zone	Automobiles per zone	Persons per zone	Y = 4.343X ₁ +0.758X ₂ - 66	0.001	0.181	0.930
		Avg. trips per household per zone	Avg. automobiles per household per zone	Avg. person per household per zone	Y = 3.458X ₁ +2.054X ₂ -2.94	0.418	0.209	0.714

The objective of this section of the study is to compare the two techniques in terms of their suitability to trip-generation analyses. This part of the research proceeded along the following lines:

1. The data were compiled on a zonal basis in the form of aggregate totals and aggregate rates.
2. The data were divided into the two spatially independent sets mentioned previously, cross section A and cross section B.
3. Regression on equations were derived (using the BIO-MED O2R stepwise regression program) for both the zonal aggregate totals

$$\text{trips per zone} = f(\text{persons per zone and households per zone})$$

and the zonal aggregate rates

$$\text{trips per household per zone} = f(\text{persons per household per zone and automobiles per household per zone}).$$

These equations were developed using the data in cross section A.

4. The equations were applied to the zonal data regarding persons and automobiles from cross section B.
5. The relationships were evaluated in terms of how well they reproduced the zonal trip data from cross section B.
6. Several equation statistics were compared on a common basis.

The equations that were developed from applying each of the two methods are given with their associated statistics in Table 2. The most outstanding difference between the two equations is in the relative size of the constants. Although both are negative in sign, the aggregate-total constant is virtually insignificant, representing one-tenth of one percent of the mean trips per zone. The constant in the rate equation, however, is almost 42 percent of the mean household trip rate. In general, relatively large constants reduce the sensitivity of the expression to the variables that are supposed to reflect a causal relationship. In the case of large negative constants, the situation is aggravated by introducing the possibility of generating negative trip values for a zone with few persons or automobiles.

It is important to note that in the statistical measures that are given in Table 2, the coefficients of determination R² and the standard errors of the estimates, divided

by the mean of the dependent variable, are not strictly comparable between the two equations. This is simply because of the differences in the formulation of the variables constituting the relationships. To perform an independent evaluation on an equivalent basis, both equations were applied to the data set from cross section B, and the solution of the rate equation for each zone

TABLE 3
ADJUSTED STATISTICS FOR EQUATIONS IN TABLE 2

Equation	R ²	Standard Error/Mean
Aggregate total	0.929	0.187
Aggregate rate	0.888	0.234

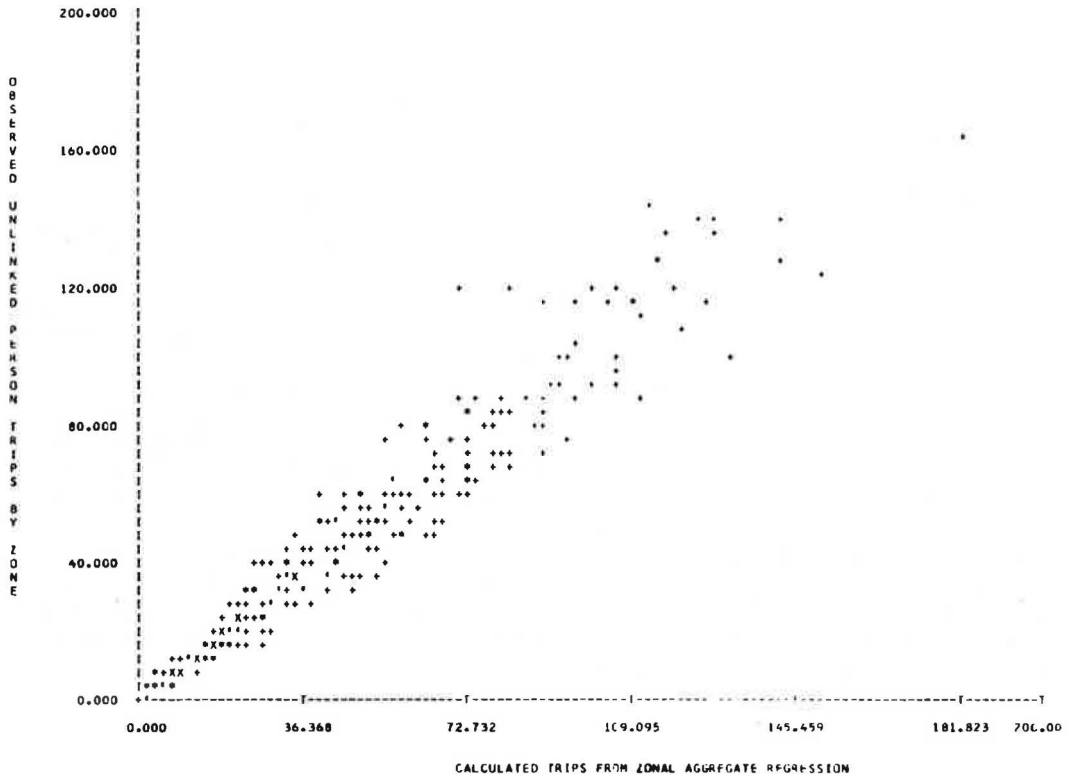


Figure 4. Comparison of observed vs predicted zonal trips generated by residents for zonal aggregate regression.

was multiplied by the number of households in the zone to yield zone totals, which were then comparable to the results obtained from the aggregate-total equation. The application of the rate equation in such a manner duplicates the process that would be followed in deriving actual forecasts from such an expression. Adjusted values of R^2 and standard errors were then computed, and the results are given in Table 3.

The rate equation, which explained only 71 percent of the variation in the data from which it was derived, accounts for 89 percent of the variation when used to estimate zonal totals from an independent data set. These adjusted statistics seem to indicate that the aggregate-total equation has a small advantage over the aggregate-rate equation. This is further evidenced by two additional analyses that were performed.

Figures 4 and 5 are plots of the total zonal trips derived from the home-interview survey vs those calculated from each of the two equations. Once again, the slight advantage of the aggregate-total equation is evidenced by a little less scatter about the 45 deg line in comparison to the results from the rate equation.

Finally, a root-mean-square (RMS) error analysis was performed by stratifying the zones in cross section B into 9 different size groups according to the volume of trips generated by residents. The root-mean-square error is a useful expression of the magnitude of differences between an array of estimated values and an array of actual values. For each predefined range of values, the RMS error is computed as follows:

$$RMS = \sqrt{\frac{\sum_{i=1}^N (\hat{Y}_i - Y_i)^2}{N}}$$

where

- \hat{Y}_i = estimated value,
- Y_i = actual value, and
- N = number of observations in the range.

Percentage of RMS error is simply the RMS error for a specific range divided by the mean value for that range (multiplied by 100).

The results, as shown in Figure 6, again reinforce the previous analyses that indicated that the aggregate-total equation is slightly superior. However, it does appear that the differences between the two equations, in terms of estimated zonal trips, become less marked at the upper volume range of trips.

The critical question arises as to which technique, aggregate total or aggregate rate, represents the better approach in trip generation. It should first be noted that only one type of rate was examined in this study. Rates such as trips per employee, trips per acre, trips per square foot, and the like are frequently used quite successfully in non-residential trip-generation analyses, but were not treated here. The comparison performed in this study of trips per household per zone vs trips per zone generated by residents indicates that the latter yields somewhat better results in terms of reproducing base-year data. The argument might be raised that the aggregate-total equation was favored because the comparisons were made on an aggregate-total basis; but this procedure, in fact, reflects the way in which these relationships are actually applied. The requirements placed on the trip-generation process, in terms of the phases to follow, are such that aggregated totals are the required output. (The trip-distribution and traffic-assignment process, which follow the generation of trips, require data that are aggregate totals related to some areal unit.)

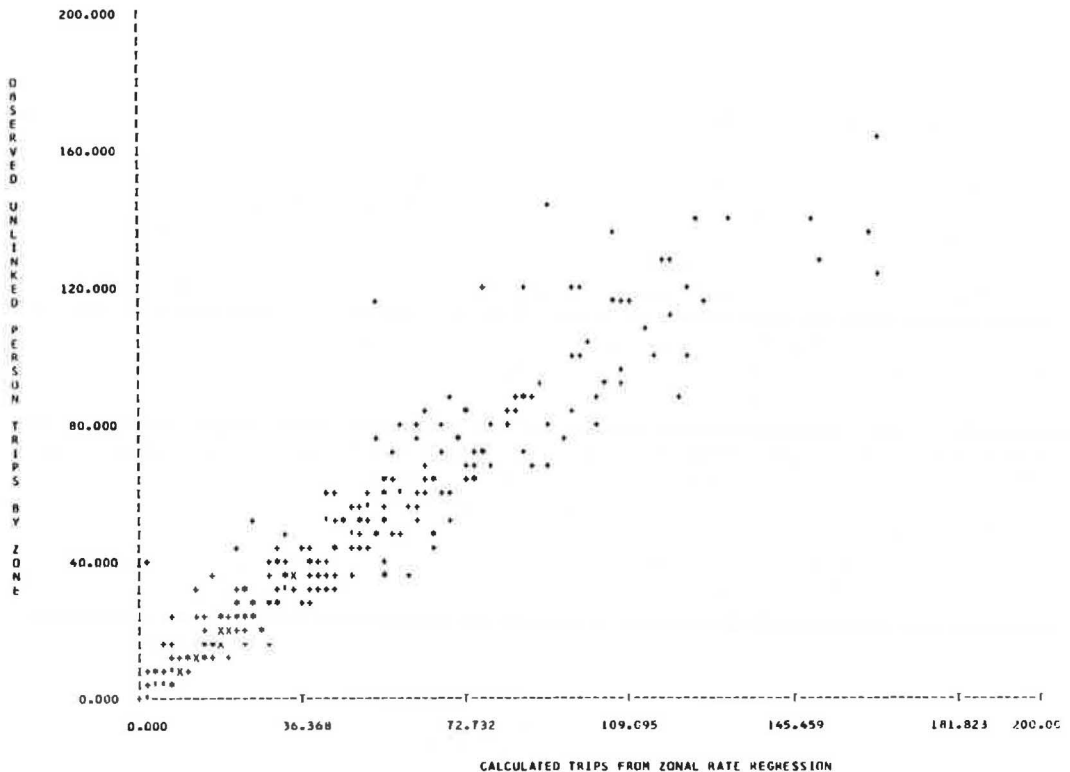


Figure 5. Comparison of observed vs predicted zonal trips generated by residents for zonal rate regression.

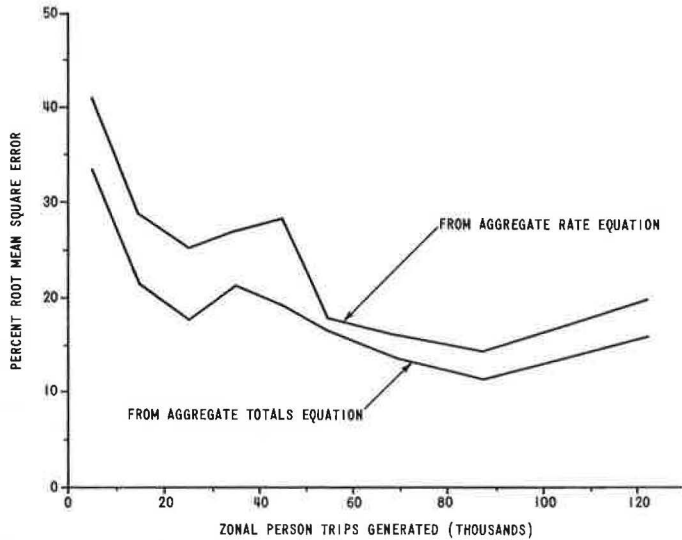


Figure 6. Root-mean-square error vs trip volume for two trip-generation techniques.

The primary advantage of dealing with rates rather than totals is the flexibility that is provided in terms of areal units of analysis. A rate relationship is not strictly tied to any particular geographic system of data aggregation, but an aggregate-total expression, primarily because of the equation constant, is tied to the zonal scheme on which it was developed, or to one that is extremely similar in terms of size and composition. Thus, if a zoning scheme is altered for some reason, or if forecasts are made for areal units that are dissimilar from those on which the equations were developed, the aggregate-rate approach must be considered.

Another practical advantage of dealing with rates is the convenience provided in working with numerical values that have an immediately recognizable meaning. For example, a median family income of \$15,000 for a particular area is more descriptive of the characteristics of the residents than an aggregate income of \$15 million.

In this analysis, both techniques dealt with aggregated data. Although the rate approach eliminated the effect of the areal aggregation configuration, the equation was developed using the same number of observations as the aggregate-total relationship. The use of rates, as described, sometimes lends the misleading notion that the analysis is being performed at a lesser degree of data aggregation. The implications of working at varying levels of aggregation are discussed in the next section.

THE EFFECTS OF DATA AGGREGATION IN TRIP GENERATION

Questions associated with the problems of data aggregation assume importance in the trip-generation phase of the transportation planning process. Almost without exception, the specified output of the trip-generation process is data that are aggregated to some areal unit. This is necessitated by the requirements of the trip-distribution and traffic-assignment techniques currently in use. There is little evidence to suggest that this condition will change in the near future. In addition, the inputs to trip-generation forecasts are generally aggregated data. Thus, the only real opportunity to work with disaggregated data is provided during the analysis stage, when the tools are being developed, because the survey data are usually available in an uncombined state.

The natural question that arises is if, during the forecasting process, both input and output are in an aggregated state, What is the real payoff in developing the tools on disaggregated data? Perhaps the following brief example can serve as a useful indication of the differences in results that can be obtained from an analysis based on uncombined data and one that uses the same data in a combined or aggregated state.

TABLE 4
INCOME AND TRIPS PER HOUSEHOLD PER ZONE
(Hypothetical Case)

Household	Zone 1		Zone 2		Zone 3		Zone 4	
	Income (thousands)	Trips	Income (thousands)	Trips	Income (thousands)	Trips	Income (thousands)	Trips
1	\$5.5	4.0	\$ 1.0	4.0	\$4.0	6.0	\$1.0	3.0
2	3.0	2.0	4.0	4.0	8.0	6.0	2.0	0.0
3	2.8	2.0	10.0	7.0			4.0	2.0
4	4.0	3.0	2.0	5.0				
5			8.0	5.0				
Average	\$3.8	3.5	\$ 5.0	5.0	\$6.0	6.0	\$2.3	1.7

Let us assume a hypothetical case in which we wish to predict average daily trips on the basis of household income. We have four analysis zones, and we are trying to develop a tool to use in forecasting trips per household for these zones at some point in the future such that aggregate totals may ultimately be derived. Our input data are arranged in matrix form (Table 4).

Figure 7 shows the difference in the types of relationships that can be derived from the same data at varying levels of aggregation. If we accept the uncombined data as the best representation of the relationships between income and trips, it is clear, by the difference between the two lines, that by aggregating the data into a small number of classes, we introduce a certain amount of bias.

The most significant consequence of aggregation is the loss of variation in the data that remains to be explained among the combined units of analysis. Naturally, this condition becomes more severe as the number of units into which the data are aggregated is diminished. In the hypothetical example, the total variation of trips per household in the basic data can be expressed as the sum of the squares of the deviation of each point from the overall mean trips per household. As given in Table 5, this number

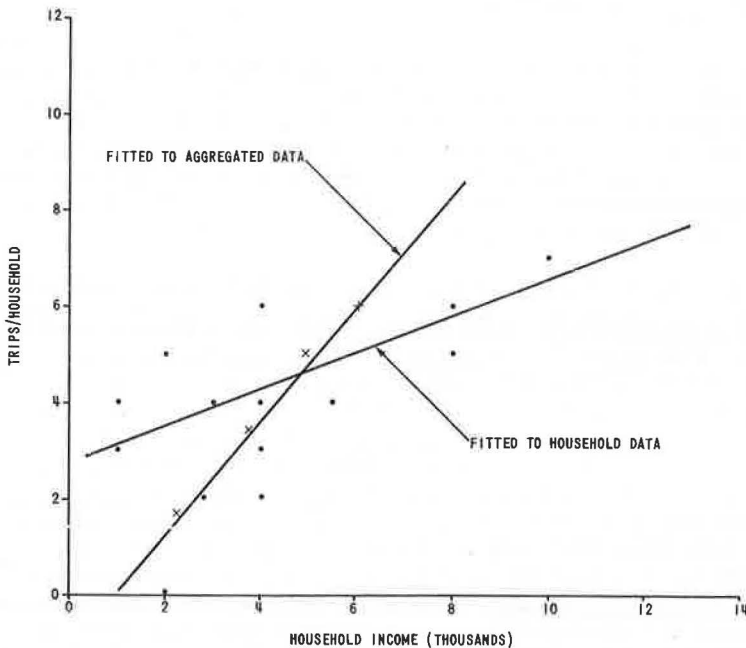


Figure 7. Relationships that may be derived from the same data at varying levels of aggregation (hypothetical case).

TABLE 5
 VARIATION WITHIN DATA AT DISAGGREGATED AND AGGREGATED LEVELS
 (Hypothetical Case)

Zone	Household	Total Trips per Household		Average Trips per Zone	
		T_h	$(T_h - \bar{T}_h)^2$	T_z	$(T_z - \bar{T}_z)^2$
1	1	4	0	3.50	0.30
	2	5	1		
	3	2	4		
	4	3	1		
2	1	4	0	5.00	0.91
	2	4	0		
	3	7	9		
	4	5	1		
	5	5	1		
3	1	6	4	6.00	3.82
	2	6	4		
4	1	3	1	1.70	5.55
	2	0	16		
	3	2	4		
Total	14	6	46	16.20	10.58
Mean		4		4.05	

turns out to be 46. The variation in the aggregated data in terms of sum of squares is 10.58. Over 75 percent of the variation in the original data is hidden within the four-zonal classification with only a relatively small fraction left to be explained between the zones.

Thus, the line placed through the aggregated zonal data, while explaining a high percentage of the zonal variation, explains relatively little of the total variation. This becomes less of a problem as the homogeneity of the data within each class increases. For example, if each of the households generated 3.5 trips in zone 1 (this assumes that fractional trips are a possibility), 5.0 trips in zone 2, 6.0 trips in zone 3, and 1.7 trips in zone 4, there would be no variation in trip-making rates within each zone, and all of the variation would be between zones. In such a case, the zone average would be an exact reflection of the intrazonal data. In a study of this type, however, the data within aggregate units are generally quite heterogeneous, as they are for this example.

The problem of relative-size differentials is also an important factor to consider. If, as is generally the case in an analysis of this type, each aggregated unit is given the same weight in the analysis, the results will tend to be biased as long as there are differences in the number of observations within the units. In our example, zone 3 carries the same weight as zone 1 in an aggregated relationship. Thus, each observation in zone 3 has twice as much influence as each has in zone 1. The results are again shown in Figure 7 where the slopes of the two lines differ sharply.

The problem arises as to which of the two relationships would be best suited in forecasting trips per household for these same zones. If it were known that there would be no relative changes in the travel and income characteristics of the households within each zone and that the relative size of the zones in terms of number of households would remain constant, the aggregated relationship could well be employed. But such situations are rarely the case. Given that changes with time are possible, it must be assumed that they will be more accurately reflected by the curve through the uncombined data, because these data are a better reflection of the true relationship between the two variables. There is no masking of variability resulting from the effect of averaging and no unequal weighting effect that is inherent by characteristic of averaged data.

Most trip-generation studies do not develop their tools on a disaggregated basis, but rely instead on the validity of the assumption that the relationships developed, using only a small fraction of the variation within a data set, are sufficiently representative of the actual relationship. The following analysis seeks to evaluate the consequences

of such an assumption by evaluating trip-generation regression equations developed at three levels of data aggregation. The analysis proceeded in the following manner:

1. Data relating to household trips, persons over five years of age per household, and automobiles per household were developed at the district, zonal, and household levels.
2. The data were separated into the two discrete sets described previously, cross section A and cross section B.
3. Equations were developed from the data in cross section A at each level of aggregation, i. e., the district rate,

$$\text{Trips per household per district} = f(\text{persons per household per district and automobiles per household per district})$$

the zonal rate,

$$\text{Trips per household per zone} = f(\text{persons per household per zone and automobiles per household per zone})$$

and the household rate.

$$\text{Trips per household} = f(\text{persons per household and automobiles per household})$$

4. The equations were applied to the data concerning persons and automobiles from cross section B.
5. Comparisons were made between the actual observed trip rates and those forecast from each of the three relationships.
6. Various statistical evaluations were made.

One need not enter into a discussion of the regression analysis in order to view the effects of data aggregation. Some of the biases that were discussed in the previous hypothetical example can be demonstrated by examining the distribution of values for one variable at the different aggregate levels. This was done for the trips-per-household variable, and the results, which are of interest, are given in Table 6.

It is immediately apparent that estimates of the mean number of trips per household in the Tri-State Study area differ significantly depending on the level of aggregation being used. The best estimate is naturally reflected by the raw, uncombined data, and turns out to be 5.87. The mean trip rates developed at the zonal and district level are 20 and 16 percent higher respectively. Although this is readily explainable, it is an excellent example of the possible errors introduced by data aggregation. (The higher mean trip rates are a result of the disproportionate weighting effect in the Tri-State area, discussed earlier, where almost half the total population is concentrated in a very small percent of the total area, i. e., New York City. Because the New York City trip rate is low and the New York City zones and districts are relatively few, the trip rates calculated on the basis of zonal and district averages are high.)

A look at the standard deviation of the person trip rates dramatically illustrates the sharp reduction in the variation within the data that resulted from averaging on the

basis of zones and districts. Although on the household level approximately two-thirds of the observations were contained within a range defined by roughly ± 100 percent of the mean, two-thirds of the zonal observations were clustered within only ± 39 percent of the mean value, and at the district level they were within 34 percent.

TABLE 6
MEANS AND STANDARD DEVIATIONS OF TRIPS PER HOUSEHOLD
FROM CROSS SECTION A AT VARYING LEVELS OF AGGREGATION

Analysis Unit	Mean Trips per Household	Standard Deviation	Mean Standard Deviation
Household	5.87	6.04	1.03
Zone	7.03	2.74	0.39
District	6.81	2.34	0.34

TABLE 7
COMPARISON OF RATE EQUATIONS DEVELOPED AT THREE LEVELS OF AGGREGATION

Analysis Unit	No. of Observations	Dependent Variable Y	Independent Variables		Equation	Constant/Mean	Standard Error/Mean	R ²
			X ₁	X ₂				
Household	5,032	Trips per household	Automobiles per household	Persons per household	Y = 3.169X ₁ +1.064X ₂ +0.242	0.041	0.856	0.309
Zone	305	Avg. trips per household per zone	Avg. automobiles per household per zone	Avg. persons per household per zone	Y = 3.458X ₁ +2.054X ₂ -2.94	0.418	0.209	0.714
District	85	Avg. trips per household per district	Avg. automobiles per household per district	Avg. persons per household per district	Y = 4.151X ₁ +1.733X ₂ -2.65	0.389	0.139	0.841

TABLE 8
ADJUSTED STATISTICS FOR EQUATIONS IN TABLE 7

Equation	R ²	Standard Error/Mean
Household	0.894	0.229
Zonal aggregate rate	0.888	0.234
District aggregate rate	0.888	0.235

The three regression equations that were developed from the data in cross section A are given in Table 7. All are rate relationships necessitated by the fact that the household equation can only be a rate relationship. The zonal equation is the identical aggregate-rate expression used in the preceding section of the paper. The standard errors and coefficients of determination given in Table 7 are not comparable, but only reflect how well the equations fit the data

from which they were derived. The R² for the household equation seems quite low until it is realized that this equation explained 31 percent of all the variation within the original data, while, for example, the district equation explained 84 percent of a relatively small fraction of the variation within data.

The equations were applied to the data from cross section B at the zonal level to achieve an equivalent basis for an independent evaluation. The explanation offered earlier in the paper for using zonal totals as a basis of comparison is repeated here: It is this quantity that, with little exception, is required as an end product of trip-generation studies. The equation statistics were adjusted, as given in Table 8, producing very interesting results. The district and zonal rate equations performed almost identically in terms of both the percentage of the variation in the zonal data, which was explained, and the relative amount of dispersion of the data about the regression line (standard error), but the household equation represented a slight improvement over each of these.

Root-mean-square errors were calculated for several trip-generation volume groups for each equation. The results are shown in Figure 8, with the RMS error expressed as a percentage of the mean of the volume group. The household equation does significantly better than either the zonal or the district aggregate equations in the lower trip volume ranges, and slightly better in the uppermost range where the average number of trips generated is approximately 120,000. In the range between 60,000 and 120,000 trips, the zonal and district equations are somewhat better in reproducing the observed zonal totals in cross section B.

Another view of the effect of data aggregation is provided by examining the coefficients of the variables in the regression. To do this required the derivation of standardized regression coefficients, called beta coefficients, so that the true relative contribution of each variable in the equation could be revealed. A beta coefficient is computed as follows:

$$B_i = b_i \frac{S_{x_i}}{S_y}$$

where

- B_i = beta coefficient,
 b_i = regression coefficient for the i th independent variable,
 Sx_i = standard deviation of the distribution of the i th independent variable, and
 Sy = standard deviation of the distribution of the dependent variable.

The meaning of the beta coefficient can be best explained by referring to the results of the analysis given in Table 9. In the equation developed at the household level, for example, a change of one standard deviation in automobiles per household would result in a corresponding change in trips per household of 0.398 of a standard deviation.

The interesting point to be made here is that the effect of intercorrelation among the independent variables changes at different levels of aggregation. One of the inherent assumptions in regression analysis is that the independent variables are not mutually correlated. When this condition is met, the regression coefficients of each variable are accurate reflections of the effect on the dependent variable of a unit change in the independent variable. Where there is a high degree of intercorrelation among independent variables, the effect of each variable is not clearly defined by the coefficients. This phenomenon is reflected by the entries in Table 9 that demonstrate (a) that the problem of intercorrelation becomes more severe at higher levels of aggregation, and (b) that, as a result, the importance to the equations of automobiles per household relative to persons per household, as reflected by the standardized beta coefficients, tends to become more highly aggregated.

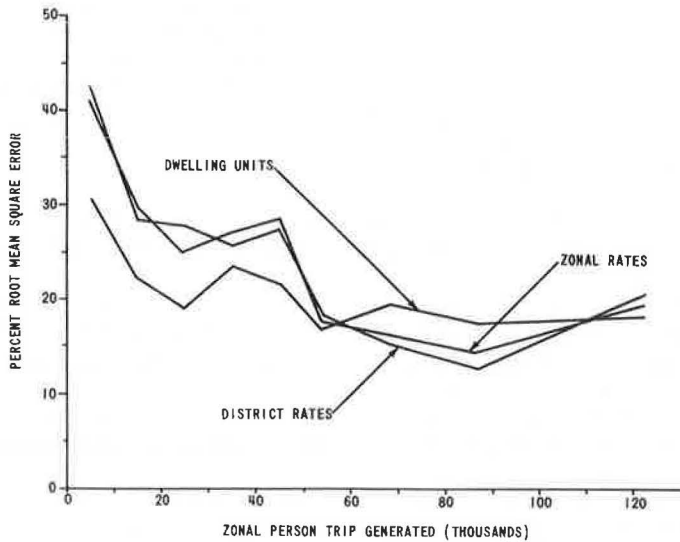


Figure 8. Root-mean-square error vs trip volume for dwelling unit analyses and for aggregate rate analyses.

TABLE 9
 EFFECT OF DATA AGGREGATION ON THE RELATIVE CONTRIBUTION
 OF THE INDEPENDENT VARIABLES

Aggregation	Beta Coefficients		B_A/D_P	Correlation Between Automobiles per Household and Persons per Household
	Automobiles per Household B_A	Persons per Household B_P		
Household	0.398	0.285	1.40	0.307
Zone	0.577	0.346	1.67	0.646
District	0.727	0.240	3.03	0.738

TABLE 10
ACTUAL TRIP RATES PER HOUSEHOLD

Automobiles per Household	Persons per Household				
	1	2	3	4	5+
0	1.23	2.54	3.74	4.35	5.01
1	2.92	4.98	6.99	7.88	10.23
2	4.25	6.93	10.39	12.46	14.67
3+	3.00	5.67	10.73	13.48	15.25

TABLE 11
PREDICTED TRIP RATES PER HOUSEHOLD FROM EQUATIONS
DEVELOPED AT DIFFERENT LEVELS OF AGGREGATION

Level of Aggregation	Automobiles per Household	Persons per Household				
		1	2	3	4	5
Household	0	1.31 (+6.50)	2.37 (-6.70)	3.43 (-8.90)	4.50 (+3.40)	5.56 (-2.50)
	1	4.48 (+53.40)	5.54 (-11.20)	6.60 (-5.60)	7.67 (-2.70)	8.73 (-13.30)
	2	7.64 (+79.80)	8.71 (+25.70)	9.77 (-6.00)	10.84 (-13.00)	11.90 (-13.30)
Zone	0	1.76 (+43.10)	3.81 (+33.30)	5.87 (+57.00)	7.92 (+82.10)	9.98 (+75.10)
	1	5.22 (+78.80)	7.27 (+4.60)	9.33 (+33.50)	11.38 (+44.40)	13.43 (+27.60)
	2	8.68 (+104.20)	10.73 (+54.80)	12.78 (+23.00)	14.84 (+19.10)	16.89 (+23.00)
District	0	1.47 (+19.50)	3.20 (+26.00)	4.93 (+31.80)	6.67 (+53.30)	8.40 (+47.40)
	1	5.62 (+92.50)	7.35 (+47.60)	9.09 (+30.00)	10.82 (+18.70)	12.55 (+9.50)
	2	9.77 (+129.90)	11.50 (+65.90)	13.24 (+27.40)	14.97 (+20.10)	16.70 (+21.60)

Note: Percentage of errors are in parentheses.

The equations were tested in terms of their ability to reproduce actual area-wide trip rates for various household types. The households in the study area were cross-classified according to the number of automobiles available and the number of persons per household. The area-wide average trip rates are given in Table 10 for each combination. Each of the three equations was then applied to each combination of automobiles and persons per household. The results, given in Table 11, clearly indicate the superiority of the household equation in predicting trip rates for data independent of areal aggregate units.

SUMMARY AND CONCLUSIONS

Two aspects of trip-generation analyses were studied in this research. The first dealt with the comparison between aggregate rates and aggregate totals, and the second with the comparison of aggregate and disaggregate trip-generation procedures.

In a comparison of (aggregate) rates vs totals, the results were calibrated and tested by using two data sets, in an effort to avoid the typical bias of evaluation caused by the use of a common data base. In addition, because rates and totals could not be directly compared, the results were standardized to a common basis for evaluation. A comparison of the two techniques produced evidence that the aggregate total equation has a slight statistical advantage over the aggregate-rate equation by virtue of such

tests as standard error of estimate, coefficient of determination, and the root-mean-square error. More significantly, however, the rate equation offers more flexibility and efficiency in analyzing the data, because it is not tied to the data scheme to which it was developed. It is recommended that aggregate rates be employed rather than aggregate totals because of this flexibility feature, for it is not unusual for analyses to be made on zonal schemes that are somewhat different from the one in which the equations were developed; and, more significantly, the zonal system for forecasting procedures may be quite different from the one utilized for equation calibration.

In the study of data aggregation, the research was directed at this primary question: Should aggregated data be utilized in trip-generation techniques or should households be treated as disaggregated units? It was noted that studies of trip generation often use aggregated data because it is assumed that the average zonal figures reflect the characteristics of the (composite) individual constituents of the zone. In many instances, however, the aggregated data mask the true variability of the data and do not represent the actual meaning of the data. On the other hand, disaggregate data limit the number and type of variables that may be employed and eliminate areal descriptions such as residential density (persons per square mile) and median household income. In addition, many of the data outputs are required on an aggregate zonal basis for trip-assignment purposes such that the disaggregate forecasts would have to be summed to yield meaningful results.

Statistical techniques were employed to measure and evaluate on a common basis the aggregate trip-generation procedures vs the disaggregate procedures. The disaggregate equations produced slightly better results than either of the aggregate equations (zones and districts) as evaluated by the standard error of estimate and the correlation coefficient. The most significant differences were found in testing the procedures for their capability for reproducing area-wide trip rates by household type. The household (disaggregate) equation produced a much lower magnitude of error when compared to the aggregate procedures. It is recommended that household disaggregate equations be utilized in trip-generation analyses, especially when proxy (disaggregate) variables may be derived for areal descriptions. Disaggregate equations have a more logical basis for producing trip-generation results; they represent the true correlation and variability between the variables and they also seem to produce slightly better results in synthesizing trip-generation characteristics than do aggregate equations.

The authors have studied the most commonly used trip-generation procedures, those utilizing multiple linear regression equations. The procedures have been viewed on a common basis for the logic and efficiency of synthesizing trip-generation results. Recommendations are made for (a) the use of aggregate rates as opposed to aggregate totals when aggregate data must be used and (b) the use of disaggregate household data as opposed to aggregate zonal or district data. For a total evaluation of trip-generation procedures, the procedures must be measured for efficiency of synthesizing present-day results as well as for relative stability over time. As data sources become available over two points in time, it is recommended that all of the suggested methods of trip generation be reevaluated and studied on a common basis utilizing the statistical measures suggested in this research paper.

Multiple-Regression Analysis of Household Trip Generation—A Critique

GERALD M. McCARTHY, Rhode Island Statewide Comprehensive Transportation and Land Use Planning Program

Multiple-regression analysis of trip generation based on data aggregated to the zonal-average level is discussed with respect to its descriptive and predictive accuracy. The validity of the major assumptions underlying this methodology is examined. Certain statistical characteristics of the zone sampling distributions are examined in order to determine (a) the accuracy and reliability of the mean as a representative measure of the individual household trip-generation rates and socioeconomic characteristics and (b) the relative homogeneity of analysis zones with respect to household trip-generation rates and socioeconomic characteristics. Sources of variation in household trip-generation rates and socioeconomic characteristics are analyzed to evaluate the effect of data aggregation on the explanative ability of trip-generation equations.

The zone-sampling distributions exhibited a considerable degree of dispersion and skewness, rather than normality, implying that zonal averages are not truly representative of the individual household traits. The majority of the residential zones were relatively heterogeneous with respect to household trip-generation rates and socioeconomic characteristics. This finding refutes the validity of the commonly accepted assumption of zonal homogeneity. The aggregation of individual household trip-generation rates to zonal averages left only a small percentage of variation for use in fitting the regression equation, thereby reducing the explanative ability of the resulting equation. When multiple-regression analysis preceded aggregation of the data, the resulting trip-generation equation explained a greater percentage of the total variation in the dependent variable without sacrificing any significant degree of accuracy in describing existing data.

•THE PRESENT STATE of transportation planning technology is characterized by relatively sophisticated statistical techniques. The rapid refinement in transportation planning technology, from its initial rule-of-thumb approach to its present state, can be attributed largely to the impetus provided by increasing travel desires and the resulting financial involvement of the federal government in the nation's transportation problems.

The present transportation planning process consists of four integrated steps: (a) inventories, (b) analysis of existing conditions and calibration of forecasting techniques, (c) forecast, and (d) systems analysis. The calibration of forecasting techniques involves the development of trip-generation and trip-distribution models. Because the accuracy of the results of the trip-distribution models is contingent on the development of accurate trip-generation information, it is not surprising that increased attention is being given to the trip-generation phase of transportation planning.

One limitation that has plagued household trip-generation methodology for some time is the inability of trip-generation equations developed for one city to duplicate accurately the zonal trip-generation data from another city. This limitation is one of the primary reasons for the need to conduct rather extensive and costly origin-destination surveys in each city or area for which future travel patterns are to be forecast. This lack of general applicability of household trip-generation equations is thought to be a result of the peculiarities of the different study areas, but sufficient research has not been focused on the present household trip-generation equation methodology to rule out the possibility of some basic fallacies in the statistical logic of the methodology.

The ease with which trip-generation equations can be developed using "canned" multiple-regression programs has resulted in widespread use of this technique. Too often, however, insufficient consideration has been given to the statistical characteristics of the zonal data being used in the multiple-regression analysis. As early as 1950, Robinson (3) pointed out the incorrectness of attempting to explain the behavior of individuals based on ecological correlations. Robinson distinguishes ecological correlations from individual correlations by defining an individual correlation as one in which the statistical object is indivisible, and an ecological correlation as one in which the statistical object is a group.

In developing household trip-generation equations from zonal averages, ecological correlations are being used to explain the trip-making behavior of individual households. Unfortunately, the incorrectness of such a procedure can be disguised in the results of the multiple-regression equation because of ecological fallacies resulting from the use of aggregated data; the equation will appear to describe rather accurately the existing zonal trip-generation data. However, descriptiveness is only one criterion of model-building. Sound statistical reasoning is also necessary if the resulting model or equation is to be truly predictive.

The hypothesis of this discussion is that the use of zonally aggregated data, such as the zonal mean, in the development of household trip-generation equations is statistically incorrect and results in a deceptively large degree of association between the independent and dependent variables in the multiple-regression equation. Furthermore, zonal aggregation obscures the true relationships between household trip-generation rates and household socioeconomic characteristics making it difficult, if not impossible, to develop general trip-generation equations—those that would be applicable to more than one study area.

The preceding hypothesis was tested by using household trip-generation and socioeconomic data from an origin-destination survey conducted in Raleigh, North Carolina, to examine the statistical characteristics of the zone samples. This, in turn, led to conclusions regarding (a) the accuracy of the zonal mean as a measure of household trip-generation rates and socioeconomic characteristics, (b) the reliability of the zone sample mean as an estimate of the zone population mean, and (c) the homogeneity of the zones as accepted by present trip-generation methodology.

An analysis of variance of household trip-generation rates and certain socioeconomic characteristics enabled the identification of the major sources of variance in the independent and dependent variables used in the multiple-regression trip-generation equation—a matter of considerable significance in the justification of the use of zonal averages in multiple-regression trip-generation equations.

The calculation of simple correlation matrices, for household trip-generation rates and selected socioeconomic characteristics based on both zonal averages and individual household data, enabled a determination of the degree to which the use of zonal averages obscures the true relationship between household trip-generation rates and household socioeconomic characteristics.

Finally, the development of multiple-regression trip-generation equations, from zonally aggregated data and also from individual household data, enabled a comparison of the accuracy of the resulting equations in explaining the variations in the dependent variable and also the accuracy of the equations to duplicate existing zonal trip-generation data. Furthermore, an analysis of the equations enabled certain tentative conclusions to be made regarding the predictive ability of the equations and also their probable general applicability.

DATA COLLECTION

The trip-generation data used in this study were obtained from the home-interview survey of the Raleigh Urban Area Thoroughfare Study conducted by Harland Bartholomew and Associates for the North Carolina State Highway Commission in cooperation with the U. S. Bureau of Public Roads. The home-interview survey covered a one-in-eight dwelling-unit sample of the entire urban area and was conducted in accordance with procedures recommended by the Bureau of Public Roads. The area within the external cordon line was subdivided into approximately 300 traffic analysis subzones, of which 184 were residential zones.

ACCURACY OF THE ZONAL MEAN

The underlying assumption justifying the use of the zonal mean in household trip-generation analysis is that it is reasonably representative of the trip-making and socioeconomic characteristics of individual households within the zone. This assumption of representativeness implies that the mean is the value around which the zone-sample distribution is centered, further implying normality of the distribution. In addition to the requirement of centrality of the mean is the requirement that the individual households should be reasonably homogeneous in order for the mean to be truly representative of the zonal data. Janes (8) recognized the importance of homogeneity with respect to household traits if zonal averages were to be correlated with volumes of traffic generation. He agreed that zonal averages would not be representative of the whole set of households if the zones were not homogeneous with respect to the household traits measured by those zonal averages.

The validity of these assumptions, with regard to household home-based trip-generation rates and selected socioeconomic characteristics, was investigated by examining certain descriptive statistical characteristics of the zone sampling distributions. The Bureau of Public Roads (10) pointed out that such investigations had long been overlooked in trip-generation research.

Zonal Household Trip-Generation Rates

For the examination of the statistical characteristics of the zone sampling distributions, 29 zones were initially selected at random from the 184 residential origin-destination zones included in the original home-interview survey. (For the purpose of simplicity, the term zone is used instead of subzone throughout this study. In the original origin-destination survey, the subzone was used to describe a further subdivision of an origin-destination survey zone.) Descriptive statistics, including the mean and standard deviation, were calculated from the data for each of the 29 zones. The sampling distributions for total home-based trips per household were plotted in the form of a frequency histogram for 6 of the 29 randomly selected zones. One zone was selected from each of the 6 study area sectors. (In the origin-destination survey, the Raleigh urban area was divided into six geographical sectors, which were further subdivided into districts and subzones.)

An analysis of these frequency histograms, two of which are shown in Figure 1, indicated two statistical characteristics of zone sample distributions that have a significant bearing on the validity of the assumption of representativeness of zonal averages. The shape of the histograms indicates that the zone sampling distributions are skewed rather than normal, and that the spread or dispersion of the distribution along the x-axis implies a certain degree of nonhomogeneity within the zones.

A determination was made of the degree to which the zonal average was the central value around which the zonal data were grouped by comparing the mean with the median value of the distribution. If the zonal data were truly centered around the mean, there would be no significant difference between the mean and the median.

Table 1 gives the degree to which the zonal average is a central value around which are grouped total home-based trip-generation rates for individual households within a particular zone. These data provide a basis for questioning the value of the zonal average as a measure of the central location of the individual household, total home-based

TABLE 1

ZONAL AVERAGE DEVIATION FROM CENTRAL VALUE OF INDIVIDUAL HOUSEHOLD TOTAL HOME-BASED TRIP-GENERATION RATES

Zone No.	Median Value, X_{50}	\bar{X}	Household With Total Home-Based Trip-Generation Rates $\leq \bar{X}$ (percent)	Household With Total Home-Based Trip-Generation Rates $\geq \bar{X}$ (percent)
1110	5.8	7.9	65.5	34.5
2162	6.8	7.3	56.0	44.0
3620	4.6	5.7	66.0	34.0
4240	6.6	8.1	57.0	43.0
5041	5.6	7.3	65.0	35.0
6052	4.9	6.0	63.0	37.0

trip-generation rates. This, in turn, raises doubt as to the zonal average's representativeness of individual household data.

Zonal Household Socioeconomic Characteristics

The statistical characteristics of the zonal sampling distribution of automobile ownership and family size per household were calculated, and the sampling distributions for household automobile owner-

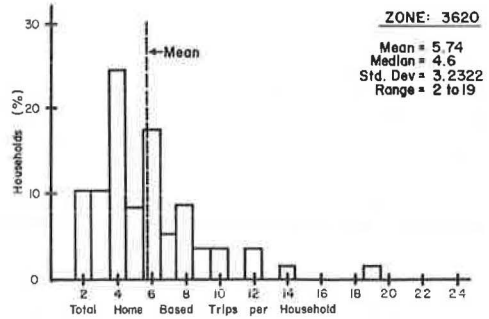
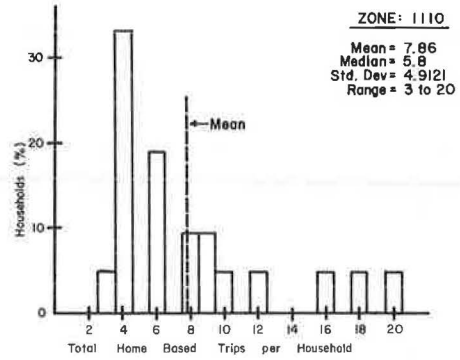


Figure 1. Total home-based trip-generation rate, zone sample distribution.

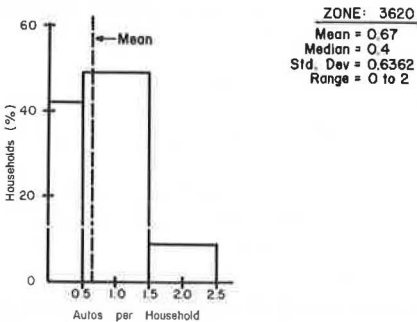
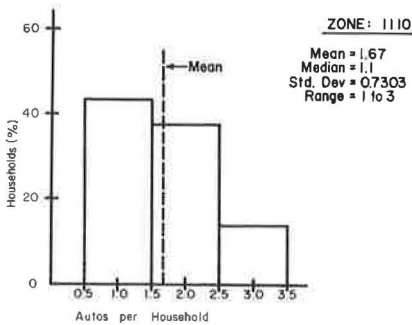


Figure 2. Household automobile ownership, zone sample distribution.

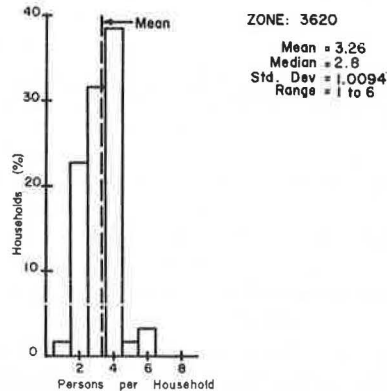
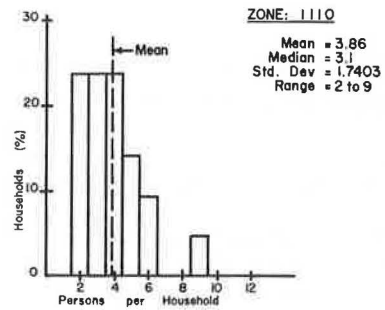


Figure 3. Household family size, zone sample distribution.

TABLE 2

ZONAL AVERAGE DEVIATION FROM CENTRAL VALUE OF INDIVIDUAL HOUSEHOLD CHARACTERISTICS OF AUTOMOBILE OWNERSHIP

Zone No.	Median Value, X_{50}	\bar{X}	Household With Automobile Ownership $\leq \bar{X}$ (percent)	Household With Automobile Ownership $\geq \bar{X}$ (percent)
1110	1.1	1.7	73.0	27.0
2162	1.4	1.7	74.0	26.0
3620	0.4	0.7	75.0	25.0
4240	0.9	1.5	67.0	33.0
5041	0.7	1.3	78.0	22.0
6052	0.8	1.3	73.0	27.0

TABLE 3

ZONAL AVERAGE DEVIATION FROM CENTRAL VALUE OF INDIVIDUAL HOUSEHOLD CHARACTERISTICS OF FAMILY SIZE

Zone No.	Median Value, X_{50}	\bar{X}	Household With Family Size $\leq \bar{X}$ (percent)	Household With Family Size $\geq \bar{X}$ (percent)
1110	3.1	3.9	69.0	31.0
2162	3.2	3.5	64.0	36.0
3620	2.8	3.3	68.0	32.0
4240	2.3	2.9	65.0	35.0
5041	2.2	2.8	68.0	32.0
6052	3.9	4.4	68.0	32.0

ship and family size were plotted in the form of frequency histograms for 6 of the initial 29 randomly selected zones.

Histograms for two of the zones, shown in Figures 2 and 3, exhibit some degree of skewness rather than normality as did the finding with respect to the trip-generation rate sampling distribution. However, the spread or dispersion of the distributions along the x-axis is not as large for the household characteristics of family size and automobile ownership as it was for household total home-based trip-generation rates, implying that zones may be more homogeneous with respect to automobile ownership and family size than they are with respect to total home-based trip-generation rates.

Tables 2 and 3 give the degree to which the zonal average deviates from the central value around which are grouped the individual household characteristics of automobile ownership and family size. These data indicate that the zonal averages for household automobile ownership and family size deviate to some degree from the central location of the zonal data, thus providing a basis for questioning the validity of the assumption of the representativeness of the zonal average.

THE RATIONALE OF AGGREGATION

The basic reason for aggregating origin-destination survey data in trip-generation methodology is that "... enough behavior must be aggregated to have statistically stable data and to discern consistent group travel behavior patterns" (10, p. 55). The same study indicates that, although the individual is, in essence, the basic trip-making unit, the magnitude of the unexplained variation that exists in the individual's travel behavior makes trip-generation analysis impractical at this level. Thus, by aggregating travel behavior to the household level, the Bureau of Public Roads (10) concluded that there is an increase in the statistical stability of data results. However, trip-generation methodology has for some time utilized the origin-destination survey zone as the level of aggregation from which trip-generation equations are developed.

The basic hypothesis of the present study is that the use of zonally aggregated household data results in the development of inaccurate predictive trip-generation equations. The findings of Robinson (3) imply that such inaccuracies are attributable to the use of aggregated data from analysis areas that have a considerable degree of nonhomogeneity. Janes states more specifically the characteristics of the basic data that contribute to inaccuracies in the trip-generation equations developed by multiple-regression techniques (8, p. 14): "Two conditions which may increase the possibility of ecological fallacy concern (1) the heterogeneity of the ecological units, in this case the survey zones, and (2) the source of variance in the dependent variables in the ecological correlation."

Therefore, this section of the discussion will examine, in greater detail, the implications of zonal aggregation with respect to zonal homogeneity and the sources of variation in the variables used in the trip-generation equation.

Zonal Homogeneity

The primary basis for utilizing zonally aggregated data has been the assumption that geographical proximity results in similarity of households with respect to trip-making and socioeconomic characteristics. The validity of this assumption is examined in the following.

Household Trip-Generation Rates—The degree to which the data in any particular zone are more or less heterogeneous than the data of the total area, from which that particular zone was formed, can be considered as a measure of the relative homogeneity of the zone. This relative homogeneity can be measured by comparing the zonal sample standard deviation to the total area sample standard deviation. This ratio, a measure of the relative zonal homogeneity, will be referred to hereafter as the relative zonal homogeneity index (RZHI). Cumulative frequencies of the RZHI for the household trip-generation rates were tabulated.

The conclusion to be drawn from an RZHI of 1.00 is that the particular zone is no more homogeneous, with respect to the household trait whose zone sampling distribution standard deviation was calculated, than the entire study area. However, because we are, in effect, making conclusions about the homogeneity of the zone population while measuring the homogeneity of the zone sample, the sampling variation in the standard deviation must be taken into account.

The sampling variation in the zone sample standard deviation was taken into account by determining, for each of the household trip-generation rate stratifications, a critical value of the RZHI. By definition, this critical value is such that any zone having a RZHI that is equal or greater in value will be considered to be significantly heterogeneous in regard to the household trait whose sampling distribution standard deviation is being considered. The mathematical logic on which this critical value is based is given in the Appendix.

Table 4 gives the critical relative zonal homogeneity index (CRZHI) calculated for each of the household trip-generation rate stratifications. Table 4 also gives the summaries of the results of applying these critical values to the tabulated cumulative frequencies for the household trip-generation rate RZHI's.

Based on the summary of the zonal homogeneity analysis given in Table 4, it can be concluded that a large enough number of zones show a significant degree of nonhomogeneity, and refute the assumption of zonal homogeneity, at least with respect to household trip-generation rates.

Household Socioeconomic Characteristics—Cumulative frequencies of the ratios of zone sample standard deviations to total study area sample standard deviations with respect to household socioeconomic characteristics were also tabulated. Based on the summary of the analysis of zonal homogeneity, as given in Table 5, it can be concluded that a large enough number of zones show a significant degree of heterogeneity to refute the assumption of zonal homogeneity, with respect to household socioeconomic characteristics.

Although the results of this analysis of zonal homogeneity, with respect to household socioeconomic characteristics, support the findings of a similar, but somewhat more

TABLE 4

CRITICAL RELATIVE ZONAL HOMOGENEITY INDEXES
AND SUMMARY OF ZONAL HOMOGENEITY ANALYSIS
FOR HOUSEHOLD TRIP-GENERATION RATES

Trip-Generation Rate Stratification	CRZHI	Zones Having RZHI \geq CRZHI (percent)
Total home-based	0.749	63.0
Total automobile-driver home-based	0.749	52.7
Automobile-driver home-based work	0.731	58.6
Total home-based work	0.749	75.8

Note: CRZHI = critical relative zonal homogeneity index; and RZHI = relative zonal homogeneity index.

TABLE 5

CRITICAL RELATIVE ZONAL HOMOGENEITY INDEXES
AND SUMMARY OF ZONAL HOMOGENEITY ANALYSIS
FOR HOUSEHOLD SOCIOECONOMIC CHARACTERISTICS

Socioeconomic Characteristics	CRZHI	Zones Having RZHI \geq CRZHI (percent)
Automobile ownership	0.755	55.4
Family size	0.755	72.9
No. of persons aged 5 and over	0.754	68.9
Income level	0.754	41.4

TABLE 6
SUMMARY OF THE ANALYSIS OF VARIANCE FOR TOTAL HOME-BASED
TRIPS PER HOUSEHOLD

Level of Aggregation	No. of Analysis Areas	TSS	WSS	Percent TSS	BSS	Percent TSS
Section	6	118,047	115,505	97.84	2,541	2.16
District	19	118,047	111,782	94.69	6,264	5.31
Zone	184	118,047	103,579	87.74	14,468	12.26
Household	4,159	118,047	0	0.00	118,047	100.00

Note: TSS = total sum of squares (total variation); WSS = within sum of squares (that portion of total variation existing within the analysis area); and BSS = between sum of squares (that portion of total variation existing between analysis areas).

limited, analysis reported by Fleet and Robertson (7), they are in direct disagreement with the findings reported by Janes (8). Fleet and Robertson reported the results of their analysis of the frequency distribution of the homogeneity index for automobile ownership only. Furthermore, they did not establish a critical value for the homogeneity index. Using a Gutman-type scale analysis procedure, Janes, on the other hand, concluded that the zones in the Champaign-Urbana, Illinois, study area were homogeneous with respect to occupation, make of automobile, and value of structure—socioeconomic variables that he considered to be pertinent in the relationship between household trip generation and household socioeconomic characteristics.

Sources of Variation

Because the multiple-regression analysis technique utilizes the variation between observations in developing the estimating equation, the amount of this between variation that is explained by the resulting multiple-regression equation is a measure of the predictive ability or reliability of the equation.

Household Trip-Generation Rates—The sources and amount of variation that are available at each level of aggregation were determined for utilization in the multiple-regression analysis. To do so, the total variation in total home-based trip-generation rates was analyzed using a one-way analysis of variance. This statistical technique separates or partitions the total variation associated with a variable, such as total home-based trip-generation rate, family size, or automobile ownership, into its component parts. Its pertinent components are the variation within analysis areas (levels of aggregation) and the variation between analysis areas.

The calculations for the analysis of variance were carried out on the Triangle University Computing Center's IBM 360 Model 75 computer using the TSAR system. TSAR (Tele-Storage-And-Retrieval) is a computer-oriented system developed and maintained by Duke University for storing, retrieving, processing, and analyzing data.

The results of the analysis of variance given in Table 6 indicate that, when zonally aggregated total home-based trip-generation rates are used in the multiple-regression trip-generation analysis, only 12.26 percent of the total variation in total home-based trip-generation rates is utilized. A further analysis of Table 6 shows that as the level of aggregation increases, the between sum of squares, expressed as a percentage of the total variation in total home-based trip-generation rates, decreases. (The between sum of squares is the variation in total home-based trip-generation rates between analysis areas.)

The significance of these findings is of particular importance with respect to the predictive ability of the trip-generation equations developed from zonally aggregated trip-generation data. It is reasonable to assume that as the percentage of the total variation utilized in the development of the trip-generation equation decreases, the predictive ability of the trip generation will decrease.

Household Socioeconomic Characteristics—An analysis of variance performed on household family-size and automobile-ownership data resulted in the summaries given in Tables 7 and 8. Although the analysis of variance summaries given in Tables 7 and

TABLE 7
SUMMARY OF THE ANALYSIS OF VARIANCE FOR HOUSEHOLD FAMILY SIZE

Level of Aggregation	No. of Analysis Areas	TSS	WSS	Percent TSS	BSS	Percent TSS
Section	6	11,423	11,055	96.78	367	3.22
District	19	11,423	10,449	91.47	974	8.53
Zone	184	11,423	8,988	78.69	2,434	21.31
Household	4,352	11,423	0	0.00	11,423	100.00

TABLE 8
SUMMARY OF THE ANALYSIS OF VARIANCE FOR HOUSEHOLD AUTOMOBILE OWNERSHIP

Level of Aggregation	No. of Analysis Areas	TSS	WSS	Percent TSS	BSS	Percent TSS
Section	6	3,004	2,705	90.04	299	9.96
District	19	3,004	2,413	80.33	590	19.67
Zone	184	3,004	2,007	66.80	997	33.20
Household	4,352	3,004	0	0.00	3,004	100.00

8 indicate that from 70 to 170 percent more of the total variation in the household characteristics of family size and automobile ownership is utilized in multiple-regression trip-generation analysis, there still remains a large percentage of the total variation that is not utilized in the multiple-regression analysis. This occurs when zonally aggregated, household family-size and automobile-ownership data are used, rather than when zonally aggregated, total home-based trip-generation rate data are used. The preceding findings, in addition to providing further evidence against the validity of the assumption of zonal homogeneity, imply that relationships or correlations developed between variables representing aggregated data will be inaccurate, because there is a considerably larger portion of variation within zones than between zones.

ECOLOGICAL FALLACIES RESULTING FROM THE USE OF AGGREGATED DATA

Household Trip-Generation Relationships

A determination was made of the effect that aggregating the home-interview survey data to the zonal average has on the relationships between certain household characteristics and total home-based trip-generation rates. This was done by determining the coefficients of correlation for individual household data and for data aggregated to the zonal-average level. The correlation coefficients for these two levels of aggregation are given in Tables 9 and 10. Table 11 summarizes these and other changes in

TABLE 9
ZONAL AVERAGE HOUSEHOLD DATA-HOUSEHOLD TRAIT CORRELATION MATRIX

Household Trait	Automobile Ownership	Family Size	No. of Persons Aged 5 and Over	Home-Based Trips per Household
Income level	0.8014	0.0736	0.0672	0.6154
Automobile ownership		0.1390	0.1266	0.6476
Family size			0.9374	0.4534
No. of persons aged 5 and over				0.4805

TABLE 10
INDIVIDUAL HOUSEHOLD DATA-HOUSEHOLD TRAIT CORRELATION MATRIX

Household Trait	Automobile Ownership	Family Size	No. of Persons Aged 5 and Over	Home-Based Trips per Household
Income level	0.5796	0.1930	0.2012	0.3800
Automobile ownership		0.2208	0.2375	0.4189
Family size			0.9072	0.4649
No. of persons aged 5 and over				0.5197

TABLE 11
PERCENTAGE OF CHANGE IN HOUSEHOLD TRAIT CORRELATIONS WHEN INDIVIDUAL HOUSEHOLD DATA ARE AGGREGATED TO THE ZONAL AVERAGE LEVEL

Household Trait	Automobile Ownership (percent)	Family Size (percent)	No. of Persons Aged 5 and Over (percent)	Home-Based Trips per Household (percent)
Income level	+38.2	-61.8	-66.6	+61.9
Automobile ownership		-37.0	-46.7	+54.1
Family size			+ 3.3	- 2.6
No. of persons aged 5 and over				- 7.5

household trait relationships resulting from the use of household data aggregated to the zonal-average level. These changes in correlation given in Table 11 emphasize the clouding effect that aggregation has on basic relationships between household travel and certain household socioeconomic characteristics as pointed out by Fleet and Robertson (7). The significance of these changes in basic relationships is that they have implications with respect to the validity and reliability of trip-generation equations developed from aggregated data.

Household Trip-Generation Equations

It has been shown that the relationships between certain household traits change when aggregated data rather than individual household data are used to develop the relationships. It is reasonable, therefore, to expect the trip-generation equations developed from aggregated data to exhibit some differences from those developed from individual household data. Table 12 gives the trip-generation equations developed from both individual household data and zonal average data.

Based on the coefficient of determination and the standard error of estimate (two standard statistical measures for evaluating multiple-regression equations), the trip-

TABLE 12
COMPARISON OF BASIC TRIP-GENERATION EQUATIONS DEVELOPED FROM NONAGGREGATED VS AGGREGATED DATA

Level of Aggregation	Trip-Generation Equation	No. of Observations	R ²	S _{y,x}	\bar{Y}	Percent S _{y,x}
Household	Home-based trips per household = (-1.0889 + 0.6394IL + 1.8328PFO + 1.4401AO - 0.1774FS)	4,158	0.38	4.18	7.49	55.81
Zonal average	Home-based trips per household = (-1.4169 + 0.7485IL + 1.6482PFO + 1.6849AO - 0.1846FS)	184	0.61	1.38	7.43	17.23

Note: R² = coefficient of determination; S_{y,x} = standard error of estimate; \bar{Y} = mean of the dependent variable; percent S_{y,x} = percentage of standard error of the estimate or S_{y,x}/ \bar{Y} ; IL = household income level; PFO = number of persons aged 5 and over; AO = household automobile ownership; and FS = household family size.

TABLE 13
COMPARISON OF ADJUSTED BASIC TRIP-GENERATION EQUATIONS

Level of Aggregation	Trip-Generation Equation	\bar{Y}	RMSE	Percent RMSE
Household	Home-based trips per household/zone = $N(-1.0889 + 0.6394IL + 1.8328PFO + 1.4401AO)$	168.28	30.82	18.31
Zonal average	Home-based trips per household/zone = $N(-1.4169 + 0.7485IL + 1.6482PFO + 1.6849AO)$	168.28	27.94	16.61

Note: RMSE = root-mean-square error; percent RMSE = percentage of root-mean-square error or $RMSE/\bar{Y}$; and \bar{Y} = mean of the dependent variable (total home-based trips per zone).

generation equation developed from zonal averages appears to be superior to the equation developed from the individual household data. However, this is not the case.

The analysis of the amount of variation that existed in household total home-based trip-generation rates (see Sources of Variance, p. 75) revealed that the within-zone variation accounted for approximately 87.7 percent of the total variation whereas the between-zone variation accounted for only 12.3 percent of the total variation when the household total home-based trip-generation data were aggregated to the zonal level. Therefore, when the coefficient of determination is adjusted to reflect the total amount of variation in total home-based trips per household that exists at the zonal-average level of aggregation, the value of R^2 drops to 7.5 percent (0.61×0.123).

It is evident, therefore, that the use of aggregated data has resulted in a deceptively high value for R^2 . A similar conclusion was reached by Fleet and Robertson (7) with respect to zonal home-based work-trip-generation equations developed from aggregated data.

Because R^2 , the coefficient of determination, measures the percentage of total variation in a dependent variable that is explained or accounted for by the combination of the

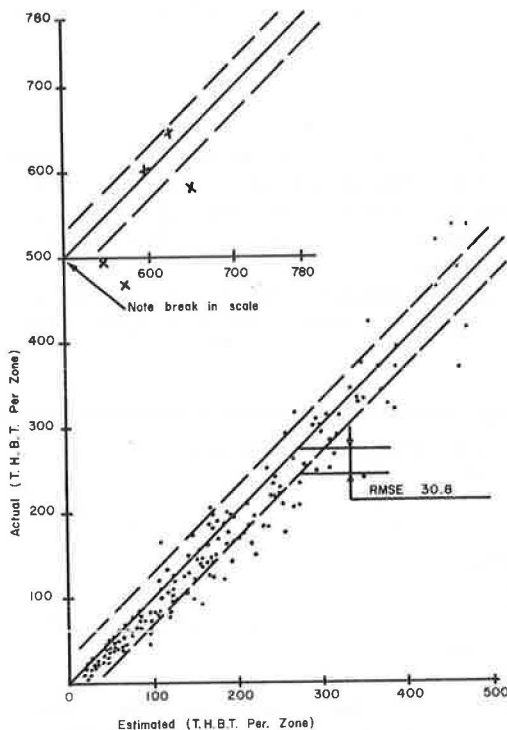


Figure 4. Actual vs estimated trips—trip-generation equation developed from individual household data.

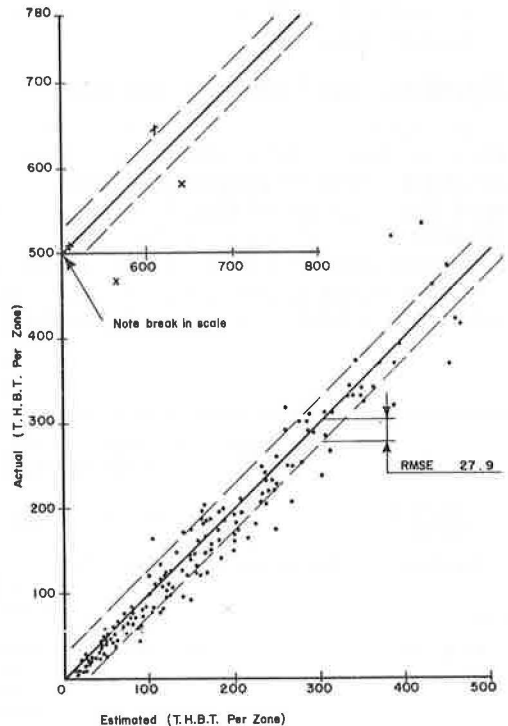


Figure 5. Actual vs estimated trips—trip-generation equation developed from individual household data aggregated to zonal averages.

independent variables included in the equation, the higher the value of R^2 , the greater will be the reliability of the association or relationship between the dependent and independent variables. Therefore, the inherent danger in such a deceptively high value of R^2 is that the trip-generation analyst could be deceived into thinking that the equation had a high degree of reliability with respect to the trip-generation relationship described, when in fact it did not.

Furthermore, there appears to be a considerable difference between the two equations in the percentages of the standard errors of the mean. However, an analysis of the results of estimates of the unexpanded home-interview trip data obtained from the two equations (as given in Table 13 and shown in Figs. 4 and 5) indicates that, in fact, the descriptive accuracy of the trip-generation equation developed from the nonaggregated data is as acceptable as that of the equation developed from the aggregated data. In fact, if family size were included in the equation (it could be argued that its F-value, 3.34, is not sufficiently less than the critical F-value, 3.84, to justify its exclusion) developed from the nonaggregated data, this equation would then provide a more accurate estimate of the unexpanded home-interview survey trip data than the equation developed from the data aggregated to the zonal-average level.

Family size was found to be more significant in the trip-generation equation developed from unaggregated data because of the fact that the correlation between family size and the number of persons five years old or older increased when the data were aggregated to the zonal-average level, thus disguising any distinction that might exist between those two variables.

The increase in correlation between family size and number of persons five years old or older caused by data aggregation resulted in a sufficient decrease in the significance of the variable (family size) in the trip-generation equation. The greater significance of family size in the trip-generation equation developed from nonaggregated data is indicated by its F-value of 3.39 as compared with its F-value of 0.21 in the zonal-average trip-generation equation.

These effects of the use of aggregated data on trip-generation equations, namely, the change in the basic trip-generation relationships, the reduction in the coefficient of determination, and the decrease in significance of certain variables resulting in their exclusion from the equation, raise doubts as to the predictive ability of trip-generation equations developed from zonally aggregated data. That these same effects may also be one of the major reasons for the failure of trip-generation equations to exhibit applicability in more than one city was first implied by Fleet, Stowers, and Swerdloff (6, p. 23).

SUMMARY

This study examined the validity of three major assumptions on which present multiple-regression trip-generation analysis methodology is based. The assumptions are as follows: (a) that the zone sample mean is an accurate representation of the traits of all the individual households in the zone, and the sample mean is a reliable estimate of the zone population mean; (b) that origin-destination survey zones are homogeneous with respect to household traits pertinent to trip-generation relationships; and (c) that valid trip-generation relationships and reliable household trip-generation equations can be developed from household trait data aggregated to the zonal-average level.

The zone sample mean is not representative of all of the households in the zone for the following reasons:

1. Zone sampling distributions are skewed rather than normal indicating the zone sample mean is not the central value around which the individual households are grouped.
2. There is a considerable amount of dispersion or heterogeneity within zones with respect to household traits that are pertinent to household trip-generation equations. The median value may be more representative of the zone sample distributions for those household traits whose zone sampling distributions show a high degree of dispersion, such as the household trip-generation distributions.

The sample means for socioeconomic household characteristics provide a more accurate estimate of their corresponding zone population means than do the zone sample

means for household trip-generation rates. However, stratification of household trip-generation rates increases the accuracy of the zone sample mean as an estimate of the zone population mean.

The heterogeneity of a considerable portion of the home-interview zones, as measured by the ratio of the standard deviation of the zone sampling distribution to that of the total study area sample distribution, is large enough to refute the validity of the assumption of zonal homogeneity commonly accepted in multiple-regression trip-generation analysis.

Aggregation of data to the zonal average level causes a major percentage to the total variation in individual household automobile ownership, family size, and total home-based trip-generation rates to be lost in the aggregation. Correlations between household socioeconomic characteristics and total home-based trip-generation rates developed from individual household data aggregated to the zonal average level differ considerably from those developed from nonaggregated individual household data.

The coefficient of determination for the total home-based trip-generation equation developed from zonally aggregated data is deceptively high because the data utilized in developing the basic trip-generation equation contain only 12.3 percent of the total variation existing in the individual household total home-based trip-generation rate data. Therefore trip-generation equations developed from individual household data aggregated to zonal averages will be based on invalid trip-generation relationships and, although they provide a reasonably accurate description of existing trip-generation rates, their predictive reliability is questionable because the amount of variation in the dependent variable that they explain is unreliably low.

Present multiple-regression, household trip-generation methodology should be modified so that multiple-regression analysis precedes aggregation of individual household data rather than the aggregation of data preceding multiple-regression analysis.

RECOMMENDATIONS FOR FURTHER STUDY

The present study points out some refinements in existing household trip-generation analysis methodology that can be expected to result in more reliable trip-generation equations with respect to their predictive ability and generality of application. However, this is considered to be only a first step toward the development of a methodology for developing household trip-generation equations that can be applied with a reasonable degree of accuracy to more than one study area.

Additional research should be carried out with respect to the assumption of linearity between dependent and independent variables in the multiple-regression trip-generation equation, and the use of the dummy variable technique should be investigated. The dummy variable technique for multiple-regression analysis was first explained by Suits (9) but has had surprisingly little utilization in trip-generation analysis. Therefore, additional household trip-generation research needs to be carried out utilizing this new technique.

REFERENCES

1. Traffic Assignment and Distribution for Small Urban Areas. U. S. Bureau of Public Roads, 1965.
2. Manual of Procedures for Home Interview Traffic Study. U. S. Bureau of Public Roads, 1954.
3. Robinson, W. S. Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, Vol. 15, 1950, pp. 351-357.
4. Oi, W. Y., and Shuldiner, P. W. *An Analysis of Urban Travel Demands*. Northwestern Univ. Press, Evanston, Ill., 1962.
5. Shuldiner, P. W. Trip Generation and the Home. *HRB Bull.* 347, 1962, pp. 40-59.
6. Fleet, C., Stowers, J., and Swerdloff, C. *Household Trip Production—Results of a Nationwide Survey*. U. S. Bureau of Public Roads, Highway Planning Technical Report 2, 1965, pp. 16-24.
7. Fleet, C., and Robertson, S. Trip Generation in the Transportation Planning Process. *Highway Research Record* 240, 1968, pp. 11-31.

8. Janes, R. W. Social Factors Associated With Traffic Generation in a Metropolitan Area of 75,000 Population. Univ. of Illinois Engr. Exp. Station, Bull. 490, 1967.
9. Suits, D. B. Use of Dummy Variables in Regression Equations. Jour. Am. Statistical Assn., Vol. 52, 1957, pp. 548-551.
10. Guidelines for Trip Generation Analysis. U. S. Bureau of Public Roads, 1967.

Appendix

TYPICAL DERIVATION OF CRITICAL RELATIVE ZONAL HOMOGENEITY INDEX

With respect to zonal sample distribution of total home-based trips per household, because

$$F = \frac{S_1^2}{S_2^2} \text{ and } F.05(df_1, df_2) = 1.783 \quad (1)$$

where

- S_1 = sample standard deviation for the total area sample (184 zones),
- S_2 = sample standard deviation for the zone being analyzed,
- df_1 = degrees of freedom for the total area sample = 4,159 - 1 = 4,158, and
- df_2 = degrees of freedom for the sample from the zone being analyzed = (4,159/184) - 1 = 22.

If

$$F = \frac{S_1^2}{S_2^2} \geq 1.783, \text{ then } S_1 \geq S_2 \quad (2)$$

where

- S_1 = population standard deviation for the total are population (184 zones) and
- S_2 = population standard deviation for the zone being analyzed.

But if

$$F < 1.783, \text{ then } S_1 < S_2 \quad (3)$$

Therefore, if

$$\frac{S_1}{S_2} < \sqrt{1.783}, \text{ then } S_1 < S_2 \quad (4)$$

$$\frac{S_1}{S_2} < 1.335$$

Therefore, if

$$S_2 \geq \frac{S_1}{1.335}, \text{ then } S_1 < S_2 \quad (5)$$

or if

$$S_2 \geq \frac{5.3238}{1.335} = 3.9912, \text{ then } S_1 < S_2$$

Therefore, for ratios of

$$\frac{S_2}{S_1} \geq \frac{3.9912}{5.3283} = 0.749 \quad (6)$$

S_1 is not significantly greater than S_2 .

Also calculated in this manner are the critical relative zonal homogeneity indexes with respect to automobile-driver home-based trips per household, total home-based work trips per household, household family size, household automobile ownership, household income level, and number of persons five years old or older per household.

Calibration of Transit Networks in Medium-Sized Urban Areas

DAVID T. HARTGEN, Planning Division, New York State Department of Transportation

•AN EFFECTIVE transportation plan for an urban area should promote complementary and not competitive uses of each travel mode. Therefore, the planning of transit facilities should include a careful analysis of the interaction of the transit network with other modes. The first step in this planning is to inventory and code the existing facilities and the characteristics of the travel on them. The elements of the transportation networks can then be represented within the computer as can also the flow of travel over them. Because the inventoried network represented in the computer must function as nearly like the real network as possible, it must be carefully checked and calibrated so that it can be used with confidence in planning future networks.

Network calibration is the process wherein the network simulated from inventory data is revised by a logical procedure so that travel routed over it by the computer is characteristically similar to that observed on the actual network. Although both are important, the process of calibrating the network can be distinguished from that of checking the network. The purpose of the latter is to ensure the accuracy of the data that describe the network's physical characteristics. The calibration process ensures that there is a comparable relationship between the network and travel as it actually exists and as it is simulated in the computer. Network checking is, therefore, a preliminary requisite to network calibration.

Network calibration is important in transportation planning for several reasons. First, it ensures that the existing operating characteristics of each mode are accurately described. This is important in planning a future system, a major portion of which will normally be composed of the existing one. Second, the estimation of future travel via each mode considers the relative level of service afforded by each mode, which is closely related to network speeds and operating characteristics.

Finally, operational problems involving network components, such as street segments, intersections, or transit routes, are more easily studied within the adjacent network structure than as isolated pieces. A calibrated network permits the study of the operation of different components of the existing system in relationship to the whole. In this way it supplements the inventoried data on components because the inventory does not supply data on component interaction. It also serves as a check on the reasonableness of component characteristics, particularly in response to travel patterns.

In the analysis of transit networks, a calibrated base network has other essential uses. The effect of changes in scheduling, routing, and fares on transit ridership is closely related to the differences in level of service on different portions of the existing network. It is extremely difficult, if not impossible, to manipulate real networks in a trial-and-error manner in order to observe the effects of improvements to that network. Simulated networks, properly calibrated, can be used to observe such effects prior to their implementation. Also, improvement of inefficient transit operation is often facilitated by study of the existing network that, if calibrated, will contain those inefficiencies at approximately the correct scale of importance.

This paper documents a method for preparing calibrated transit networks, using network inventory data and travel patterns observed in several upstate New York cities.

10/71 10/01

TRANSIT NETWORK CODING FORM
EXISTING - COMPOSITE

Coded by Paul Wood Date 7/15/68

System OR&A

Checked by RTH Date 7/15/68

ADDITIONS DELETIONS

Study SMTS

MAP NUMBER	DIST	LINK IDENTIFICATION								LINK LENGTH	TRAVEL SPEED	LINK TYPE	PLANNING LINK CODE	SCALE TIME	PEAK HOUR VOLUME	CAPACITY CLASS	DAILY BUS VOLUME	FREE TIME	ROUNDED LENGTH	SYSTEM	STREET AREA	LINK NUMBER
		"A" NODE				"B" NODE																
		ROUTE	NODE	ROUTE	NODE	ROUTE	NODE	ROUTE	NODE													
01	27	0006400	26006401	025	25	1							120	00				1230	2	OR&A		
01	27	26006401	26006400	025	25	1							37	00				1230	2	OR&A		
01	27	26006401	26005701	025	25	1							18	2	00	28		0600	0	OR&A		
01	27	2601450	26006401	025	25	1							16	2	00	28		0535	1	OR&A		
01	27	26006401	27006401	000	27	1							90	00				3000	0	OR&A		
01	27	27006401	27006450	025	27	1							10	2	00	16		0333	4	OR&A		
01	27	27006450	27014450	025	27	1							11	2	00	16		0361	5	OR&A		
01	37	27014450	27014450	025	27	1							11	2	00	16		0208	3	OR&A		
01	37	27014450	27015102	025	27	1							10	2	00	16		0292	4	OR&A		
01	37	27015102	27015351	025	27	1							14	2	00	16		0416	5	OR&A		
01	27	27006450	26006450	000	27	1							50	00				1660	0	OR&A		
01	27	0606450	26006450	025	27	1							6	3	00	82		0200	3	OR&A		
01	27	0606450	26006450	015	27	1							3	3	00	41		0100	2	OR&A		
01	27	0606450	26006350	025	27	1							6	3	00	41		0200	3	OR&A		
01	27	0606350	26014450	015	27	1							3	3	00	41		0100	2	OR&A		
01	37	06014450	26014450	025	27	1							10	3	00	41		0333	4	OR&A		
01	37	06014450	26006450	025	27	1							6	3	00	41		0260	3	OR&A		
01	27	0606450	26006450	015	27	1							3	3	00	41		0100	2	OR&A		
01	27	00074400	26014450	030	27	1							17	00				0560	3	OR&A		
01	37	06014450	26014400	030	27	1							44	00				1460	3	OR&A		
01	27	0006300	27014450	045	27	1							27	00				0900	5	OR&A		
01	37	27014450	26006300	045	27	1							67	00				2253	5	OR&A		
01	27	0006300	26006350	045	27	1							27	00				0900	5	OR&A		
01	27	0606350	26006300	045	27	1							67	00				2238	5	OR&A		

Figure 1. Transit network coding form.

TRANSIT NETWORK INVENTORY

The first step in the preparation of the calibrated transit network is an inventory of the existing routes. Information on the level of transit service is collected for each route segment. This normally includes scheduled peak and off-peak headway, number of daily buses in each direction, speed, fare, route locations, transfer points, and points of route turnbacks or branches.

The transit network is described by series of links defined by two nodes. The node number is composed of eight digits—the first three indicate the route number, the last five, the zone number and node number within that zone. For example the link

320 115 02-320 115 03

indicates route number 320, going from zone 115, node 02, to zone 115, node 03 (1). Additional information about the link is then coded as shown in Figure 1. Normally, the values of free time (travel time in hours) and rounded link length would be calculated from other coded data.

There are three major types of links in the transit network.

1. Mainline links (coded 1 or 2 in column 27 of the form shown in Figure 1) are the portions of the network over which buses actually travel; they are usually two-way. Depending on its operating characteristics, a mainline link may represent any one of a number of transit modes (local bus, express bus, or rail rapid transit.)

2. Access links (coded 3 or 5 in column 27) represent the paths of access from mainline links to zone centroids or loading nodes.

ularly during the peak hour. If it does, then it should be included. Studies in several cities in upstate New York indicate that a reasonable lower limit is a route having one bus during peak hour and four buses during the day in each direction.

The temporal or frequency-of-service aspect of a transit network can be measured by the average time a transit user spends waiting for a vehicle once he arrives at the bus stop or station. Also included is the waiting time at transfer points, which may be approximated by the number of transfers occurring over a given period.

The operation of the transit network is measured by travel speed once the transit rider has boarded the vehicle. The operating speed of the vehicle itself is unaffected by network coverage or by frequency of transit service. Door-to-door speed, on the other hand, is affected by these variables, and it is difficult to isolate their effect on overall network speed. Therefore, because it is a more independent measure, vehicle operating speed is used instead of door-to-door speed.

Travel pattern refers to the routing of trips over the network; it is based on trip length or trip duration. If the coded network responds to travel in a reasonable manner, travel distances and travel times should be similar to those on the actual network.

Measures of Network Operation

The process of calibration involves the comparison of estimates of transit network operation with known values, and subsequent revisions to improve agreement between the two. This comparison can be done by assigning known transit trip interchanges determined from the home-interview survey to the transit network, and observing the travel pattern. The parameters observed should measure a variety of aspects of system operation. Table 1 lists the most commonly used control values (in order of rela-

TABLE 1
SOURCES OF VALUES OF VARIABLES USED IN TRANSIT
NETWORK CALIBRATION

Variable	Source of Control Value	Source of Estimated Value	Aspect of Network Measured	Percent Error Allowable Between Control and Estimated Values
Operating speed ^a	Company records	Transit network summaries	Speed	5
Route-miles per square mile ^a	Company records	Transit network summaries	Coverage	10
Distribution of travel time ^a	Home interview	Assignment	Trip duration	Similar to control at 0.05 level ^c
Distribution of travel distance ^a	Home interview	Assignment	Trip distance	Similar to control at 0.05 level ^c
Person-hours of travel ^b	Home interview	Assignment	Frequency	5
Person-miles of travel ^b	Home interview	Assignment	Coverage	10
Average trip time ^b	Home interview	Assignment	Trip duration	5
Average trip length ^b	Home interview	Assignment	Trip distance	10
Door-to-door speed ^b	Home interview	Assignment	Speed, coverage, frequency	10
Transfers ^a	Company records	Assignment	Frequency	15
Screenline counts ^a	Survey	Assignment	Trip pattern	20

^aVariables that are independent of each other.

^bVariables that depend on the values of other variables.

^cUsing chi-square or Kolmogorov-Smirnov tests.

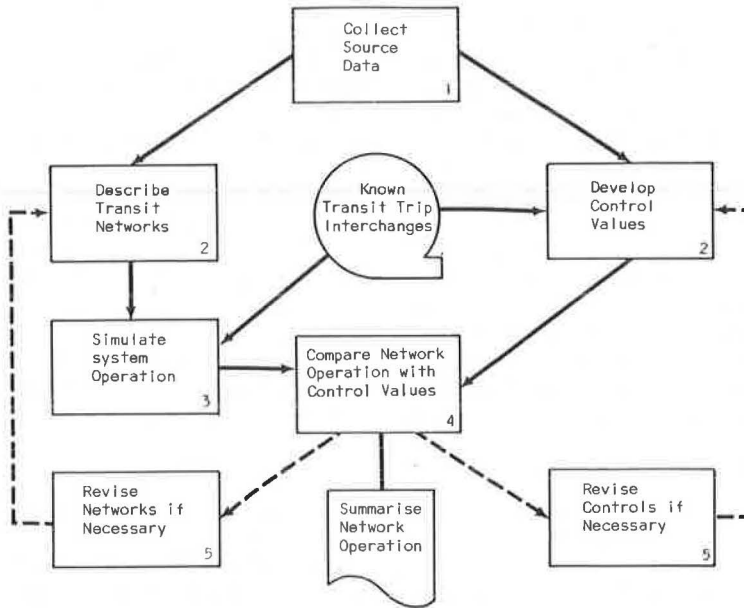


Figure 3. Transit network calibration method.

tive importance) and the allowable tolerance between these and the estimated values in the calibration process. The speed and coverage of the mainline portion of the network, as measured by operating speed and route-miles per square mile, can be calibrated before any attempts are made to simulate travel patterns.

Generally, estimates of transit travel time are more accurate than those of travel distance. This is because trip length is usually calculated from knowledge of the zone-to-zone airline distance using an over-the-road conversion factor, both of which may contain considerable error, especially for trips less than two miles in length. On the other hand, aggregate estimates of travel time are not affected by such variations. Therefore, a maximum error of 10 percent is allowed for variables involving trip length, but errors of less than 5 percent are required for variables involving time. These errors may seem unreasonably restrictive, but close agreement in the calibration of the existing network will increase confidence in travel-time trees and other network-derived data used in analysis of future travel patterns.

Because the number of transfers and screen-line volumes actually counted may vary considerably by day, less accuracy should be required when comparing their estimated and control values. Maximum errors of 15 and 20 percent respectively are recommended. Another useful check is passenger volumes at scattered locations on screen lines.

The similarity of the distributions of travel time and distance may be conveniently compared with a conventional chi-square or Kolmogorov-Smirnov test. A level of significance of 0.05 or 0.10 should be used.

Calibration Process

The method by which the estimated values are compared with control values is schematically shown in Figure 3. Its five basic steps are as follows:

1. Information on transit operation, obtained from company records, includes schedules and route layouts, coverage and bus speeds by route, transfer volumes, and screen-line counts.

2. The transit network is coded from data on scheduled speeds and routing supplied by the company. Network operation variables are derived from company records, and

travel data are obtained from the home-interview file. These provide control values against which the network operation can be measured.

3. The actual network operation is simulated by assigning known transit trip interchanges over the coded network. Estimated values of network variables are obtained from this assignment.

4. The operation of the simulated network is then compared with that of the actual network.

5. Revisions are made to the simulated network if its operation varies beyond the allowable limits with that of the actual network.

Revising the network, of course, implies that one is satisfied with the reliability of control values. In some cases, however, a revision of the controls, not the network, may be justified. This is particularly true in the early phases of calibration, when the collection of source data may not be complete. For instance, if operating speeds over a given portion of route are unavailable and the analyst elects to represent that portion as a function of surrounding portions, he may wish to revise the control speed based on observations of network operation through that segment.

When used in calibrating a network in a specific urban area, this process, simply described here, must, of course, be expanded and warped to fit the particular requirements of that area. Certain portions of source data may be unreliable or lacking altogether, or coding procedures may not permit detailed descriptions of the transit network. On the other hand, special surveys may provide information deemed reliable enough to develop estimates of control values. An on-bus survey, for instance, can yield considerably more data on transit usage than a conventional home-interview survey. These factors should be considered when preparing estimates of system operation.

An Example

Figure 4 shows the calibration process used in the transportation study for the Capital District, the area that includes the New York cities of Albany, Schenectady, and Troy. (The numbers in the lower right corner of the boxes in Figure 4 are used to key these steps to the discussion that follows in which similar numbers appear in parentheses.) The analysis of transit in this region is complicated by the fact that each of the three urban centers is served by a different transit company. Although there is a limited amount of intercity service, the vast portion of the transit service is within city limits. All three companies serve about 70,000 daily riders.

The calibration procedure used in this region is influenced to some extent by these considerations. Because most travel is intracity, the transit network in each urban center is only slightly influenced by the other two networks and may be treated separately. For this arrangement, it seemed reasonable to calibrate each network by city—first, the mainline portion, then access portions, and finally transfer portions.

Data obtained from the three transit companies included schedules, frequency of service, and operating speeds for each route. From these, all routes having at least one bus during the peak hours and four daily buses (each way) were coded as part of the transit network. This accounted for about 75 percent of the total reported route-miles of operation. Visual checks with the real system (1) indicated that additions to the transit network were required, which necessitated reducing the lower limit of route service to one peak-hour bus and two daily buses.

Addition of these routes (2) resulted in the network shown in Figure 5. This network contains about 90 percent of the total route-miles of operation reported by the companies. The control estimate of total route-miles, in this case, was unchanged although additions were made to the network. Alternatively, one might have eliminated that portion of the total route-miles that had very infrequent service and, thus, altered the control value. This course of action would be justified when dealing with very extensive networks in which only a very small portion of routes provide poor service.

Next, data on average bus-miles and bus-hours were developed from the coded transit network (3) and compared with published company statistics. From these figures estimates of average system-wide bus operating speeds were prepared. The results

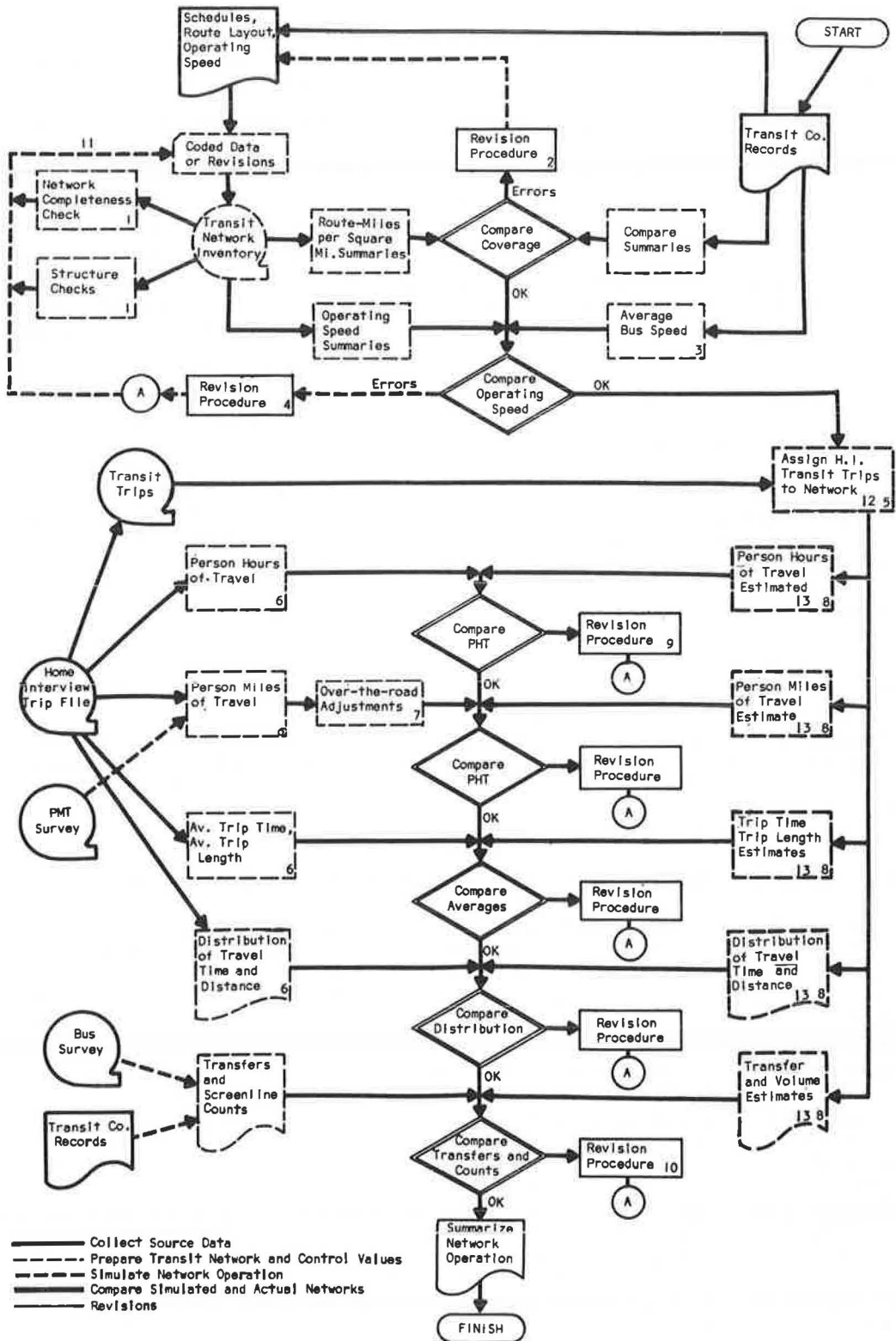


Figure 4. Calibration procedure for the Capital District transit network.

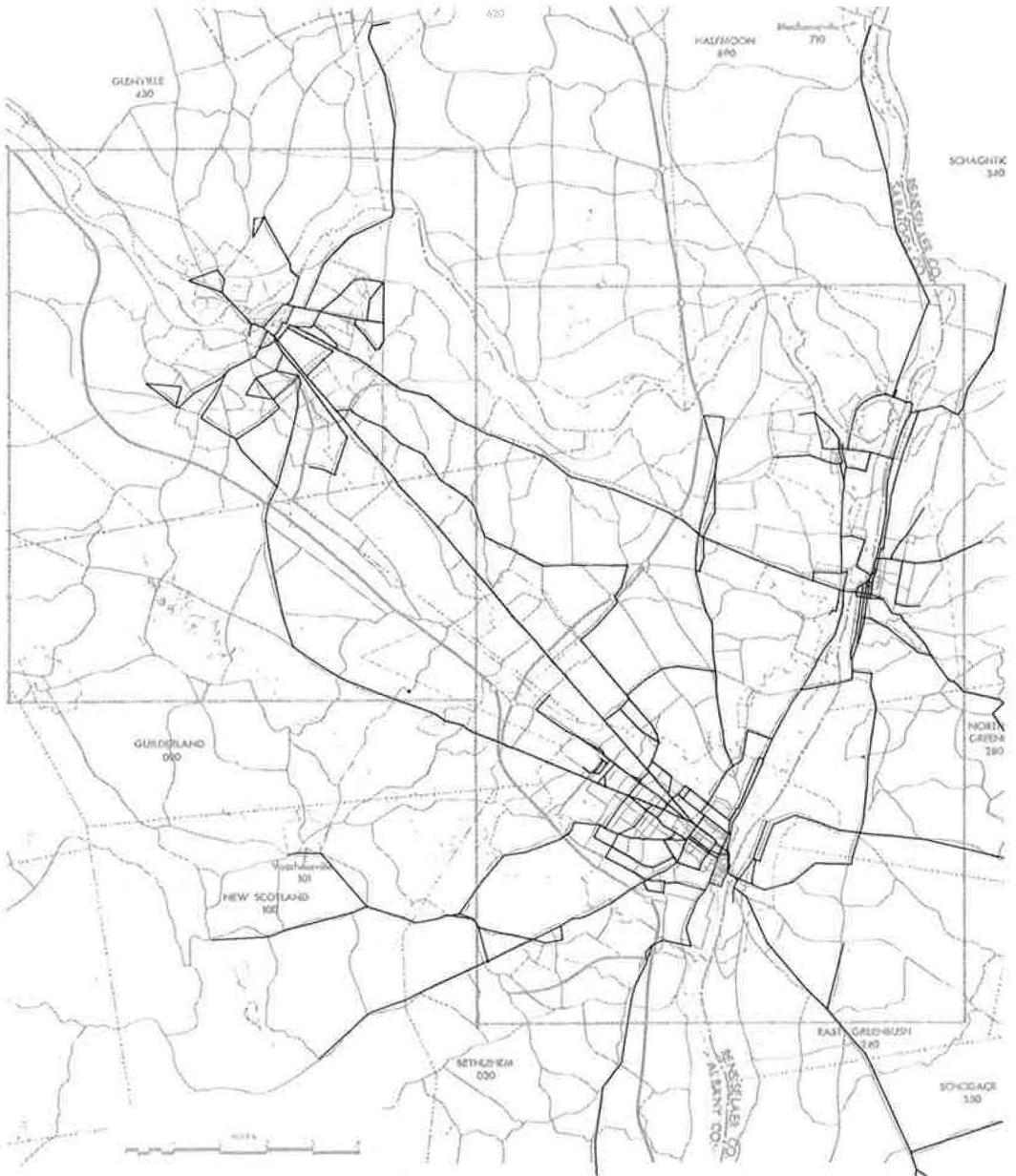


Figure 5. Base transit network—Capital District transportation study.

are given in Table 2. This analysis showed that the characteristics of the mainline portion of the network agreed well with system characteristics.

In many cases, however, the actual operating speed of the network may be faster than that of the real transit networks. This is because the network, although correctly described from published company schedules, does not include delays in operation caused by traffic congestion, breakdowns, or frequent stops. If the mainline speed of the coded network is inordinately high, it should be examined for errors and omissions of this type. Appropriate revisions (4) can then be made.

TABLE 2
COMPARISON OF ACTUAL AND ESTIMATED MAINLINE OPERATING CHARACTERISTICS
OF THE TRANSIT NETWORK

City	Bus-Miles			Bus-Hours			Operating Speed		
	Actual	Estimated	Percent Error	Actual	Estimated	Percent Error	Actual	Estimated	Percent Error
Albany	12,910	12,120	- 6.1	1,300	1,250	- 3.8	9.9	9.7	-2.0
Schenectady	3,650	3,350	- 8.2	318	300	- 5.7	11.4	11.2	-1.8
Troy	<u>1,270</u>	<u>1,120</u>	<u>-11.8</u>	<u>164</u>	<u>140</u>	<u>-14.7</u>	<u>7.8</u>	<u>8.0</u>	<u>+2.6</u>
Total	17,830	16,590	- 6.9	1,782	1,690	- 5.2	10.0	9.8	-2.0

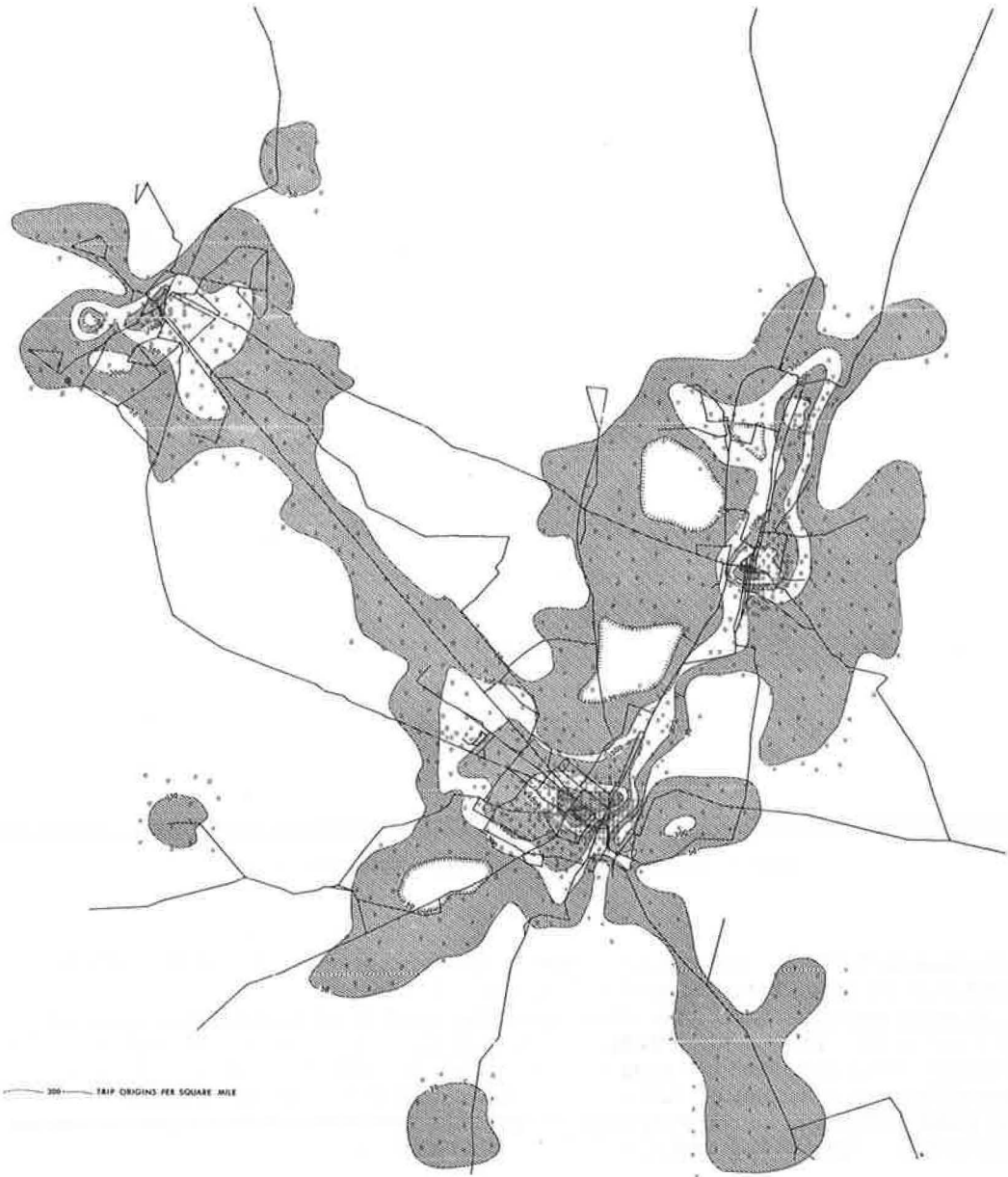


Figure 6. Distribution of transit trip origins in relationship to base transit network.

In the Capital District, the pattern of transit trips is closely aligned to the location of the transit network. Figure 6 shows the base transit network and the spatial distribution of transit trip origins (shown by the pattern of dots). Density contours indicate that the limit of existing service is in the range of 50 to 100 transit trip origins per square mile. This is about 5 percent of the total trip-origin potential. Using this technique, the analyst can check the reasonableness of the spatial location of each transit route. These criteria may also be used as guides in planning the extent of the future network.

The calibration may now proceed to the next stage, that of simulating travel over the network. This is usually done by assigning known transit trip interchanges to the network (5) and observing the resulting travel pattern. Control values used in this portion of the calibration may be derived from trip data obtained in the home-interview survey or from other supplemental studies.

Two important pieces of data are the distributions of travel time and distance, from which the average trip duration, average trip length, total person-hours of travel, and total person-miles of travel are obtained (6). The reported trip lengths are adjusted (7) to reflect over-the-road travel (3). Control values derived from the Capital District home-interview file are given in Table 3.

The outputs of the first transit assignment (8) are estimates of each of the control values. Table 3 shows that these estimates are, for the most part, not acceptable. The estimate of person-miles of travel (PMT) is within 10 percent of the control value, but person-hours of travel (PHT) is considerably outside. Because travel speed over the mainline portion of the network has already been calibrated (Table 2) in the pre-assignment phase, a low PHT estimate at this point would indicate an inordinately high speed on access links. This is confirmed by comparison of the estimated door-to-door speed (8.3 mph) with the control value (5.8 mph).

TABLE 3
COMPARISON OF CONTROL AND ESTIMATED VALUES OF VARIABLES BEFORE
AND AFTER TWO ASSIGNMENTS

Variable	Values				Percent Error			
	Control	Assignment Estimates			Control	Assignment Estimates		
		Pre	1st	2nd		Pre	1st	2nd
Operating speed, mph ^a	10.0	9.8	—	—	5.0	-2.0	—	—
Albany	9.9	9.7	—	—	5.0	-2.0	—	—
Schenectady	11.4	11.2	—	—	5.0	-1.8	—	—
Troy	7.8	8.0	—	—	5.0	+2.6	—	—
Dist. of travel time	— ^a	—	— ^a	— ^a	— ^c	—	— ^b	— ^c
Person-hours of travel	37,231 ^d	—	28,533	37,061	5.0	—	23.3	-0.5
Person-miles of travel	216,750 ^d	—	234,843	228,373	10.0	—	+ 8.3	+5.4
Average trip time, min	32.1	—	24.6	32.0	5.0	—	-23.3	-0.5
Average trip length, min	3.1	—	3.4	3.3	10.0	—	+ 8.3	+5.4
Door-to-door speed, mph	5.8 ^d	—	8.3	6.1	10.0	—	+43.1	+5.2
Transfers	8,600 ^e	—	5,700	9,100	15.0	—	-33.7	+5.8
Total transit trips	69,500	—	—	—	—	—	—	—

^aSee Figure 5.

^bDifferent at 0.05 level using Kolmogorov-Smirnov test.

^cSimilar at 0.05 level using Kolmogorov-Smirnov test.

^dFrom home-interview survey.

^eFrom Table 2.

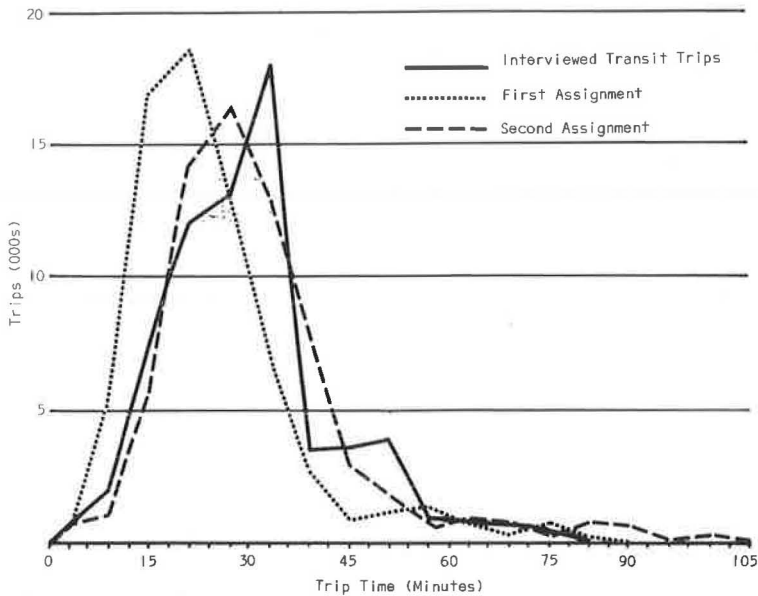


Figure 7. Distribution of transit travel time.

Access speed may be high for a number of reasons. The analyst may have over-estimated the number of persons boarding transit vehicles from automobiles (park-and-ride, kiss-and-ride) and, therefore, provided too much access at automobile speeds. Also, for those areas where automobile access is legitimate, the automobile speeds themselves may be too high. This might result from use of average or free automobile speeds when, in fact, a major portion of transit travel occurs during peak hours when automobile speeds are restrained because of congestion. Again, walk speeds on access links may be overestimated. Finally, wait time, which is included in the calculation of access speed, may be underestimated. This might occur when congestion or other delays increase transit vehicle headways to values greater than those reported in schedules.

At any rate, a major revision to the network at this point (9) should be a reduction of access speeds. The reduction factor required should be somewhat less than the ratio of speeds (0.70) because a major influence of network speed is due to mainline movement, not to access movement. In the Capital District, time spent on access links represents about 61 percent of total travel time; the adjustment factor, therefore, should be 0.61×0.70 or 0.43. Note that we have determined to make all of this adjustment in travel time, not in trip length, even though the estimate of average trip length is also in error by 8 percent. This course of action is justified, as Figure 7 shows, because the distribution of travel time estimated from the first assignment is similar in shape to the distribution of interviewed or control data, but is shifted to the left. An across-the-board factor applied to all travel times will adjust the distribution mean without materially altering its shape.

One other adjustment that should be made at this point is in transfer volumes. The estimated value is 34 percent low, indicating that travel over these links is extremely difficult. The problem may be simply an overestimate of transfer time. Transfers normally require a headway time of about half that of the route being boarded, but some scheduling of arrivals at transfer points can reduce this considerably, particularly for routes with headways greater than 30 minutes. In this case, analysis showed that such an error had occurred throughout the system. Therefore, reducing the time required to make transfers by a factor of $5700/8600$, or 0.66, is justified and should result in a closer agreement (10).

These revisions are then applied to the access and transfer portions of the network respectively (11). Known transit trips are then reassigned (12) to the adjusted network and new estimates of PHT, PMT, average trip time and length, and transfers are obtained (13). As data given in Table 3 show, this second assignment indicates that the network is now simulating actual network travel fairly well. Virtually all error in estimates of total PHT and average trip time has been eliminated, and the error in PMT has been reduced to 5 percent. The distribution of travel time, as obtained from the assignment, is similar to the control distribution (Fig. 7) at the 0.05 level, using the

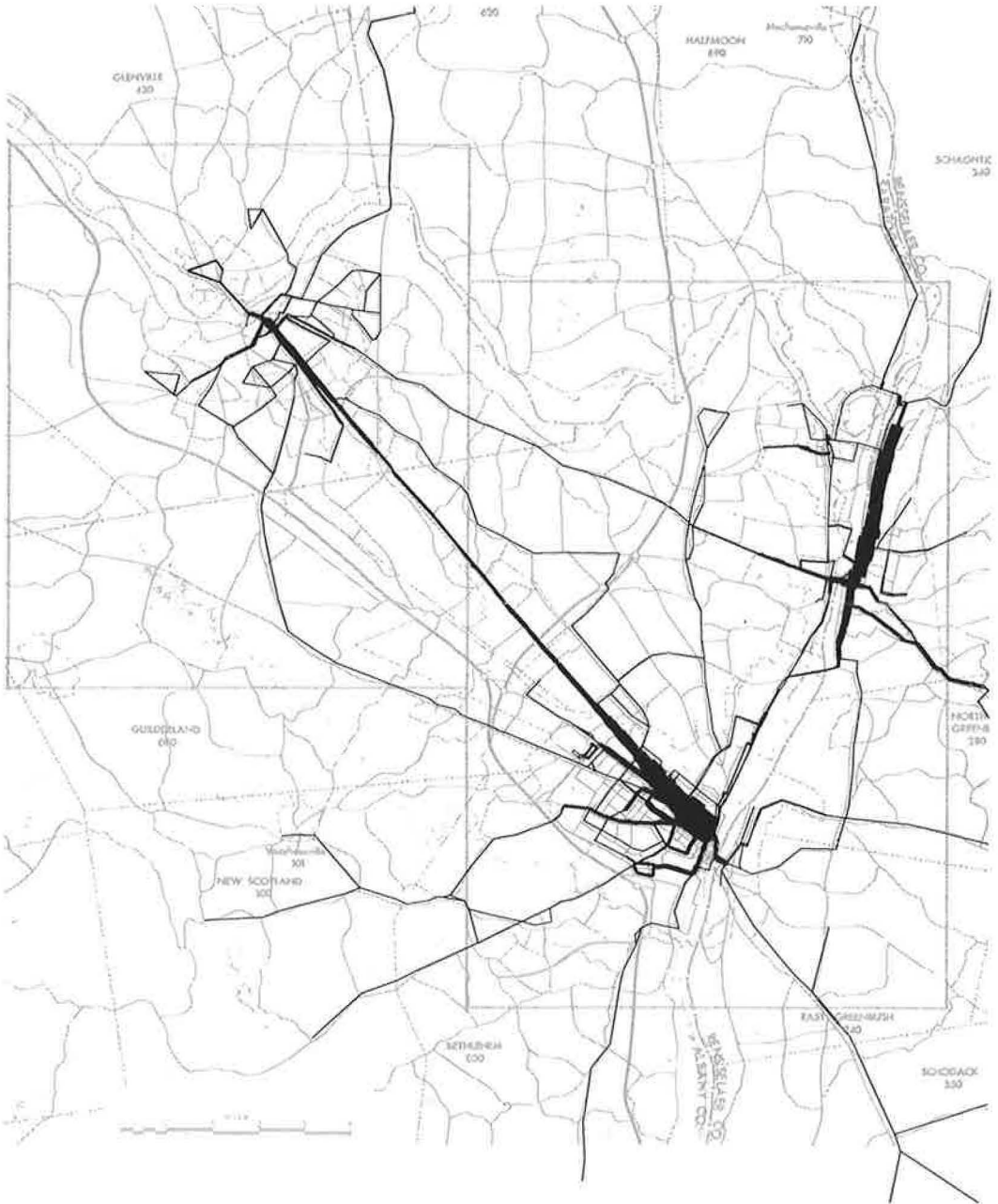


Figure 8. Travel over base transit system.

Kolmogorov-Smirnov test. Factors applied to transfer time result in a slight overestimate of transfer volumes, but the estimate is well within allowable limits. Unless further refinement in PMT is desired, the network should now be considered calibrated.

Travel over the base network is shown in Figure 8. The flow increases uniformly toward each of the city centers, with the Albany center attracting about two-thirds of all travel. Intercity movement is confined to the Albany-Schenectady and Albany-Troy corridors. However, all portions of the network, except the outermost routes, carry some travel.

The calibrated network may now be used to analyze characteristics of the actual network that it represents. Summaries of route-miles, bus-miles, and average loadings per mile, by zone, are useful in locating areas of further concentration. Sets of travel-time trees for each zone show how the network is loaded by assigned trips. Measures of interzonal separation can also be developed for use in the analysis of modal-split and accessibility models.

SUMMARY

The calibration of transit networks plays an important part in the transportation planning process. Successful network calibration depends on the reliability of available source data and the ability to simulate the real network and travel over it. As many different aspects of system operation as possible should be analyzed to ensure similarity between the actual and the simulated networks. Variables chosen to represent the characteristics of the network in the calibration process should measure route coverage, frequency of service, and travel speed.

The initial phase of calibration involves the mainline portion of the network. Checks on bus-miles of operation and average operating speed ensure that this portion of the network is correctly described. This phase should be completed before any attempt is made to simulate travel.

Patterns of travel over the transit network can then be analyzed for similarity with actual travel characteristics. This phase of calibration is concerned with simulating operating characteristics by revising the network so that the distributions of travel time and distance are similar to independently derived distributions. Other useful measures are average trip time and length, person-miles of travel, person-hours of travel, transfers, and screen-line counts.

In general, a transit network of average complexity can be calibrated in two to three assignments, given, of course, previous agreement between actual vehicle speeds and those posted or reported by the operating agency.

REFERENCES

1. Fifield, David C. Transit Network Inventory Manual. Planning Div., New York Dept. of Transportation, Albany, unpublished.
2. Petersen, Stephen G. Walking Distances to Bus Stops in Washington, D.C. Residential Areas. Traffic Engineering, Dec. 1968, pp. 28-34.
3. Woods, K. B., ed. Highway Engineering Handbook. McGraw-Hill, New York, 1960, Figure 4-24, pp. 4-46.

Automobile Occupancy Projections Using A Modal-Split Model

FRANKLIN SPIELBERG, Cleveland-Seven County Transportation-Land Use Study

The use of a diversion-curve, modal-split model is proposed as a method of converting person trips on a highway network to vehicle trips. The model is programmed for the CDC 3600 computer. Investigation was done to determine significant variables for predicting vehicle usage. Median family income at the production end of the trip, orientation of the trip to the central business district or to locations other than the central business district, travel time, and trip purpose were used as variables. Curves relating the percentage of highway person trips made by drivers to these variables were developed and the model was tested with base-year data. The total error in predicted vehicle trips was less than 0.5 percent indicating that this type of model is well suited to prediction of vehicle usage. Projections to a future year indicated that significant decreases in automobile occupancy will occur.

●A MAJOR DESIRED output of the transportation planning process is a highway network that will meet the vehicular travel demand in a given design year. The usual approach to this problem is as follows:

1. Base-year data are collected.
2. A series of generation, distribution, modal-split, and assignment models are calibrated.
3. Projections of independent variables are made to the design year.
4. Total person trips are generated and distributed.
5. A modal split is performed separating transit trips from total person trips.
6. An automobile occupancy factor is applied to the remaining trips to convert them from person trips to vehicle trips.
7. The vehicle trips are assigned to a test network.

At this point, based on established criteria, the network is evaluated. It is either recommended as it is for further consideration by the planning body, or it is modified. If modifications are indicated, the testing process begins again at either the assignment or the distribution phase.

Although not generally recognized, the automobile occupancy factor used in converting highway person trips to highway vehicle trips can be an important factor in determining the adequacy of a proposed system. Given 5 million highway trips, an occupancy factor of 1.5 would yield 3.33 million vehicle trips, and a factor of 1.3 would yield 3.85 million trips—an increase of 16 percent in the number of vehicle trips that must be served. Thus, the projection of automobile occupancy rates can be critical in system design.

At one time, the usual method of arriving at design-year occupancy was to compute average automobile occupancy from base-year data.

$$\text{Average persons per car} = (\text{automobile-driver trips} \\ + \text{automobile-passenger trips}) / \text{automobile-driver trips}$$

This factor was applied uniformly to the projected highway person trips.

As the planning process became more sophisticated, it was realized that automobile occupancy would vary with the purpose of the trip. A factor, therefore, was computed for each trip purpose, using base-year data.

$$\text{Average persons per car}_{(N)} = [\text{automobile-driver trips}_{(N)} + \text{automobile-passenger trips}_{(N)}] / \text{automobile-driver trips}_{(N)}$$

where N is purpose of the trip.

This factor was used to convert the projected person trips to vehicle trips on a trip-purpose basis. It, however, introduced additional error because the trip purpose of the passenger is frequently not the same as the purpose of the driver. Where it was felt that the average automobile occupancy would change over time as the characteristics of the region changed, a subjective judgment was made of the magnitude and direction of the shift, and similar factors were applied.

This line of reasoning, that the automobile occupancy would vary with the characteristics of the region, led to the use of different occupancy factors for CBD-oriented trips and for non-CBD-oriented trips, as opposed to a uniform factor for all travel in the region. A logical extension of this thinking led the Twin Cities Transportation Study to develop a model that predicted automobile occupancy rates for work trips based on the income at the production end of the trip. A second equation predicted occupancy for all other trips based on only production zone income.

These equations form what is, in fact, a modal-split model. It divides trips into automobile-driver and nonautomobile-driver trips on the basis of the individual interchanges and the characteristics of the trip ends. It reduces errors resulting from the application of uniform factors to large areas. For the Cleveland-Seven County Transportation-Land Use Study (SCOTS), it was decided to investigate this modal-split model approach to automobile occupancy projection.

MODEL

Model Formulation

The modal-split model that was chosen for use by SCOTS was programmed for the CDC 3600 computer. It is a diversion-curve model in which a series of curves are developed that split a given trip table into two trip tables. The percentage split is read from a curve. The ordinate of the curve is the percentage to be allocated to one table and the remaining trips are allocated to the second table. The abscissa is designed to be the ratio of travel time between the mode associated with the first table and the mode associated with the second table.

Associated with each trip are (a) a production code that relates to the value of a parameter at the production end of the trip, (b) an attraction code that relates to a parameter at the attraction end of the trip, (c) a range code that relates to an interchange parameter, travel time, and (d) purpose. A separate diversion curve may be used for each combination of trip purpose (up to 11 purposes), production code (up to 4 codes), attraction code (up to 4 codes), and range code (up to 4 codes). Thus a total of $11 \times 4 \times 4 \times 4$ or 704 curves may be used, if sufficient base data exist to develop and apply the number of curves. In actual practice, the number of curves used is much smaller.

Curve Development and Calibration

Given this formulation of the model, the next steps are the determination of significant production, attraction, and interchange parameters and the values on the diversion curves. A series of regressions were performed to examine the contribution of each set of production zone parameters to the total variance in automobile occupancy. Because the model attributes total variation to four variables, excluding trip purpose, large values of correlation coefficients were not anticipated; a correlation coefficient

on the order of 0.25 was considered sufficient to indicate significant production parameters. For simplicity and speed in model operation, it was decided to use the same production parameter for all purposes, although this is not necessary.

Of the production parameters examined, three appeared significant—median family income, automobiles per person, and dwelling units per acre. Of the three, median family income appeared to be the most significant for the majority of purposes, and it was selected to be the production zone parameter. Because the model allows only four levels of the parameter, a stratification was made into high income (above \$9,550), medium (\$6,550 to \$9,550), and low (under \$6,550).

An interesting sidelight from the analysis is that the availability of transit service does not have the expected impact on automobile occupancy. Areas with transit service showed higher occupancy rates than those without service. This indicates that the income and automobile ownership characteristics are more significant than transit service in determining automobile occupancy and that trips made by automobile passengers would have been automobile-driver trips had the additional vehicles been available.

For the attraction-area parameter, it was decided to use a simple CBD, non-CBD split. This approximates a measure of the ease of parking. Subsequent investigation has indicated that, for certain trip purposes, accuracy could be increased by an additional stratification to include non-CBD areas in which parking is difficult. This modification may be introduced into later applications of the model.

In determining an interchange parameter, we felt that travel time would be significant. The question arose, however, if it would be a sufficient measure. Base-year trips, obtained in the origin-destination study, were stratified by purpose (home-based including work, shop, social-recreation, school, and miscellaneous, which includes personal business, medical-dental, and eat meal; and nonhome-based), mode (automobile driver, automobile passenger), production code, attraction code, and travel time. For travel time, the skim-tree time from the base-year highway network was used.

PAGF 1			
AUTO DRIVER AND PASSENGER TRIPS BY TIME AND INCOME P CODE AND A CODE			
	DRIV	PASS	
NCHD			
HIGH			
MISC			
3	43344	6772	50116
4	192237	44309	236546
5	186028	36665	222693
6	446544	98606	545150
7	527422	146788	674210
8	385671	148550	534221
9	478504	120283	598787
10	462923	147336	610259
11	314651	99719	414370
12	271134	72938	344072
13	235484	90568	326052
14	196635	58812	255447
15	179661	70874	250535
16	143379	44643	188022
17	114903	46107	161010
18	117393	53631	171024
19	115956	54328	170284
20	95412	36211	131623
21	76849	33423	110272
22	73773	32719	106492
23	57827	21562	79389
24	66537	32128	98665
25	54473	20119	74592

Figure 1. Example of analysis table.

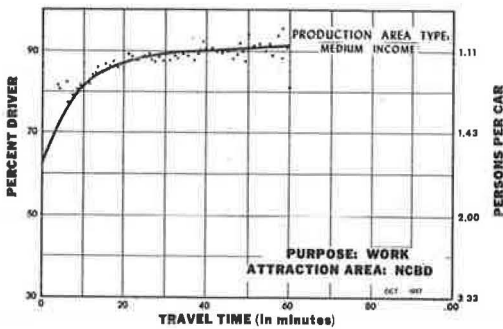


Figure 2. Example of base-year data points with hand-fitted curve.

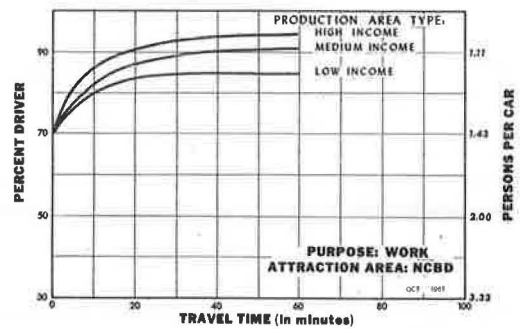


Figure 3. Model diversion curves for automobile driver-passenger split, work purpose to non-CBD area.

Thus, data were obtained on the total number of trips for a given purpose, of a given interchange type and travel time, for each of the two modes. Figure 1 shows an example of these tables.

From these tables, the percentage of total highway trips that were automobile-driver trips was computed for each travel-time increment. The sets of points were plotted with percentage of drivers as the ordinate and travel time as the abscissa. The driver percentage, the reciprocal of automobile occupancy, was used to maintain compatibility with the form of the modal-split model. After the points were plotted, a curve was hand-fitted to them. Figure 2 is an example of one of these charts.

Not all the groupings yielded easily fitted sets of points because, in many cases, an insufficient number of trips fell into the given strata. A guideline was established, therefore, for fitting the curves to the points. This was that, for a given purpose-attraction code combination, the diversion curves for all income groups would have the same shape. The applicability of the rule was apparent for those purposes that had sufficient data in all groupings to show clearly defined curves.

The fit obtained on the majority of these curves indicated that travel time alone would be acceptable as the single interchange parameter. The diversion curves obtained from this analysis are shown in Figures 3 through 11. Curves for school trips are not included. Analysis indicated that acceptable relationships for school trips could not be developed from strictly automobile driver-passenger data because a large number of school trips are made by other modes and because, for most automobile-passenger school trips, the driver's purpose is generally other than that of going to school. This was not felt to be a problem because of the small number of automobile-driver trips with a true school purpose.

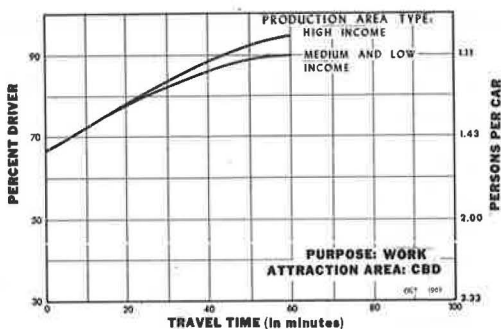


Figure 4. Model diversion curves for automobile driver-passenger split, work purpose to CBD area.

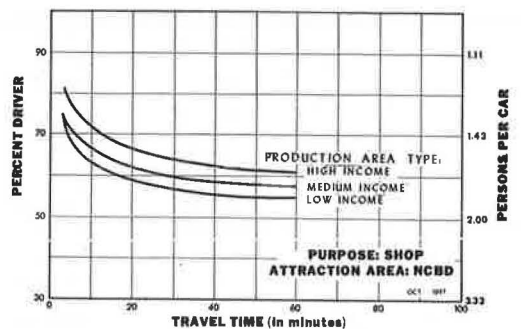


Figure 5. Model diversion curves for automobile driver-passenger split, shop purpose to non-CBD area.

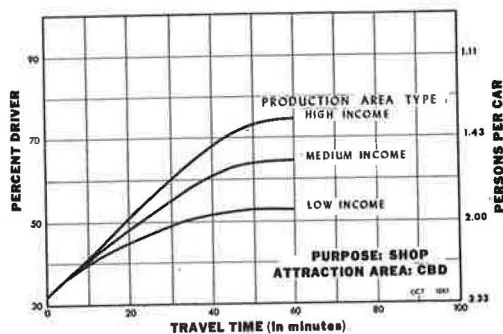


Figure 6. Model diversion curves for automobile driver-passenger split, shop purpose to CBD area.

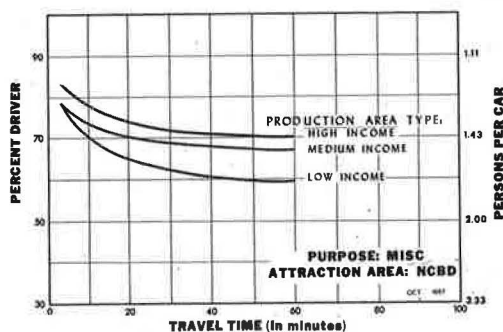


Figure 7. Model diversion curves for automobile driver-passenger split, miscellaneous purpose to non-CBD area.

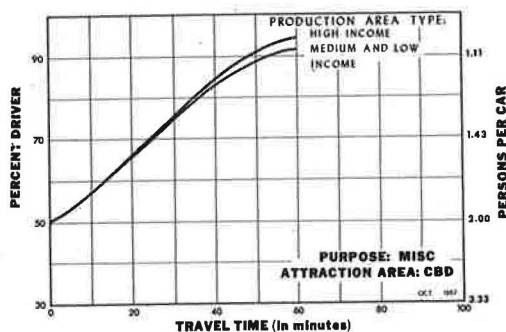


Figure 8. Model diversion curves for automobile driver-passenger split, miscellaneous purpose to CBD area.

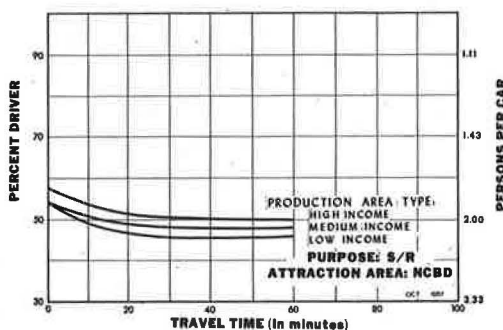


Figure 9. Model diversion curves for automobile driver-passenger split, social-recreation purpose to non-CBD area.

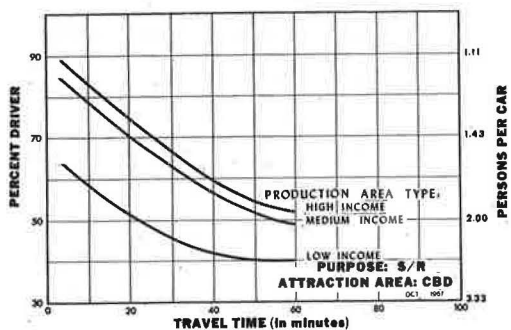


Figure 10. Model diversion curves for automobile driver-passenger split, social-recreation purpose to CBD area.

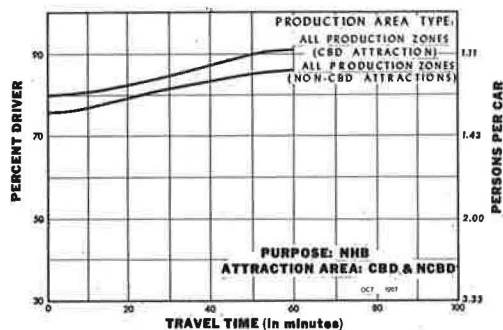


Figure 11. Model diversion curves for automobile driver-passenger split, nonhome-based purpose to non-CBD and CBD area.

Interpretation of Diversion Curves

The curves may be divided into two groups: rising curves for which the driver percentage rises with travel time and falling curves for which the driver percentage falls with travel time. Falling curves seem to be associated with casual trips, i. e., all social-recreational trips as well as shopping and miscellaneous trips with destinations outside the CBD. Rising curves are associated with definite trip purposes, i. e., work and shopping within the CBD. This indicates that longer casual trips are not made unless several persons are involved. Casual passengers, who "go along for the ride," tend to avoid longer trips when the primary trip-maker (the driver) has a definite trip purpose.

TESTS OF TIME MODEL

Total Trips

The prime test of the model was its ability to reproduce base-year trips. For this test, trip tables were prepared of automobile-driver plus automobile-passenger trips for each of the six purposes and for total trips. These tables represented all internal person trips by highway among the 986 internal zones as found in the base-year, origin-destination survey. A modal split was made, and automobile-driver trips predicted by the model were compared to those found in the survey. To facilitate comparison, the tables were compressed to 117 districts, yielding 13,689 possible interchanges, and stratified by volume groups to indicate any improvement in projection of larger volumes. In addition to the usual root-mean-square (rms) statistic, a t-statistic was incorporated into the comparison program to indicate if there were statistically significant differences between the projected interchanges and the actual interchanges. The critical t-value varies with the actual number of interchanges in the volume group being considered. For most volume groups, however, when the absolute value of t is less than 1.97, we have no reason to reject at the 5 percent level the hypothesis that the predicted value is the same as the observed. This comparison is given in Table 1. Figure 12, which shows the percentage of rms error plotted against the volume group, indicates that 68 percent of the projections were within ± 10 percent for volumes greater than 600. Estimates of this accuracy are well within the tolerance found in transportation planning models.

TABLE 1
DISTRICT-TO-DISTRICT VOLUMES OBTAINED FROM O-D SURVEY AND ESTIMATED BY MODEL
FOR AUTOMOBILE-DRIVER TRIP INTERCHANGES

Volume Groups	O-D Volume		Model Volume		Average Interchange Difference ^a	Standard Deviation	t-Value	Percent rms Error
	Total	Average Interchange	Total	Average Interchange				
	(1)	(2)	(3)	(4)				
0 to 1	0	0.0	2,688	1.0	-0.99	3.17	-16.27	0.00
3 to 3	162	3.0	200	3.7	-0.70	2.17	-2.39	75.90
4 to 4	772	4.0	963	5.0	-0.99	4.00	-3.44	102.98
5 to 5	525	5.0	679	6.5	-1.47	4.10	-3.67	87.02
6 to 6	1,758	6.0	2,037	7.0	-0.95	4.31	-3.78	73.53
7 to 7	1,372	7.0	1,452	7.4	-0.41	3.75	-1.53	53.84
8 to 8	864	8.0	1,115	10.3	-2.32	11.32	-2.13	144.45
9 to 9	2,358	9.0	2,637	10.1	-1.06	5.10	-3.38	57.94
10 to 10	3,460	10.0	3,595	10.4	-0.39	5.06	-1.43	50.75
11 to 15	8,959	12.9	9,393	13.5	-0.62	6.32	-2.61	49.26
16 to 20	12,305	18.2	12,772	18.9	-0.69	7.61	-2.36	41.96
21 to 25	11,735	22.8	11,853	23.1	-0.23	8.69	-0.60	38.06
26 to 30	13,196	27.9	13,486	28.5	-0.61	10.49	-1.27	37.66
31 to 35	13,188	32.9	13,485	33.6	-0.74	11.82	-1.25	36.02
36 to 40	14,766	37.9	14,835	38.0	-0.18	10.40	-0.34	27.46
41 to 45	12,736	43.0	12,565	42.4	0.58	11.29	0.88	26.28
46 to 50	13,180	47.9	12,654	46.0	1.91	10.78	2.94	22.85
51 to 60	26,720	55.4	26,205	54.4	1.07	12.51	1.88	22.64
61 to 70	25,870	65.2	25,456	64.1	1.04	13.85	1.50	21.31
71 to 80	25,443	75.3	25,506	75.5	-0.19	15.33	-0.22	20.37
81 to 90	23,208	85.6	23,221	85.7	-0.05	18.80	-0.04	21.95
91 to 100	25,470	95.4	25,531	95.6	-0.23	19.45	-0.19	20.39
101 to 150	107,755	122.6	106,846	121.6	1.03	19.02	1.61	15.54
151 to 200	97,093	174.0	97,302	174.4	-0.37	26.11	-0.34	15.01
201 to 250	76,978	224.4	76,233	222.3	2.17	28.05	1.43	12.53
251 to 300	85,675	274.6	85,194	273.1	1.54	31.68	0.86	11.55
301 to 350	78,249	323.1	75,876	321.5	1.58	35.36	0.89	10.90
351 to 400	83,781	375.7	82,971	372.1	3.63	34.01	1.59	9.10
401 to 450	68,589	423.4	67,898	419.1	4.27	40.10	1.35	9.52
451 to 500	65,845	473.7	65,405	470.5	3.17	43.03	0.87	9.11
501 to 1,000	385,914	689.1	379,582	677.8	11.51	57.20	4.68	8.46
1,001 to 2,000	367,980	1,378.2	366,463	1,372.5	5.68	95.28	0.97	6.93
2,001 to 3,000	238,222	2,430.8	238,939	2,438.2	-7.32	144.22	-0.50	5.94
3,001 and over	1,040,442	6,267.7	1,038,548	6,256.3	11.41	346.32	0.42	5.53
Total or average	2,932,570	214.2	2,923,585	213.6	0.66	46.42	1.65	21.67

^aColumn 2 minus column 4.

TABLE 2
 AUTOMOBILE-DRIVER TRIPS BY PURPOSE OBTAINED
 FROM O-D SURVEY AND ESTIMATED BY MODEL

Purpose	O-D Survey	Model Estimations	Model as Percent of O-D
Nonhome-based	609,702	626,076	102.6
Home-based			
Work	805,525	806,875	100.1
Shop	626,118	622,694	99.4
Social-recreation	401,943	402,791	100.2
Miscellaneous	449,710	446,966	99.3
Total, includes school	2,933,034	2,923,985	99.6

Trips by Purpose

In addition to the comparison of the total trip projection, an analysis was made of the estimates for each of the trip purposes to find any bias in the individual curves.

Table 2 gives the total number of automobile-driver trips by purpose as found in the origin-destination survey and as estimated by the model. Figures 13 through 17 show the relationship of the percentage of rms error to volume for each of the purposes. It is felt that the estimates of trips by purpose is

adequate, but further analysis is being done on nonhome-based trips to eliminate the overestimate in that category.

As mentioned previously, it was not possible to develop reasonable curves for school trips. For this reason, they were split using a uniform factor in the model test and were included in the total trip comparison. They are not included in the purpose-by-purpose comparison.

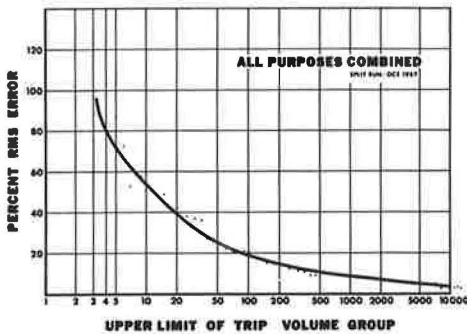


Figure 12. Percentage of rms error for estimated automobile-driver trips, all purposes.

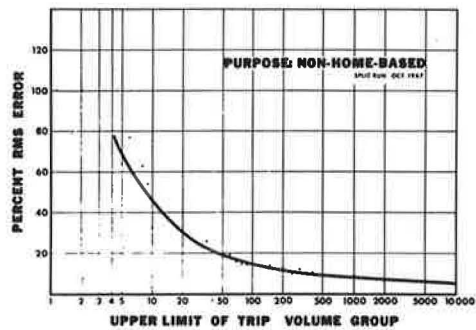


Figure 13. Percentage of rms error for estimated automobile-driver trips, nonhome-based purpose.

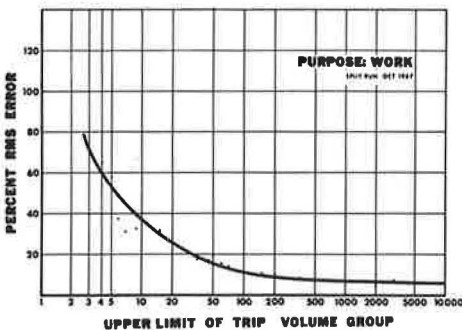


Figure 14. Percentage of rms error for estimated automobile-driver trips, work purpose.

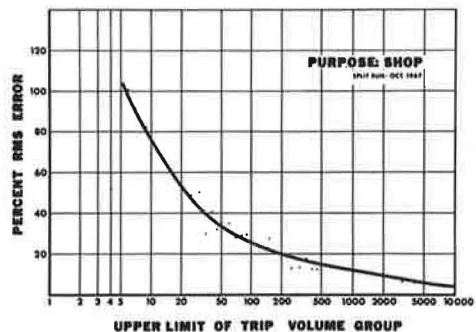


Figure 15. Percentage of rms error for estimated automobile-driver trips, shop purpose.

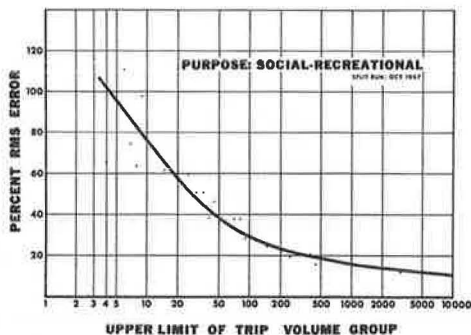


Figure 16. Percentage of rms error for estimated automobile-driver trips, social-recreation purpose.

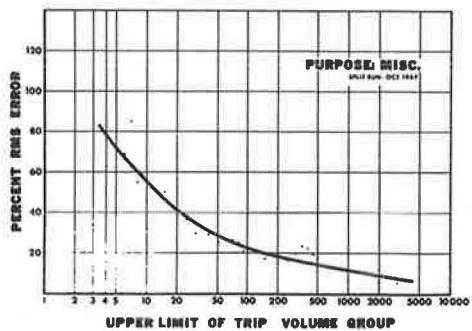


Figure 17. Percentage of rms error for estimated automobile-driver trips, miscellaneous purpose.

Small Areas and CBD

Further investigation was performed to determine if the model was correctly estimating the number of automobile-driver trips to various parts of the region. Table 3 gives the number of productions and attractions of automobile-driver trips by purpose as found in the origin-destination survey and as predicted by the model for a selected

TABLE 3
AUTOMOBILE-DRIVER TRIPS BY PURPOSE OBTAINED FROM
O-D SURVEY AND ESTIMATED BY MODEL FOR SMALL AREA

Purpose	Zone ^a	Productions		Attractions	
		O-D Survey	Model Estimations	O-D Survey	Model Estimations
Nonhome-based	All	3,122	3,367	3,113	3,379
	714	277	320	314	365
	715	239	282	250	309
	716	442	459	416	450
	720	696	744	684	670
	721	261	251	260	257
	722	1,207	1,311	1,189	1,328
Home-based	All	10,233	10,209	5,497	5,271
	714	1,303	1,342	511	457
	715	1,954	1,953	272	273
	716	1,914	1,927	543	562
	720	2,827	2,834	1,062	1,039
	721	1,421	1,371	469	506
	722	814	782	2,620	2,434
Shop	All	1,552	1,163	461	395
	714	180	77	17	28
	715	276	153	16	26
	716	530	385	153	61
	720	337	558	136	158
	721	174	127	64	52
	722	45	63	75	70
Social-recreation	All	1,446	1,323	842	773
	714	229	153	119	111
	715	205	141	55	54
	716	439	426	207	255
	720	344	356	297	215
	721	144	133	106	55
	722	85	114	58	83
Miscellaneous	All	2,145	1,848	1,350	1,361
	714	257	211	98	100
	715	363	310	132	137
	716	431	393	187	156
	720	577	535	339	325
	721	313	250	152	162
	722	184	149	462	481

^aThese zones are in the Hough area.

TABLE 4
 AUTOMOBILE-DRIVER TRIPS BY PURPOSE OBTAINED FROM O-D SURVEY AND
 ESTIMATED BY MODEL FOR CBD

Purpose	Productions			Attractions			Internal		
	O-D Survey	Model Estimations	Model as Percent of O-D	O-D Survey	Model Estimations	Model as Percent of O-D	O-D Survey	Model Estimations	Model as Percent of O-D
Nonhome-based	45,096	44,474	98.5	41,528	41,673	100.1	995	960	96.5
Home-based									
Work	1,408	1,698	120.5	80,942	81,581	100.7	0	30	—
Shop	98	236	241.0	13,954	13,364	95.8	0	4	—
Social-recreation	298	344	115.0	10,738	12,022	112.0	0	0	—
Miscellaneous	482	758	157.0	20,086	20,612	97.5	0	23	—
Total, includes school	47,432	47,529	100.4	169,926	170,165	100.1	1,000	1,018	101.8

area. It is not expected that the model would reproduce the values for individual zones as well as those for the entire region, but the comparison does indicate that it functions well in all areas with little or no systematic bias.

Table 4 gives the estimates for the CBD. For those purposes with significant volumes, the model functions quite well, although there are some larger percentage discrepancies in low-volume cases. The total estimate is again quite satisfactory.

APPLICATION OF THE MODEL

After it was established that the model could indeed reproduce the base-year trip patterns, the next step was the projection of future automobile-driver trip patterns. The trip-generation model produced the total aggregate trip ends that were then allocated to the zones using the direct-trip allocation model (1), converted to productions and attractions by individual purposes using an interface model, and linked to form interchanges using a gravity model. A transit-nontransit modal split was performed using the same model with different parameters, and the nontransit table was used as input into the automobile-driver modal-split model.

Production-area income codes for the design year were established from projections by census tract of median family income. The groupings used for the future year were low (below \$9,000), medium (\$9,000 to \$12,000), and high (over \$12,000).

The attraction-area codes were the same as those in the base year, CBD and non-CBD. In order to simplify use of the model, it was necessary to produce dummy curves that assigned 100 percent of the truck-taxi trips to the vehicle-driver trip table.

The results of the projection of automobile-driver trips as represented by the average number of persons per car are given in Table 5. The results indicate that average automobile occupancy is not only dropping but dropping at different rates for the various purposes. The significance of this change can be shown in the following example: If the average occupancy rate had been 1.42 rather than 1.32, the number of automobile-driver trips would have been decreased by an amount nearly equal to the total number of trips projected to use public transit in the design year.

In this application, the transit-split model and the driver-split model were run in series because this offered a logical flow (Fig. 18). The trips are first divided into highway and

TABLE 5
 AUTOMOBILE OCCUPANCY BY PURPOSE OBTAINED
 FROM O-D SURVEY AND ESTIMATED BY MODEL

Purpose	Average No. of Persons per Car	
	O-D Survey	Model Estimations
Nonhome-based	1.18	1.09
Home-based		
Work	1.51	1.38
Shop	2.00	1.68
Social-recreation	1.37	1.27
Miscellaneous	1.30	1.21
Average	1.45	1.32

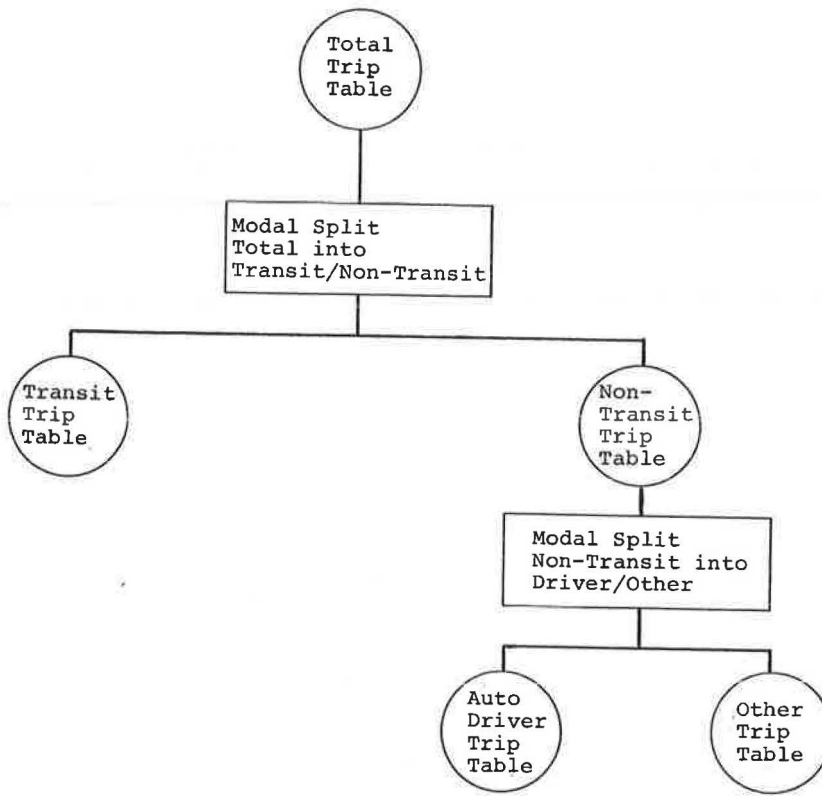


Figure 18. Series operation of modal-split model application.

transit trips. The highway portion is then further divided into automobile-driver trips and all others. This is the same procedure used in the past when an occupancy factor was used. One difficulty in this method is that the final output of the second split (automobile-driver trips) is dependent on the first split. This increases the probability for error in the final output.

There are other equally logical approaches using multimodal-split models that might be explored. One would be to run the models in parallel (Fig. 19). With this procedure, each output would be independent of the other split, and this would probably decrease the error in the final trip tables. It would, however, require the additional normalization because each of the models uses different parameters, and there is no assurance that the total of the three output trip tables will be equal to the original trip table.

Another approach would be a series operation with the trips first being split into automobile-driver trips and others. The other trips would then be split into transit and nontransit trips. This approach might be more desirable because the number of vehicle trips is far more sensitive to the driver split than to the transit split. The logic of this choice pattern would have to be analyzed further.

No tests have been performed to compare these various methods of operation. The obvious test would be to start with a full base-year table and split it in each of the three ways described. The method that produced the least overall error would seem most desirable for use. It also might give insight into the true modal-choice, decision-making process.

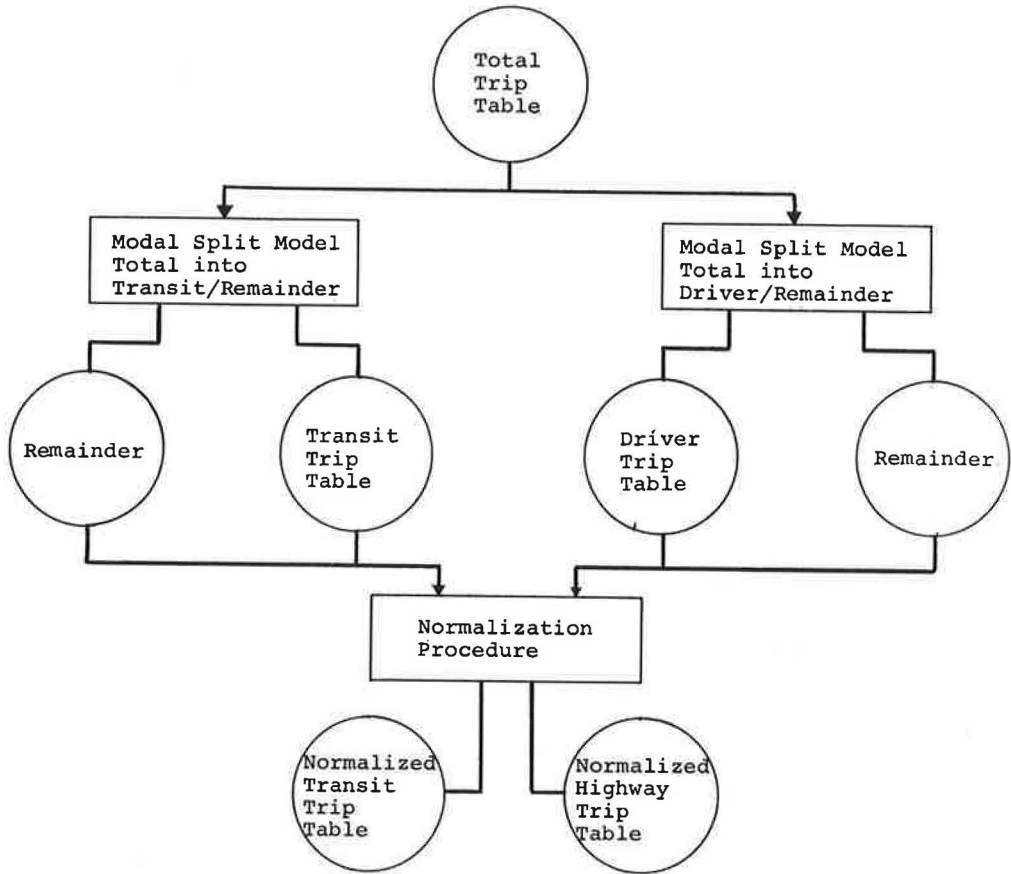


Figure 19. Parallel operation of modal-split model application.

CONCLUSIONS

The following major conclusions may be drawn from this study:

1. A modal-split model is a desirable method for predicting automobile occupancy.
2. It is relatively easy to develop and calibrate diversion curves that will accurately reproduce base-year data.
3. A great deal of effort has been spent on analysis of transit modal split, but there has been relatively little work on automobile-driver modal split. The latter, however, is a far more significant item in estimating future vehicular travel.

The great shift in average automobile occupancy indicates that we can no longer be satisfied with conversion of person trips to vehicle trips by factors held consistent from the base year. A full analysis is needed to determine not only the magnitude of this shift but also the causes. If the causes are known, it may prove possible to design systems that will promote higher occupancy levels, allowing the entire highway networks to serve more efficiently.

REFERENCE

1. Lathrop, G. T., Hamburg, J. R., and Young, G. F. Opportunity-Accessibility Model for Allocating Regional Growth. Highway Research Record 102, 1965, pp. 54-66.

License Plate Traffic Survey

HOWARD McCANN and GARY MARING, Current Planning Division,
U.S. Bureau of Public Roads

This report describes a license plate, origin-destination traffic survey that was conducted in the area of Boston, Massachusetts, on January 13 and 14, 1968. During the survey, vehicle license plate numbers were recorded at four highway stations. A computer search determined the names and addresses of vehicle owners. Mail questionnaires were sent to 4,910 vehicle owners requesting information on these trips. Background information, operating procedures during the survey, results of the survey, and conclusions are presented.

•THE COLLECTION and analysis of origin-destination data are essential to the highway planning process. Home interviews and roadside interviews are currently the basic sources of these data.

A method of surveying in which license plates are recorded is made economically possible by use of a computerized vehicle registration system. In this scheme license plates are recorded as vehicles pass highway stations, and questionnaires are sent to vehicle owners requesting information on the specific trips. The method depends on a rapid and inexpensive look-up of vehicle owners' names and addresses. Automated vehicle registration files, with plate numbers and owner identification on data processing cards or magnetic tape, make the look-up feasible.

License plate origin-destination studies were conducted by the California Division of Highways as part of a freeway traffic survey in Los Angeles, and by the Bay Area Transportation Study in San Francisco. Available information indicates that the surveys were successful.

In order to evaluate the procedure more fully, a field test was conducted in Massachusetts, where the Registry of Motor Vehicles maintains an automated vehicle registration file. Two methods of recording plate numbers were tested, visual and photographic, prior to the actual survey. These methods were evaluated under the conditions of high vehicle speeds and heavy traffic volumes.

During the visual tests, one man equipped with field glasses read the numbers into a tape recorder. In another visual test, the numbers were recorded on coding sheets by a second man. Both visual tests had two basic flaws: (a) all license numbers could not be recorded when platoons of vehicles passed the station; and (b) multiple recording errors are likely to be introduced between the time the license plates are viewed and the time the keypunching of the plate numbers into data processing cards is completed.

Kodak Cine-Special 16-mm motion picture cameras were used in the photographic tests. The cameras were positioned both on the shoulder of the road and on the overpass structures. Several film types, shutter speeds, and lenses were used.

The camera position on the overpass structures appeared best, although shoulder operation could be feasible under certain conditions. The best pictures were obtained with a 1/400 sec shutter speed, a 6-in. telephoto lens, and a fast black and white film. Kodak TRI-K Reversal film, type 7278, was used: this film has an ASA rating of 200.

U.S. DEPARTMENT OF TRANSPORTATION
 FEDERAL HIGHWAY ADMINISTRATION
 BUREAU OF PUBLIC ROADS
 WASHINGTON, D.C. 20591

Dear Car Owner:

The Bureau of Public Roads is conducting research to help engineers improve intercity highways. Your help is needed to tell us where your car was coming from and going to on _____ when it was seen traveling out-bound from _____. Listed below are several questions concerning that trip. We would appreciate it very much if you would answer these questions and mail this form back to this office. No postage is required.

We cannot complete this important research unless you help by returning this information. Your cooperation is most important. You may be assured that the requested information concerning the travel of your vehicle will be held in the strictest confidence.

Thank you for your help.

Sincerely yours,

F. C. Turner

F. C. Turner
 Director of Public Roads

Questionnaire

1. What was the main purpose for this trip? (check one)
 - Earning a living
 - Family business
 - Social, recreational
 - Educational, civic, religious

2. Composition of travel party:

Total number of occupants including driver	_____
Age group	Number
Under 5	_____
5 - 18	_____
19 - 65	_____
Over 65	_____

3. In what place (city or township, and State) did your overall trip begin?
 - City or township _____
 - State _____

4. If this was an overnight trip, where did your trip start on the above date?
 - City or township _____
 - State _____

5. What place was the overall destination of this trip? This is normally the farthest point to which the vehicle was driven.
 - City or township _____
 - State _____

6. If the trip continued beyond the above date, at what place did you stop on the evening of this date?
 - City or township _____
 - State _____

Figure 1. Questionnaire for drivers of passenger cars.

U. S. DEPARTMENT OF TRANSPORTATION
 FEDERAL HIGHWAY ADMINISTRATION
 Bureau of Public Roads
 Washington, D.C. 20591

Vehicle License _____

Dear Truck Owner:

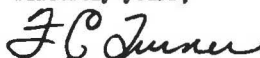
The Bureau of Public Roads is conducting research to help engineers improve intercity highways. Your help is needed to tell us where your truck was coming from and going to on _____ when it was seen traveling outbound from _____.

Listed below are several questions concerning that trip. We would appreciate it very much if you would answer these questions and mail this form back to this office. No postage is required.

We cannot complete this important research unless you help by returning this information. Your cooperation is most important. You may be assured that the requested information concerning the travel of your vehicle will be held in the strictest confidence.

Thank you for your help.

Sincerely yours,



F. C. Turner
 Director of Public Roads

Questionnaire

1. Was this an overnight trip? Yes No
2. Where did the trip start and end on the above date?
Start City or township _____ State _____
End City or township _____ State _____
3. Where did your overall trip start and end?
Start City or township _____ State _____
End City or township _____ State _____
4. Was this trip made to or from a loading or unloading point? Yes No
5. During this trip, was the vehicle empty or loaded? Empty Loaded
6. What was the weight of the vehicle on this trip?
 Gross weight _____ weight of load _____
 Empty weight: Power unit _____ Trailer _____
 Are these scale weights or estimated weights ?
7. If loaded, what commodity was carried?
 Farm, forest and fish products , Sand, gravel, ores, petroleum, minerals ,
 Processed foods, beverages and tobacco , Manufactured products ,
 Mixed freight , Waste material
8. Axle arrangement of vehicle (check one)









<p>a. <input type="checkbox"/>  single <input type="checkbox"/> dual <input type="checkbox"/></p> <p>b. <input type="checkbox"/> </p> <p>c. <input type="checkbox"/> </p>	<p>d. <input type="checkbox"/> </p> <p>e. <input type="checkbox"/> </p> <p>f. <input type="checkbox"/> </p>	<p>g. <input type="checkbox"/> </p> <p>h. <input type="checkbox"/> </p> <p>i. <input type="checkbox"/> Other</p>
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
9. Body type:
 Pickup , flatbed , rack or livestock , dump , van ,
 refrigerator van , tank , Other
10. Is this multi-stop vehicle used for delivery or pickup? Yes No

Figure 2. Questionnaire for drivers of trucks.



Figure 3. Location of survey sites.

Sharp images of state names could be produced when the camera was set to cover one lane. When two lanes were covered the state names were not clearly legible although the numbers were legible. Kodak Recordak 310 film readers were used to view the film. This is a desk-size reader, with a 9- by 12-in. screen.

The photographic method of recording plate numbers was more complex than the visual method, but the chance of error was less in that cards were keypunched from the basic data source. In addition, all vehicles could be recorded regardless of traffic volumes or speeds. These advantages led to the decision to use the photographic method during the Massachusetts survey. During the tests and the survey, the film was exposed manually when vehicles passed focus points on the highway. Bursts of several frames were exposed for each vehicle. An electronic switch is being developed that exposes one frame as a vehicle crosses a road tube.

SURVEY PROCEDURES

Questionnaire Design

Two questionnaires were prepared, one for owners of passenger cars and one for owners of trucks (Figs. 1 and 2). The questionnaires for automobile owners requested information concerning trip purpose, composition of travel party, and origin and destination of trip. The survey was related to intercity travel; therefore, the origin and destination were requested by city or township and state. The questionnaire for truck owners requested information on trip origin and destination, commodity, weight, axle arrangements, and body type.

Survey Sites and Plate Number Recording

The four survey sites selected were located on I-93 and I-95 north of Boston, on I-95 south of Boston, and on the Massachusetts Turnpike west of Boston (Fig. 3). The sites were considered representative of high-volume intercity highways. The survey was conducted for several hours on Saturday and Sunday, January 13 and 14, 1968. These days were chosen in anticipation of higher traffic volumes.

License plates were recorded of vehicles traveling outbound from Boston. Cameras were used at all sites except at the site on the Massachusetts Turnpike where plate



Figure 4. Position of cameras on overpass structure.

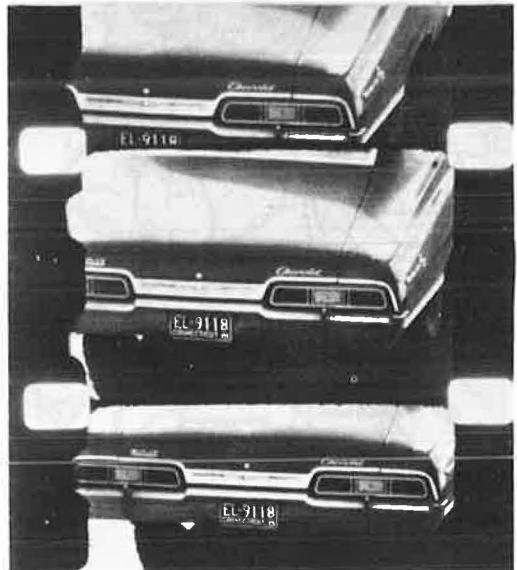


Figure 5. Photograph of license plate.



Figure 6. Film reader and keypunch operation.

numbers were recorded manually when vehicles stopped at toll gates. Figure 4 shows camera positions on the overpass structure at I-95 south of Boston.

Keypunching Operation

The film was developed on the evening of January 14, and keypunching started that night. Both tasks were performed by private companies. Figure 5 shows a view of the film produced during the survey. Five Recordak film readers were rented from Kodak and were used by the keypunch operators to view the film. Figure 6 shows the reader and the keypunching operation.

One data processing card was punched for each vehicle. For Massachusetts vehicles, nine digits were punched on each card—six digits for the license number, one digit for station location, and two digits for job identification. For out-of-state vehicles, the state name abbreviation was also punched. The operators manually advanced the film and were able to punch about 300 cards per hour. Occa-

sionally, plate numbers could not be distinguished on the film because the camera lens opening was improperly set, the vehicles changed lanes, or the plates were dirty. Keypunching was completed that night and the computer search started the next morning, January 15, at the Massachusetts Registry of Motor Vehicles.

Computer Search for Names and Addresses of Vehicle Owners

During the computer operation, the input cards were first converted to magnetic tape. The data records were sorted by license plate number to match the order of plate numbers in the state's master file. The sorted tape was then compared with the master tapes. When a match occurred, the vehicle owner's name and address was placed on a tape that was used later for printing.

The Massachusetts vehicle file has about three million entries on 35 reels of magnetic tape. Each entry contains 168 characters. About seven hours of computer time is required to make a search through the entire file. This time is substantially the same regardless of how many plate numbers must be matched and addresses determined.

TABLE 1
QUESTIONNAIRES MAILED

State	Number
Massachusetts	4,130
Automobile owners	4,025
Truck owners	105
Connecticut	150
Maine	41
New Hampshire	150
New Jersey	39
New York	98
Rhode Island	287
Vermont	15
Total	4,910

TABLE 2
RESPONSE RATE FOR MASSACHUSETTS
VEHICLE OWNERS

Vehicle Owners	Number	Percent
Passenger car owners		
Total questionnaires mailed	4,025	100.0
Returned	2,609	64.8
Usable	2,427	60.3
Unusable because of recording errors	69	1.7
Undelivered	32	0.8
Unusable for other reasons ^a	81	2.0
Trucks owners		
Total questionnaires mailed	105	100.0
Returned	62	59.0

^aSuch as vehicle was leased or questionnaire was incomplete.

TABLE 3
RESPONSE RATE FOR OUT-OF-STATE
VEHICLE OWNERS

State	Mailed	Returned	Response Rate (percent)
Connecticut	150	90	60.0
Maine	41	32	78.0
New Hampshire	150	88	58.7
New Jersey	39	24	61.5
New York	98	47	48.0
Rhode Island	287	178	62.0
Vermont	15	7	46.7

After the computer search was completed, the vehicle owners' names and addresses were printed. Both a listing and gummed labels were printed from the tape. Of the 4,428 Massachusetts plate numbers read into the computer, 4,130 were matched with numbers in the file. The remaining 298, 6.7 percent of the vehicles, were missed either because there were errors in keypunching or because there were no records in the computer file. The computer file is updated frequently, but a small lag exists at all times.

The state's operating costs for its GE 415 computer are approximately \$75 per hour, which represents all cost items including the leasing fee. The total cost to search the file, for any number of license plates, is therefore about \$500. The extensive search time for relatively few entries would be greatly reduced with a random-access storage device, such as a data cell, disk, or drum. With this device, records can be accessed directly rather than serially. In states with this computer capability, the retrieval of relatively few addresses could be accomplished in considerably less time than seven hours.

Questionnaire Mailing Operation

Gummed labels were manually attached to questionnaires. The first batch was mailed 30 hours after filming was completed. The remaining were mailed over a period of several days to measure the response rate as a function of the elapsed time between the trip and the receipt of the questionnaire. The mailing to Massachusetts residents was completed on January 20, one week after the first day of the survey.

License plate numbers of out-of-state vehicles were furnished to the appropriate states. Mailing for the out-of-state vehicles was completed on January 26. The number of questionnaires mailed to Massachusetts residents and to residents of other states is given in Table 1.

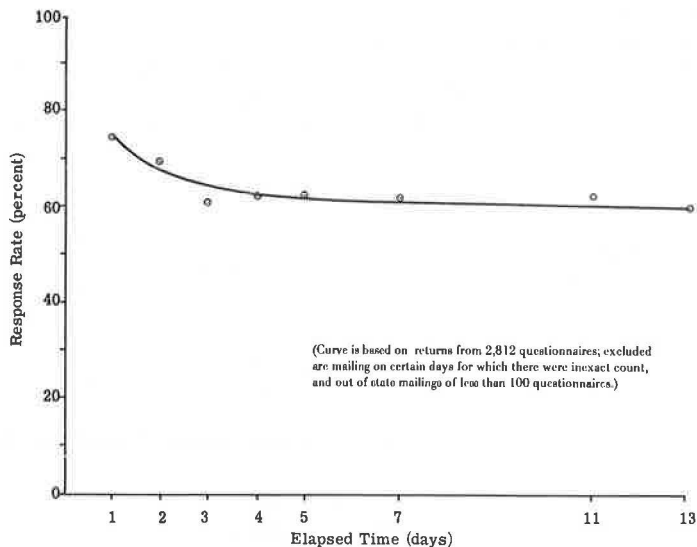


Figure 7. Response rate of automobile owners by elapsed time between survey and day of mailing.

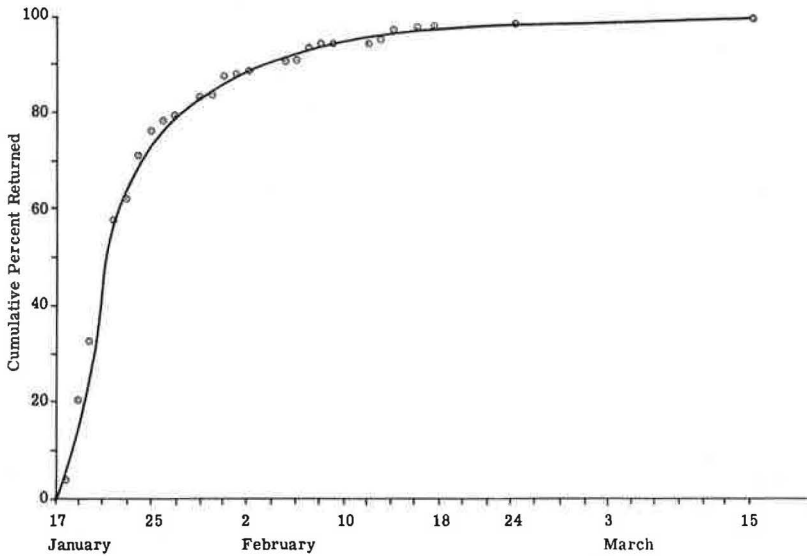


Figure 8. Date of response by automobile owners for 700 questionnaires mailed January 15, 1968.

RESULTS OF THE SURVEY

Table 2 gives the overall response rate for Massachusetts vehicle owners. The automobile-owner response rate of 64.8 percent was reduced to 60.3 percent when unusable responses were eliminated. Table 3 gives the overall response rate for out-of-state vehicle owners.

Figure 7 presents the overall automobile-owner response rate as a function of time elapsed between the survey and the mailing of the questionnaires. The response rate decreased from 74.0 percent for the questionnaires mailed 1 day after the survey to 60.0 percent for the questionnaires mailed 13 days later. Figure 8 shows the time of response for a group of 700 questionnaires mailed on January 15. Eighty percent of the eventual responses were received in Washington within 12 days after the mailing in Boston.

Characteristics for nonrespondents must be considered in the expansion of sample results. One item of information, the distribution of the distance from their homes to locations observed, was known for all Massachusetts vehicle owners in the survey. A dissimilarity in this distribution for respondents and nonrespondents would indicate possible differences in trip length distribution. Such a finding would prohibit a uniform expansion of sample data. A chi-square analysis showed that the home-to-station distributions for the respondents and nonrespondents could be accepted as similar on the 0.10 confidence level (Tables 4 and 5 in the Appendix). Summaries of the data reported by the respondents are given in Tables 6 and 7 in the Appendix.

CONCLUSION

This pilot study indicates that the license plate traffic survey is a feasible method of obtaining travel characteristics of motorists. The response rate indicates the motorists' willingness to cooperate in this type of survey. A limited analysis of the distribution of distance from registered address to highway station revealed no apparent difference between the respondents and nonrespondents.

The method used to record license plate numbers depends on specific conditions. Where vehicle volumes and speeds are high, the photographic procedure appears to be the best.

To be feasible, the survey requires a rapid and inexpensive look-up of vehicle owners' names and addresses. Such a look-up is practical only in those states with an automated vehicle registration file.

In designing such a survey, one must consider those vehicles for which data will be difficult to obtain. These include vehicles registered in other states or those owned by leasing companies, government agencies, or private companies.

License plate surveys have potential in other types of transportation studies. The plate look-up provides a mailing address for each vehicle observed at any location. Thus a trip table could be prepared for vehicles observed at parking lots in shopping centers, business districts, or recreational areas. Questionnaires could be sent to a small sample of vehicle owners to provide information on trip purpose, vehicle occupancy, and other characteristics.

ACKNOWLEDGMENTS

Richard E. McLaughlin, Massachusetts Registry of Motor Vehicles, provided the services of the Registry in making the computer search for vehicle owners' names and addresses. The Massachusetts Turnpike Authority permitted personnel to be stationed at the toll booths. The Massachusetts Department of Public Works provided special traffic counts on the highways in the survey. The motor vehicle departments of Connecticut, Maine, New Hampshire, New Jersey, New York, Rhode Island, and Vermont mailed the questionnaires for vehicles from these states.

From the Bureau of Public Roads, Leon Litz, Methods Branch, aided in the technical design of the study and in discussing and resolving problems relative to the study; Phillip Robinson, Massachusetts Division, assisted extensively during the survey; and William Hall and George Crum, Publications and Visual Aids Branch, were responsible for the photography during the survey.

Appendix

(The tables on the following pages give the results of a chi-square analysis of the survey data and travel characteristics of vehicle owners who responded to the questionnaires.)

TABLE 4
CHI-SQUARE ANALYSIS—STATION I-95S CALCULATIONS

Home to Station (miles)	Respondents		Nonrespondents		Total Observed
	Observed	Expected	Observed	Expected	
0 to 4	38	38.1	23	22.9	61
5 to 9	91	86.9	48	52.1	139
10 to 14	109	117.5	79	70.5	188
15 to 19	56	60.6	41	36.4	97
20 to 29	92	83.8	42	50.3	134
30 to 39	25	21.9	10	13.1	35
40 to 49	11	10.0	5	6.0	16
50 and over	3	6.3	7	3.8	10
Total	425	—	255	—	680

TABLE 5
CHI-SQUARE ANALYSIS—ALL STATIONS

Station	Chi-Square	Degrees of Freedom	Level of Significance
I-95S	11.11	7	0.10
Mass. Turnpike	7.90	8	0.10
I-95N	4.16	6	0.50
I-93N	3.20	6	0.70

Note: Chi-square = $\sum \frac{(O-E)^2}{E} = 11.11$, where O is observed and E is expected. With 7 degrees of freedom, this is significant to the 0.10 level.

TABLE 6
TRAVEL CHARACTERISTICS OF PASSENGER AUTOMOBILE OWNERS

Characteristic	Percent	Characteristic	Percent
Purpose of trip		Age group of occupants (cont'd)	
Earning a living	17.6	19 to 64	72.9
Family business	35.1	65 and over	4.2
Social, recreational	39.9	Trip length distribution, miles	
Educational, civic, religious	7.4	0 to 4	0.0
Number of occupants		5 to 9	0.4
1	34.0	10 to 14	3.4
2	35.3	15 to 19	9.0
3	13.4	20 to 29	25.1
4	9.7	30 to 39	15.3
5	4.6	40 to 49	11.5
6 and over	3.0	50 to 99	19.8
Age group of occupants		100 to 249	14.0
0 to 4	6.5	250 to 499	1.3
5 to 18	16.4	500 to 999	0.2
		1,000 and over	0.0

TABLE 7
TRAVEL CHARACTERISTICS OF MASSACHUSETTS TRUCK OWNERS

Characteristic	Percent	Characteristic	Percent
Overnight trip		Type of commodity	
Yes	3.5	Farm, forest	7.7
No	96.5	Gravel, petroleum	34.6
Empty or loaded vehicle		Proc. foods, beverages	19.2
Empty	16.1	Manufactured goods	28.9
Loaded	64.3	Mixed freight	5.8
Both	19.6 ^a	Waste material	—
Trip to or from loading or unloading point		Other	3.8
Yes	85.5	Vehicle used for pickup or delivery	
No	14.5	Yes	68.1
Gross weight of vehicle, thousand lb		No	25.5
0 to 5	19.2	Both	6.4 ^a
5 to 10	15.4	Axle arrangement (Fig. 2)	
10 to 20	5.8	2S	45.5
20 to 30	11.5	3A	3.6
30 to 40	3.9	2S1	7.3
40 to 50	9.6	2S2	25.4
50 to 60	5.8	3S1	—
60 to 70	9.6	3S2	16.4
70 and over	19.2	2-2	—
Weight of load, thousand lb		3S1-2	—
0 to 5	41.7	Other	1.8
5 to 10	6.3	Body type	
10 to 20	6.2	Single pickup	10.9
20 to 30	10.4	Dual pickup	—
30 to 40	6.2	Flatbed	—
40 to 50	25.0	Rack or livestock	3.6
50 to 60	4.2	Dump	—
60 to 70	—	Van	34.5
70 and over	—	Refrigerator van	7.3
		Tank	36.4
		Other	7.3

^a Both blocks were occasionally checked on these questions because the respondents were not sure on which leg of the round trip they were observed.