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-

Paper sponsored by Committee on Origin and Destination.

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	0.898	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

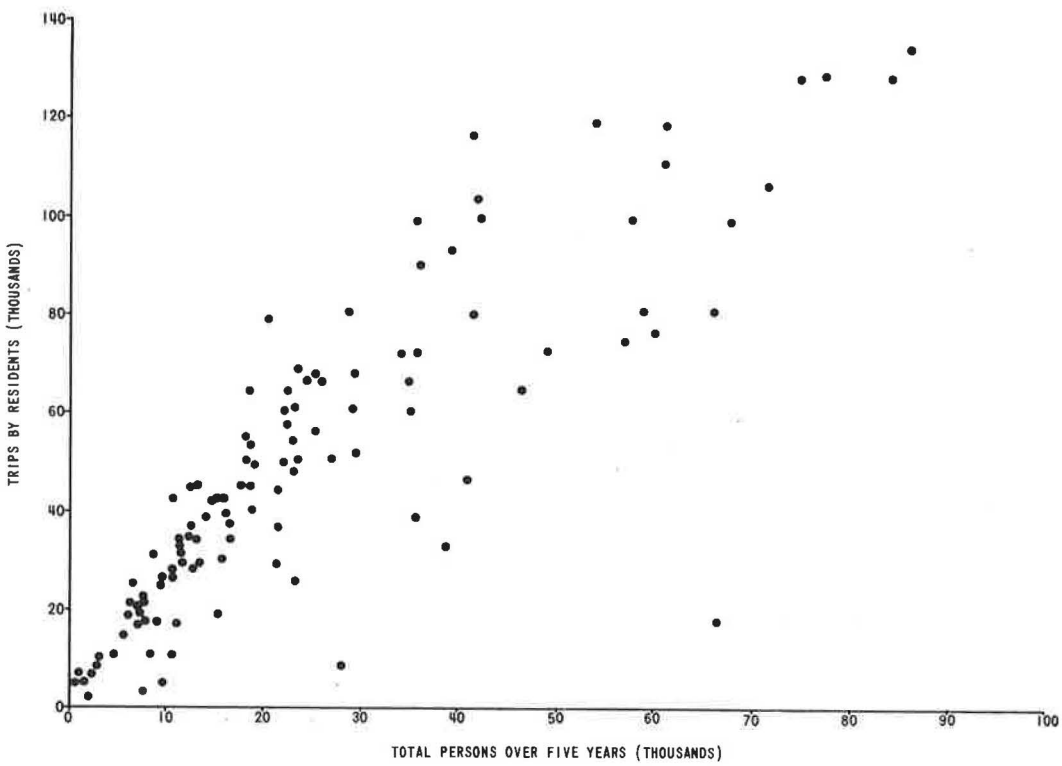

Figure 1. Relationship between trips per zone generated by residents and persons per zone.

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

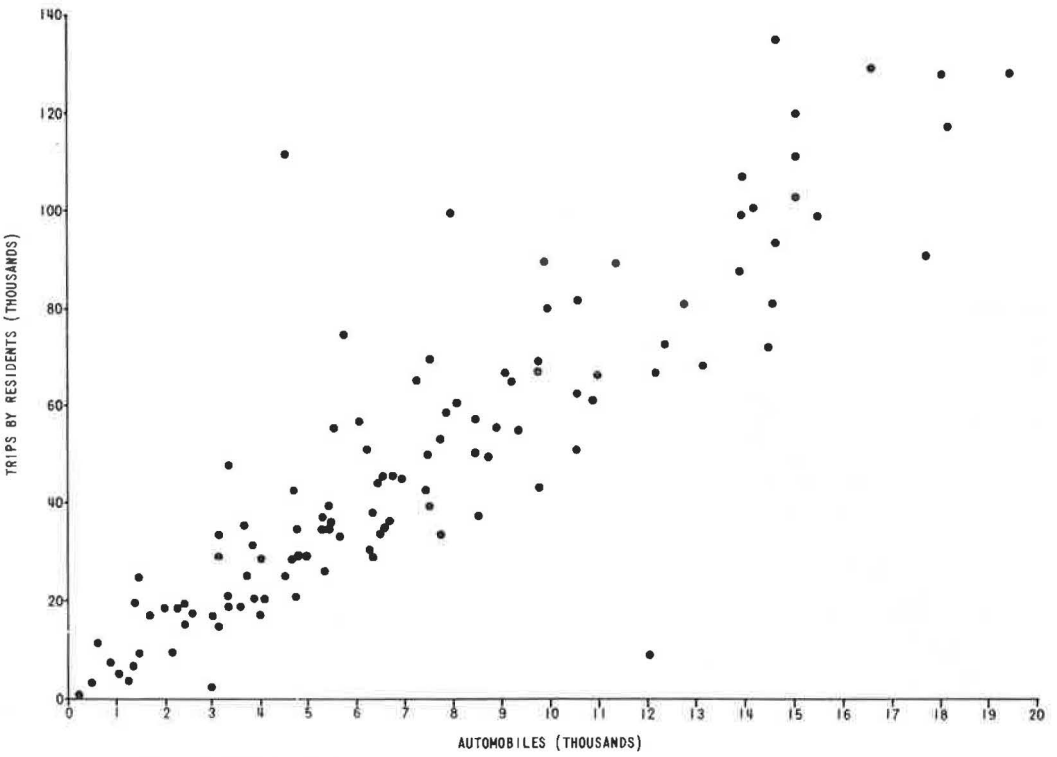

Figure 2. Relationship between trips per zone generated by residents and automobiles per zone.

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.

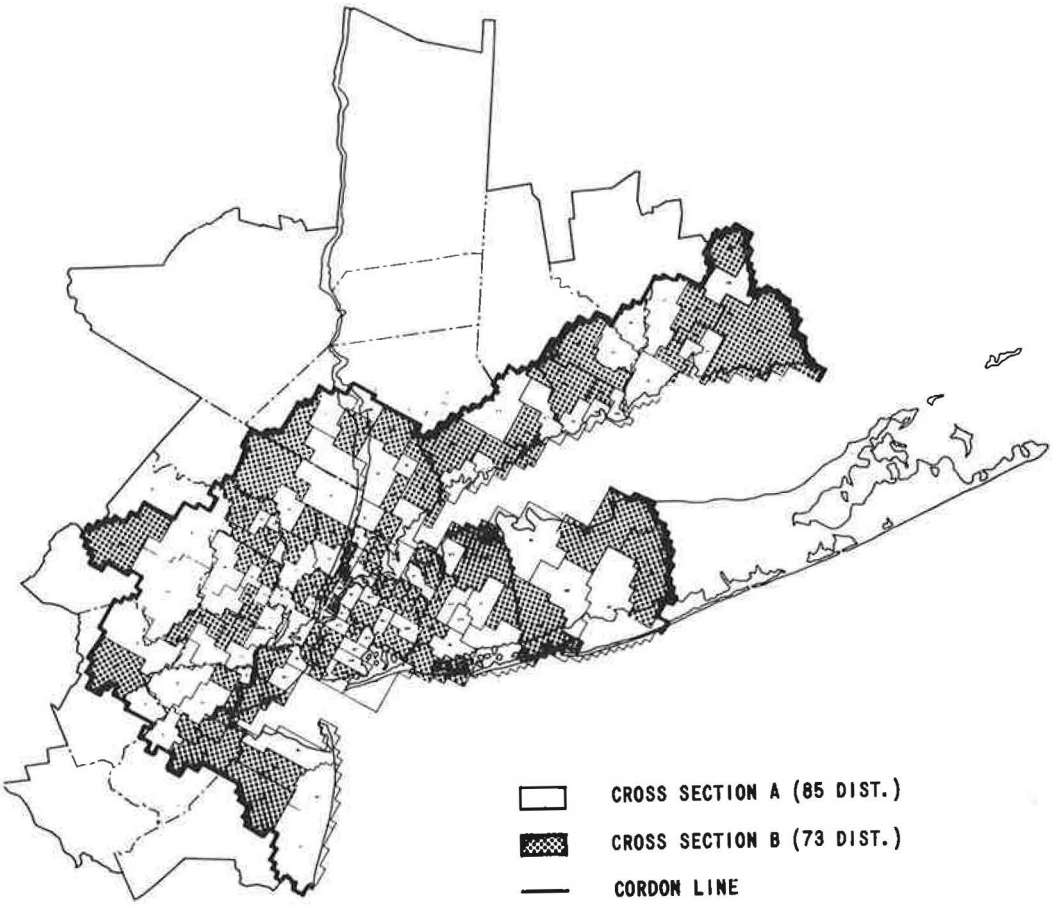

Figure 3. Spatial distribution of the two data sections.

TABLE 2

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

Analysis Unit	No. of Observations	Dependent Variable Y	Independent Variables X_1	Independent Variables X_2	Equation	Constant/ Mean	Standard Error/ Mean	R^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

trips per zone = f (persons per zone and households per zone)

and the zonal aggregate rates

trips per household per zone = f (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 R^2 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

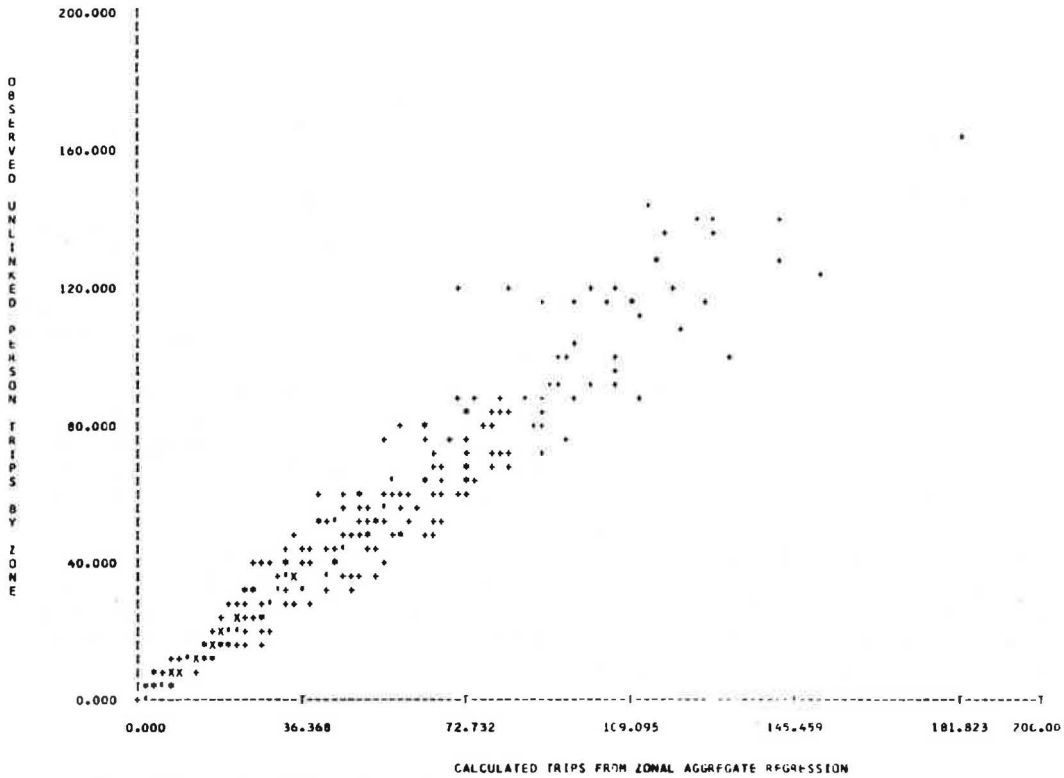

Figure 4. Comparison of observed vs predicted zonal trips generated by residents for zonal aggregate regression.

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:

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

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 nonresidential 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.)

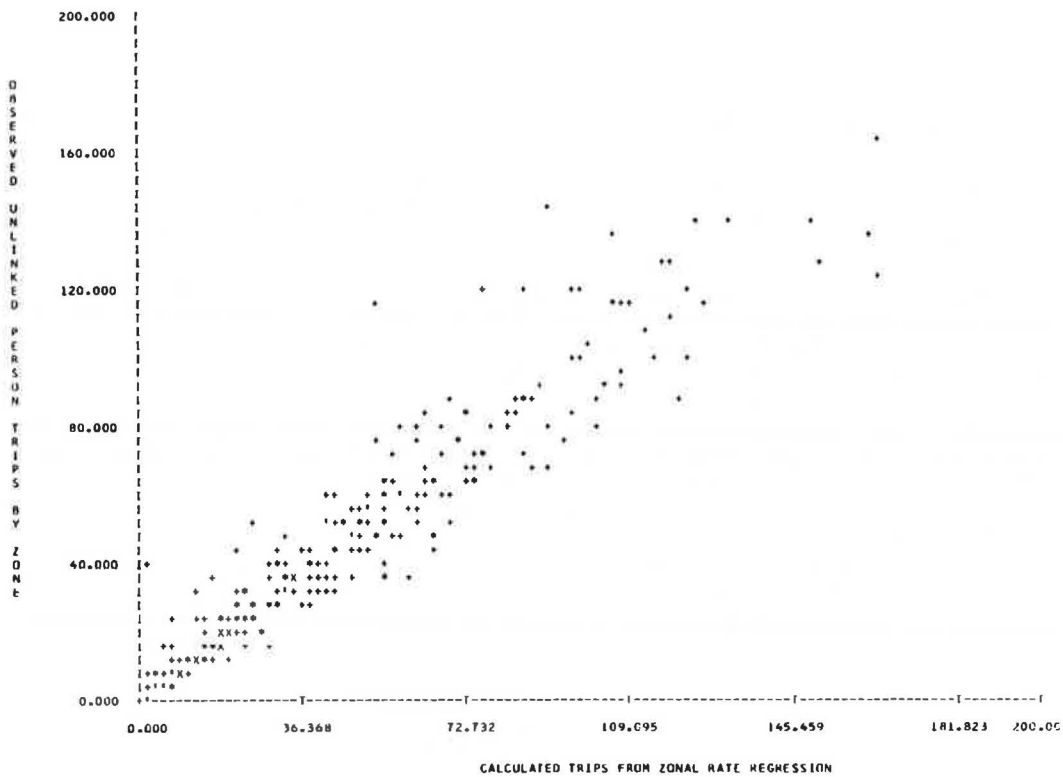

Figure 5. Comparison of observed vs predicted zonal trips generated by residents for zonal rate regression.

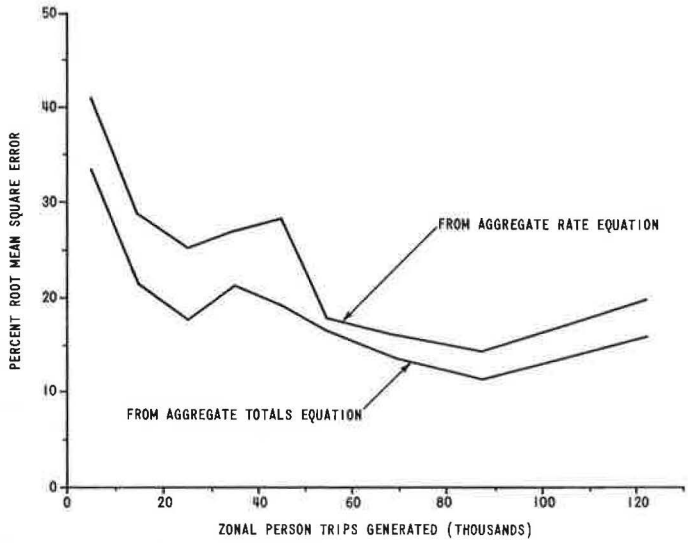

Figure 6. Root-mean-square error vs trip volume for two trip-generation techniques.

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, of 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

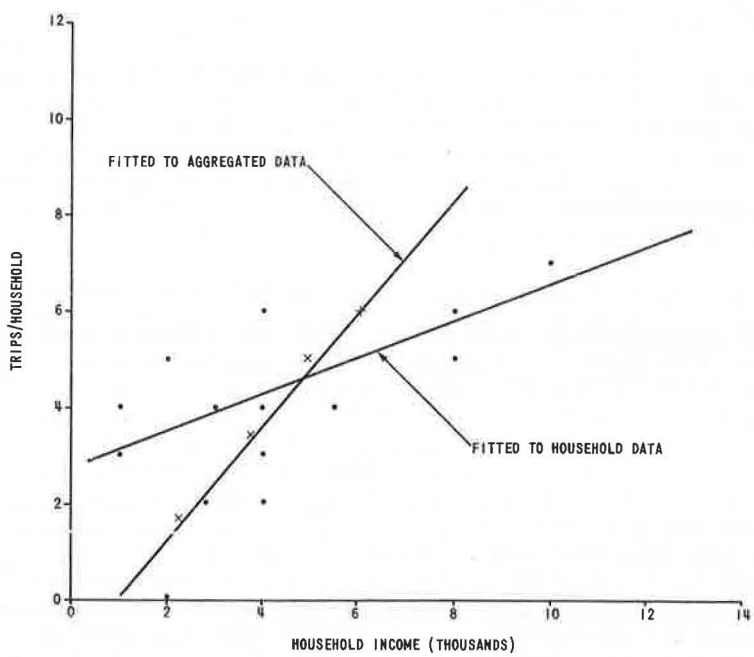

Figure 7. Relationships that may be derived from the same data at varying levels of aggregation (hypothetical case).

TABLE 5

VARIATION WITHIN DATA AT DISAGGREGATED AND AGGREGATED LEVELS

(Hypothetical Case)

Zone	Household	Total Trips per Household		Average Trips per Zone	
		T_h	$(T_h - \overline{T}_h)^2$	T_z	$(T_z - \overline{T}_z)^2$
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,

Trips per household per district = f (persons per household per district and automobiles per household per district)

the zonal rate,

Trips per household per zone = f (persons per household per zone and automobiles per household per zone)

and the household rate.

Trips per household = f (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 X_1	Independent Variables X_2	Equation	Constant/ Mean	Standard Error/ Mean	R^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 R^2 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:

$$B_i = bi \frac{Sx_i}{Sy}$$

where

B_i = beta coefficient,
b_i = regression coefficient for the ith independent variable,
Sx_i = standard deviation of the distribution of the ith 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.

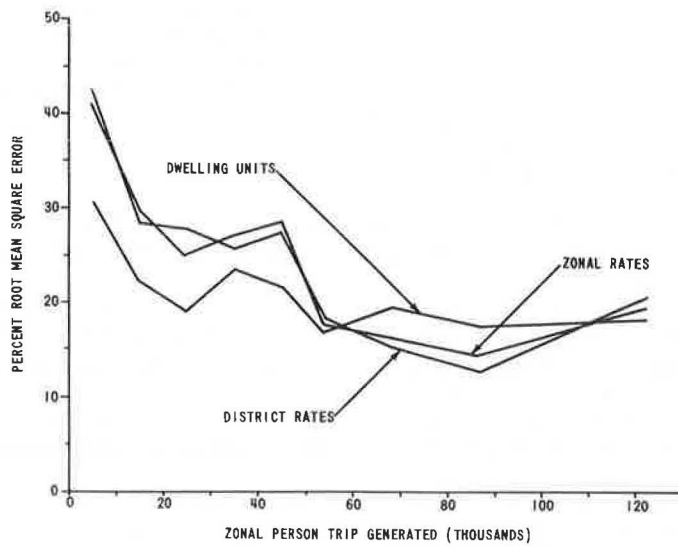

Figure 8. Root-mean-square error vs trip volume for dwelling unit analyses and for aggregate rate analyses.

TABLE 9

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

Aggregation	Beta Coefficients		B_A/B_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 disagreggate 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.