

# Trip Generation by Cross-Classification: An Alternative Methodology

PETER R. STOPHER AND KATHIE G. McDONALD

An alternative methodology for calibrating cross-classification models, namely multiple classification analysis (MCA), is described. This technique, which has been available in the social sciences for some time, does not appear to have been used in transportation planning before, although it appears to be able to overcome most of the disadvantages normally associated with standard cross-classification calibration techniques. The MCA procedure is described briefly, and its merits—in terms of statistical assessment, ability to permit comparisons among alternative models, and lack of susceptibility to small samples in individual cells—are discussed in detail. In addition, the method is based on analysis of variance (ANOVA), which provides a structured procedure for choosing among alternative independent variables and alternative groupings of the values of each independent variable. These procedures are contrasted with standard procedures for cross-classification that estimate cell values by obtaining the average value of the dependent variable (e.g., a trip rate) for those samples that fall in the cell and are unable to use any information from any other cell. The process of selecting independent variables and selecting groupings of the chosen variables by ANOVA is illustrated with a case study. In this study the way in which this process works, and the degree to which there is statistical information provided to guide the analyst's judgment, is shown. In the case study the confirmation of intuitive selections of variables is noted, and also a more surprising result is produced that shows that the best household grouping is one that combines two- and three-person households. A second case study illustrates the use of MCA to calculate trip rates. A comparison of the conventional procedure of cell-by-cell averaging, a MCA design that does not account for interactions among the independent variables, and a MCA design that corrects for interactions is given. It is shown that the MCA allows trip rates to be computed for some cells that are empty of data, and that MCA removes some possibly spurious rates that arise in the conventional method from small sample problems in some cells. It is concluded that MCA provides a strong methodology for cross-classification modeling and that the procedure is effective in surmounting most of the drawbacks of conventional estimation of such models.

In the 1950s and 1960s most of the transportation planning studies developed trip-generation equations that used linear regression, particularly for person trip-production models. Linear regression was so strongly favored that it was the central method in the FHWA guide to trip-generation analysis (1). Initially, most of the trip-production models were formulated to provide an estimate of zonal trips as a function of zonal variables that describe households. These models were increasingly the subject of criticism, particularly because of the loss of variance from the extremely aggregate nature of these models (2,3). As a result, household models of trip production were developed, in which the dependent variable became average daily trips per household, possibly by purpose, as a function of attributes of the household. These models remained, however, predominantly linear-regression models.

In a few instances an alternative method of modeling trip generation appeared. This method was known in the United States as cross-classification and in the United Kingdom as category analysis (1,4). This method went through the same development as the linear-regression models, with the earliest procedures being zonal trip estimators and subsequent models being based on household rates. For the most part, however, the household-based cross-classification models were still aggregate in that the classes were defined by average zonal values for household characteristics, and the trip rates were applied simply to the total number of households in the zone. Thus a cross-classification model based on household size and car ownership might have the first variable classified into ranges, such as less than 1.5 persons per household, 1.5 to 2.5 persons per household, 2.5 to 3.5 persons

per household, and more than 3.5 persons per household; car ownership was defined similarly in ranges. Then the average zonal values of each variable would be determined and a look-up table would be used to select one cell rate for the zone based on these average values.

Although the cross-classification method was widely used in Europe, it was used in relatively few instances in North America. However, with the growing interest in and use of disaggregate modal-choice models, there has been a resurgence of interest in the cross-classification model, formulated now in a substantially more disaggregate form. Currently, the model uses categorized variables, such as household size, vehicle ownership, and so on, as integer values to describe individual households. The rates in the cells of the table are then average rates for households of that type. The correct application of the model is to estimate the number of households in each category within a zone and to multiply the trip rates by those numbers of households. In general, this procedure leads to greater disaggregation than any other method of modeling trip generation, and has the potential to provide more policy responsiveness than alternative methods.

It is important to note that the standard method for computing cell rates is to group households in the calibration data to the individual cell groupings and total, cell by cell, the observed trips by purpose groups. The rate is then the total trips in a cell by purpose divided by the number of households in the cell. In mathematical form it is as follows:

$$t_{mn}^p = T_{mn}^p / H_{mn} \quad (1)$$

where

$t_{mn}^p$  = trip rate for the pth purpose for households of type mn,  
 $T_{mn}^p$  = observed trips made by households of type mn for purpose p, and  
 $H_{mn}$  = observed number of households of type mn.

The advantages that can be claimed for the disaggregate cross-classification methods are as follows:

1. Cross-classification methods are independent of the zone system of a region,
2. They do not require prior assumption about the shape of the relationships (which do not even need to be monotonic, let alone linear),
3. Relationships can differ in form from class to class of any one variable (e.g., the effect of household size changes for zero car-owning households can be different from that of one car-owning households), and
4. The cross-classification model does not permit extrapolation beyond its calibration classes, although the highest or lowest class of a variable may be open-ended.

The models also have several disadvantages, which are common to all traditional cross-classification methods:

1. There is no statistical goodness-of-fit measure for the model, so that closeness to the calibration data cannot be ascertained;
2. Cell values vary in reliability because of different numbers of households being available in each cell for calibration;
3. For the same reason as the preceding problem, the least-reliable cells are likely to be those at the extremes of the matrix, which may also be the most critical cells for forecasting;
4. There is no effective way to choose among variables for classification or to choose best groupings of a given variable, except to use an extensive trial-and-error procedure not usually considered feasible in practical studies; and
5. The procedure suppresses information on variances within a cell (5).

An alternative computational method is put forward and illustrated in the balance of this paper. This method--multiple classification analysis (MCA)--is well known to quantitative social scientists, but appears not to have been used by transportation analysts. As will be shown, MCA overcomes most of the disadvantages of cross-classification models without compromising their advantages.

#### MULTIPLE CLASSIFICATION ANALYSIS

MCA is based on a simple extension of analysis of variance (ANOVA), and ANOVA (6) also provides a statistically powerful procedure for selecting the variables and their categories for the cross-classification models. MCA is a rather simple development out of ANOVA, with application primarily for two-way and greater ANOVA problems.

Although a number of alternative methods have been suggested for analyzing cross-classification models and for determining cell values (7), there remains little change in the practice of estimating cross-classification cell values. Generalized linear models and regressions with dummy variables have been suggested as alternative methods, but they have not found wide acceptance in practice. The method suggested here is more readily accessible than most others because it is contained in some statistical packages that are available to transportation planners. Nevertheless, like many of the other methods that have been suggested recently, there is no treatment of this method in the statistical texts most frequently used by engineers and by courses taken by transportation planners. Indeed, no reference to the method could be found in any of the statistical texts most likely to be found on the bookshelf of a transportation planner or an engineer. Therefore, a brief description of the method is provided here.

Consider a two-way ANOVA design in which the dependent variable is a continuous variable, such as a trip rate, and the two independent variables are two integer variables that describe households, such as household size and vehicle ownership. First, a grand mean can be estimated for the dependent variable, where this grand mean is estimated over the entire sample of households. Second, group means can be estimated for each group of each independent variable, without regard for the other; in other words, means are computed from the row and column sums of the cross-classification matrix. Each of the group means can be expressed as a deviation from the grand mean. Observing the signs of the deviations, a cell value can now be estimated by adding the row and column deviations of the cell to the grand mean.

An example may help to clarify this. Suppose the dependent variable is home-based work trips, and the independent variables are cars owned and household size. The grand mean is 1.49 trips per household. Deviations for cars owned are -0.97 for zero cars, -0.26 for one car, and +0.88 for two or more cars. Deviations for household size are -1.06 for one person, -0.33 for two persons, +0.49 for three persons, +0.55 for four persons, and +0.70 for five or more persons. For a household with one car and three people, the trip rate would be estimated as 1.72 ( $= 1.49 - 0.26 + 0.49$ ). That is, it is the grand mean plus the deviation for one car plus the deviation for three persons. Note that, in contrast to standard transportation cross-classification models, the deviations are computed not only for households in the cell three persons with one car, but rather the car deviations are computed over all household sizes, and the household deviations are computed over all car ownerships.

If interactions are present, then these deviations need to be adjusted to account for the interactive effects. This is done by taking a weighted mean for each of the group means of one independent variable over the groupings of the other independent variables, rather than a simple mean, which assumes that variation is random over the data in a group. These weighted means will decrease the sizes of the adjustments to the grand mean when interactions are present. The cell means of a multiway classification are still based on means estimated from all the available data, rather than being based on only those data points that fall in the multiway cell. Furthermore, there is no over-compensation resulting from a false assumption of total lack of correlation between the independent variables.

Because it is based on ANOVA, MCA also has statistical goodness-of-fit measures associated with it. Primarily, these consist of an F statistic to assess the entire cross-classification scheme, an eta-square statistic (8) for assessing the contribution of each classification variable, and an R-square for the entire cross-classification model. These measures provide a means to compare among alternative cross-classification schemes and to assess the fit to the calibration data.

Without pursuing some further advantages offered by the statistical context within which MCA is applied, it is apparent that MCA overcomes effectively several of the disadvantages cited for other types of cross-classification models. First, there are statistical goodness-of-fit measures available for the MCA models that permit selection from among alternative classification schemes and that permit overall assessment of fit to the calibration data. Second, the cell values are no longer based only on the size of the data sample within a given cell; rather the cell values are based on a grand mean derived from the entire data set, and two or more class means are derived from all data in each class of the classification variables, where the intersection of those classes defines the cell of interest. This also tends to reduce the uncertainty of forecasting outlying households. For example, if a critical cell is the five or more person household with two or more cars available, for which the original data might have provided less than 2 percent of the sample, MCA will provide a cell rate that is based on the grand mean (from all the data) adjusted by deviations for all five or more person households and all two or more car households, where the first of these might comprise 10 percent or more of the data and the second more than 20 percent. Clearly, there is far greater reliability in this cell rate than would be obtained from traditional methods.

## SELECTING CLASSIFICATION VARIABLES AND CLASSES

In current computer software packages that compute an MCA (9), the MCA is usually provided after performing ANOVA. In turn the use of ANOVA provides the appropriate method for selecting variables and classes within variables. After developing a series of hypotheses about possible variables and classes of variables that might be used for the cross-classification scheme, a series of ANOVAs can be performed, from which several pieces of information are obtained that indicate better or worse classification schemes.

Several pieces of information are provided by a standard ANOVA that enable this evaluation to be made. First, there is an F statistic available for each main effect and for the interaction effects. A highly significant F statistic for the main effects indicates that the variable is strongly associated with the trip-rate variations in the data. A highly significant F statistic for the interaction effects suggests that the independent variables may be too highly intercorrelated to be useful, and it is likely to be necessary to choose among alternative independent variables and reduce as much as possible the interaction effects. There is also an overall F statistic for the entire cross-classification scheme that indicates the extent of covariation between the trip rates and the set of classified independent variables.

By trial-and-error procedures, or nested hypotheses, it is also possible to compare alternative independent variables and to compare alternative classifications. Of course, as the number of classes is changed, there is a consequent change in the number of degrees of freedom of the ANOVA problem and a consequent change in the expected F statistic. Obviously, this must be taken into account in assessing alternative schemes, but it then becomes possible to determine the amount of information loss occurring by aggregating classes, or the amount of added information obtained by disaggregating classes.

Thus ANOVA provides a structured and statistically sound procedure for selecting both the independent variables and the best groupings of those variables from those available. There is no claim of optimality in this, and clearly there are countervailing tendencies from aggregating and disaggregating variables, which demand the application of judgment to the results rather than blind acceptance of the statistical indicators. Also, the method is only as good as the initial and subsequent hypotheses of model structure. This may be interpreted as an advantage to the method over linear regression. The latter method permits too readily the abrogation of judgment to stepwise or similar regression procedures that may build models that appear to perform well, based on statistical measures and the R-square values, but which make no conceptual sense, whereas the application of ANOVA is far more demanding of the structuring of conceptually sound hypotheses, particularly because of its rather low efficiency in selecting good structures from blind application.

Finally, with each ANOVA it is possible to obtain the MCA results. These can also be revealing because they provide the additional statistics of an R-square and the eta-square for each variable, and they indicate the size of the deviations from the grand mean provided by each class of each independent variable. These data items may illuminate, clarify, or support the results from the ANOVA and should generally lead to a more rapid closure on a good structure for the model.

In summary, the use of the ANOVA that accompanies the MCA procedure resolves the remaining disadvantage of traditional cross-classification methods,

namely the lack of a sound method for choosing among alternative variables and alternative classes within a variable.

There is, however, one disadvantage incurred as a result of the use of MCA. MCA averages the effect of the relationships of one variable over classes of the other variables. Because the deviations are based on row and column means, there is no longer the capability for the shape of the relationship to differ from class to class of each variable as exists in traditional cross-classification methods. There does remain, however, no limitation on the average shape of the relationship for each independent variable, which still is not required even to be monotonic, let alone linear. This appears to be a relatively small price to pay for the advantages obtained, particularly when taking into account that many of the variations in functional form between classes in traditional models may derive from spurious small-sample effects.

## USE OF ANOVA TO SELECT VARIABLES AND CLASSES

A case study application of this method used data on 2,446 households from a metropolitan area in the Midwest. For initial variable selection, several candidates were identified and classifications were proposed for each of these variables. As a precursor to the multiway analyses, one-way ANOVAs were performed between trip rates and each candidate variable.

There are two bases for selecting variables in travel-forecasting models that hold true for any model. This first is conceptual or behavioral justification that the variable has a causal effect on the phenomenon being modeled, and the second is statistical justification that the variable shows a significant and measurable empirical association with the phenomenon being modeled.

Given 30 years of travel forecasting at the regional level, considerable experience and information exists now on variables that affect trip production, so that extensive concept formulation is not necessary. Based on past experience, the following variables were considered:

1. Household size (persons per household),
2. Automobile ownership or availability,
3. Housing type,
4. Household life cycle or structure,
5. Number of workers,
6. Number of licensed drivers,
7. Income, and
8. Area type.

Each of these variables is described briefly, together with its expected effects on trip production.

Household size is defined as the number of persons in the household without regard to age. Household size is expected to cause increases in tripmaking for all trip purposes, although not in a uniform manner. Trips per person is expected and has been shown to be relatively stable; hence the more people in the household, the more trips are likely to be made by the household.

Automobile ownership or availability is measured as the number of automobiles, vans, or lightweight trucks usable for personal travel by household members, either owned by the household or available to members of the household. A well-documented phenomenon is that acquisition of a vehicle increases substantially the number of trips and motorized trips made by a household. This arises both from substitution of vehicular trips for walk trips and from satisfaction of previously unsatisfied demand for travel. The tripmaking rate of increase is nonlin-

ear, with a decreasing rate of increase with increasing automobiles. Vehicle availability is likely to be the more appropriate measure than ownership because it is a more accurate measure of the potential to satisfy demand for vehicular trips.

Housing type is usually defined as single-family or multifamily dwellings, and hotel and motel units when tourists and nonresidents are to be included. It has a weak conceptual link, deriving principally from density considerations and some aspects of vehicle availability associated with vehicle storage space.

Recent research (10) suggests that a household-structure variable correlates more strongly with trip rates than almost any other variable. The categories of this variable are described elsewhere (see paper by McDonald and Stopher elsewhere in this Record), as are the arguments for its conceptual effect on tripmaking (10), and they are not described in this paper.

Number of workers may be defined as all workers, or as full-time workers only, where worker is restricted to work outside the home. Clearly, the number of workers will be in direct proportion to and is causative of the number of household work trips. Also, as more members of a household of a given size work, the number of trips for all other purposes is likely to be fewer, except for non-home-based trips, because more activities are likely to be undertaken on the way to or from work.

To the extent that a household has more licensed drivers than vehicles, more licensed drivers than workers, and more vehicles than workers, the number of licensed drivers would be expected to have a positive relationship to all nonwork trip purposes.

Income is usually defined as income groups of fairly broad income ranges. As income increases (all other things being equal), it is expected that tripmaking would increase because purchasing trips requires available monetary budgets and, as these increase, so does the potential to satisfy previously unsatisfied demand.

Area type has been defined in a variety of ways and is designed to differentiate between areas with markedly different intensities of development and activity. Therefore, either explicitly or implicitly, it is related to employment and residential densities. Where densities are higher, motorized trips are likely to be fewer because opportunities for satisfying activities are closer and both congestion and parking price may be significantly higher, whereas parking availability is lower. In addition, various services and home deliveries may be more available, thus reducing the need for some trips. The effect of area type is likely to be greatest on discretionary travel (home-based social-recreational, home-based other) and least on mandatory travel (home-based work or school).

The purpose of the one-way ANOVAs was both to determine which variables appeared to have the strongest relationships to tripmaking by purpose and to determine the best grouping of data to use. The results of these procedures were as follows.

1. Number of cars available was consistently one of the most significant variables for all trip purposes. It always performed better than number of cars owned.

2. Household size was also consistently a significant variable for all trip purposes.

3. Area type, which was defined as two groups--high density of either residences or employment, and low density of both residences and employment--was ranked third in significance across most trip purposes.

4. Housing type, denoted as single family and

multifamily, ranked about fourth in significance across most trip purposes.

5. Household structure, which was defined in terms of the relationships among household members, presence or absence of children, and some aspects of both household size and ages of members, was found to be inferior to household size alone and to number of cars available.

6. Other variables examined included number of workers, number of licensed drivers, and income. Each of these variables was significant for at least one purpose in the most disaggregated form of the variables, but they did not perform satisfactorily across a majority of the purposes.

In experiments on groupings, the results were as follows.

1. Vehicle ownership or availability could be specified as zero, one, and two or more without significant loss of power of the variable.

2. The optimal grouping of household size appeared to be one, two and three, four, and five or more. Examination of some other recent models (11) revealed a small difference in tripmaking rates for most purposes between two- and three-person households, which tended to confirm this grouping.

3. Income is best grouped into low (less than \$15,000), medium (\$15,000 to \$34,999), and high (more than \$35,000) categories.

4. Household structure should be grouped into five categories: single-person households, one-parent households, adult households with children and more than one adult, adult households without children and more than one adult, and households of unrelated individuals.

5. Number of workers can be grouped so as to aggregate households of four or more workers into one class, yielding categories of zero, one, two, three, and four or more.

6. Number of licensed drivers can also be aggregated to a set comprising zero, one, two, three, and four or more.

These results should not be considered indicative of general rules of classification. They are for the case study data and are provided here to illustrate the way in which ANOVA can be used for this type of analysis. Details of the runs are not provided here, because the results were derived from use of six trip purposes and involved running a rather large number of ANOVAs. Furthermore, it is not the purpose of this paper to produce specific recommendations on the structure of trip-generation models or to develop conclusions about the inclusion of one or another variable in the model. This is left to other papers that may use the approach described here to make more detailed studies of the performance of alternative variables. Despite the number required to be run, neither setup time to run them nor central processing unit (cpu) time on the computer to complete them were large.

The results of some of the multiway ANOVAs used to select the cross-classification scheme are given in Tables 1-4. The data in Table 1 give five purposes by using car ownership, housing type, and household size, whereas the data in Table 2 are the same except for the use of car availability in place of ownership. For all purposes except shopping, the F statistics are higher, although not significantly so, in most cases. The R-squares for the MCA tables and the eta-squares for the vehicle variable follow the same pattern. There are also two fewer significant interaction terms for car availability than for car ownership. This led to the selection of car

availability in preference to car ownership, thus confirming the results from the one-way ANOVAs.

The data in Table 3 give the replacement of the partly insignificant housing type by total employment. Only the home-based work model is clearly better in this specification, the models for all other purposes being virtually indistinguishable from the model with housing type. The data in Table 4 give the use of income in place of housing type.

Confirming the NCHRP results (10), income is apparently able to add little once vehicle availability is included. In all purposes, none of the statistical measures for the ANOVAs is as good for this specification as for the one that uses housing type.

An additional interesting result is given in Table 5. In the ANOVAs presented in Tables 1-4, household size was left disaggregated for two- and three-person households. In Table 5 the best speci-

Table 1. ANOVA results for model structure 1.

Statistic	Purpose				
	HBWORK	HBSHOP	HBSOCR	HBOTHR	NHB
F	28.0	6.0	5.7	33.8	10.5
df					
Within group	2,240	2,240	2,240	2,240	2,240
Between groups	29	29	29	29	29
Significant	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>
R <sup>2</sup>	0.255	0.065	0.059	0.291	0.103
Eta-square					
Vehicles owned	0.34 <sup>b</sup>	0.14 <sup>b</sup>	0.09 <sup>b</sup>	0.10 <sup>b</sup>	0.16 <sup>b</sup>
Housing type	0.06 <sup>b</sup>	0.05 <sup>b</sup>	0.01	0.02	0.05 <sup>b</sup>
Household size	0.25 <sup>b</sup>	0.16 <sup>b</sup>	0.20 <sup>b</sup>	0.50 <sup>b</sup>	0.22 <sup>b</sup>
Significant interactions	Vehicles owned and household size	None	None	Vehicles owned and household size; housing type and household size	Vehicles owned and household size

Note: Independent variables are vehicles owned, housing type, and household size. F = F-score, df = degrees of freedom, HBWORK = home-based work, HBSHOP = home-based shopping, HBSOCR = home-based social-recreation, HBOTHR = home-based other, and NHB = non-home-based trips.

<sup>a</sup>Significant at 99 percent or beyond.

<sup>b</sup>Significant at 95 percent or beyond.

Table 2. ANOVA results for car availability.

Statistic	Purpose				
	HBWORK	HBSHOP	HBSOCR	HBOTHR	NHB
F	29.5	5.9	6.0	35.1	11.4
df					
Within group	2,292	2,292	2,292	2,292	2,292
Between groups	29	29	29	29	29
Significant	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>
R <sup>2</sup>	0.261	0.062	0.060	0.295	0.113
Eta-square					
Vehicles available	0.36 <sup>b</sup>	0.12 <sup>b</sup>	0.10 <sup>b</sup>	0.11 <sup>b</sup>	0.20 <sup>b</sup>
Housing type	0.05 <sup>b</sup>	0.05 <sup>b</sup>	0.00	0.01	0.04
Household size	0.24 <sup>b</sup>	0.16 <sup>b</sup>	0.19 <sup>b</sup>	0.50 <sup>b</sup>	0.21 <sup>b</sup>
Significant interactions	None	None	Vehicles available and household size	Housing type and household size	None

Note: Independent variables are vehicles available, housing type, and household size. Statistics and purposes are defined in Table 1.

<sup>a</sup>Significant at 99 percent or beyond.

<sup>b</sup>Significant at 95 percent or beyond.

Table 3. ANOVA results with employment.

Statistic	Purpose				
	HBWORK	HBSHOP	HBSOCR	HBOTHR	NHB
F	37.0	4.3	5.2	25.9	9.5
df					
Within group	2,402	2,402	2,402	2,402	2,402
Between groups	42	42	42	42	42
Significant	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>	- <sup>a</sup>
R <sup>2</sup>	0.376	0.058	0.061	0.295	0.126
Eta-square					
Vehicles available	0.22 <sup>b</sup>	0.15 <sup>b</sup>	0.11 <sup>b</sup>	0.10 <sup>b</sup>	0.16 <sup>b</sup>
Workers	0.40 <sup>b</sup>	0.04	0.02	0.05	0.14 <sup>b</sup>
Household size	0.16 <sup>b</sup>	0.17 <sup>b</sup>	0.20 <sup>b</sup>	0.49 <sup>b</sup>	0.19 <sup>b</sup>
Significant interactions	Workers and vehicles available; workers and household size	None	Household size and workers; household size and vehicles available	Workers and household size	Workers and household size

Note: Independent variables are vehicles available, workers, and household size. Statistics and purposes are defined in Table 1.

<sup>a</sup>Significant at 99 percent or beyond.

<sup>b</sup>Significant at 95 percent or beyond.

Table 4. ANOVA results with income.

Statistic	Purpose				
	HBWORK	HBSHOP	HBSOCR	HBOTHR	NHB
F	23.7	4.1	3.2	22.8	10.5
df					
Within group	2,153	2,153	2,153	2,153	2,153
Between groups	41	41	41	41	41
Significant	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>
R <sup>2</sup>	0.298	0.053	0.046	0.284	0.119
Eta-square					
Vehicles available	0.21 <sup>b</sup>	0.13 <sup>b</sup>	0.08 <sup>b</sup>	0.08 <sup>b</sup>	0.13 <sup>b</sup>
Income	0.31 <sup>b</sup>	0.00	0.02	0.07 <sup>b</sup>	0.18 <sup>b</sup>
Household size	0.19 <sup>b</sup>	0.15 <sup>b</sup>	0.08 <sup>b</sup>	0.49 <sup>b</sup>	0.17 <sup>b</sup>
Significant interactions	None	None	None	Income and household size	Income and household size; vehicles available and household size

Note: Independent variables are vehicles available, income, and household size. Statistics and purposes are defined in Table 1.

<sup>a</sup>Significant at 99 percent or beyond.

<sup>b</sup>Significant at 95 percent or beyond.

Table 5. ANOVA results with aggregated household size.

Statistic	Purpose				
	HBWORK	HBSHOP	HBSOCR	HBOTHR	NHB
F	34.2	7.2	7.3	41.2	13.9
df					
Within group	2,298	2,298	2,298	2,298	2,298
Between groups	23	23	23	23	23
Significant	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>	<sup>a</sup>
R <sup>2</sup>	0.244	0.061	0.058	0.284	0.112
Eta-square					
Vehicles available	0.37 <sup>b</sup>	0.12 <sup>b</sup>	0.11 <sup>b</sup>	0.12 <sup>b</sup>	0.20 <sup>b</sup>
Housing type	0.05 <sup>b</sup>	0.05 <sup>b</sup>	0.00	0.01	0.04
Household size	0.19 <sup>b</sup>	0.15 <sup>b</sup>	0.19 <sup>b</sup>	0.49 <sup>b</sup>	0.21 <sup>b</sup>
Significant interactions	Vehicles available and household size	None	Vehicles available and household size	None	None

Notes: Independent variables are vehicles available, housing type, and household size. Statistics and purposes are defined in Table 1.

<sup>a</sup>Significant at 99 percent or beyond.

<sup>b</sup>Significant at 95 percent or beyond.

fication from the previous structures is used, but with the two- and three-person households aggregated into a single group. Because there is a decrease in the number of degrees of freedom, it is expected that the F score will increase. However, the increase is larger than would be expected just from this effect. Housing type still appears to be an ineffective variable, but the use of the more aggregated household size appears to be indicated quite clearly.

DERIVATION OF CROSS-CLASSIFICATION TRIP-GENERATION MODELS

A useful example of the MCA procedure is provided by the use of some data from a trip-generation modeling process used in San Juan, Puerto Rico (12). Figure 1 provides a set of trip rates computed in the standard procedure by using individual cell means. Note that cells 9 and 21 do not have trip rates because the available data lacked observations in these two cells. Figure 2 shows the numbers of households in each cell, and it can be seen that these range from a low of 4 to a high of 133. This range indicates clearly a significant range of reliability in the estimates of rates. If conventional wisdom is adopted, in that a mean and variance can be estimated with some element of reliability from a minimum of 50 observations, 14 of the 24 possible cells are estimated with too few data points.

As the next step in the procedure, a manual estimation of a noninteractive MCA was undertaken. This was done at the time because of the lack of availa-

bility of the computer software to undertake a full MCA, but it is useful because it traces out the procedure for MCA. First, a grand mean was computed for the entire set of home-based work trips; it was found to be 1.49. Then deviations were computed for each of the three variables. For the four household-size groups, the group means were found to be 0.33, 1.26, 1.85, and 1.84; for the two area types, they were 1.41 and 1.60; and for the three vehicle-ownership groups, they were 0.65, 1.51, and 2.36. The deviations are computed in each case by expressing the group means as values that deviate from the grand mean. To compute the cell value for area type 1, vehicle ownership of 1, and household size of four persons, the value is 1.98 (= 1.49 + 0.11 + 0.02 + 0.36). The complete set of cell values is shown in Figure 3. Note that there are values now in both cell 9 and cell 21.

Several points are worth noting from a comparison of Figures 1 and 3. First is the one already mentioned of the existence of rates for the empty cells of Figure 1 that appear in Figure 3. Second, some counterintuitive progressions in Figure 1 are removed or decreased substantially in Figure 3. These progressions appear to have been caused by problems from the small sample size. From examining the data in Figure 2, it can be seen that the grand mean is estimated from 1,178 observations, and that the least-reliable deviation (for one-person households) is based on 81 observations. All other deviations are based on more than 120 observations. Although there are still some large variations in the sample size used to compute the deviations, the range of 81

Figure 1. Conventional trip rates: home-based work.

CROSS class		Persons/ DU			
Area Type	Vehicles /DU	1	2,3	4	5+
	1 Rural Low Density	0	0.00	0.48	1.35
1		1.50	1.46	1.88	1.65
2+		-	2.10	2.23	2.36
2 Urban High Density	0	0.10	0.62	1.00	0.70
	1	0.80	1.29	1.58	1.69
	2+	-	2.19	2.70	2.59

Figure 3. Noninteractive MCA trip rates: home-based work.

CROSS class		Persons/ DU			
Area Type	Vehicles /DU	1	2,3	4	5+
	1 Rural Low Density	0	0.00	0.52	1.12
1		0.45	1.38	1.98	1.96
2+		1.30	2.23	2.83	2.81
2 Urban High Density	0	0.00	0.33	0.93	0.91
	1	0.26	1.19	1.79	1.77
	2+	1.11	2.04	2.64	2.62

Figure 2. Number of households by cell of cross-classification.

CROSS class		Persons/ DU			
Area Type	Vehicles /DU	1	2,3	4	5+
	1 Rural Low Density	0	17	60	17
1		4	88	48	69
2+		0	42	40	70
2 Urban High Density	0	40	110	34	40
	1	20	133	55	93
	2+	0	58	43	63

Figure 4. Full MCA trip rates: home-based work.

CROSS class		Persons/ DU			
Area Type	Vehicles /DU	1	2,3	4	5+
	1 Rural Low Density	0	0.12	0.62	1.01
1		0.86	1.36	1.75	1.70
2+		1.63	2.13	2.52	2.47
2 Urban High Density	0	0.10	0.60	0.99	0.94
	1	0.84	1.34	1.73	1.67
	2+	1.61	2.11	2.50	2.45

to 689 observations represents a much less-significant variation in reliability than in the data used for Figure 1.

Figure 4 presents the results from a full-interaction MCA for the same data. There are clearly some major interactions in this specification of the model, as shown by the differences in the rates between Figures 3 and 4. The anomalous decrease in rate between four and five or more person households remains and is of a similar order of magnitude, which suggests that this result is structured in the data. For the remaining differences, some rates are higher than before, whereas others are lower. As is

expected from the theory, the range of trip rates is lower in Figure 4 than in Figure 3 because accounting for interactions decreases the net effect of each variable. Thus the highest trip rate in Figure 3 is 2.83, whereas the highest rate in Figure 4 is 2.52. Similarly, the lowest value has increased from 0.00 in Figure 3 to 0.10 in Figure 4. Perhaps the most marked difference in the two figures is between the one and two or more vehicle households. The large differences at all household-size values between these two have decreased markedly in Figure 4, and the values of the one-vehicle households are substantially higher in the one-person households,

and lower in the largest households for Figure 4 compared with Figure 3.

Some statistical comparisons among the results serve to illustrate the differences better than can be seen from a visual inspection. First, root mean square (RMS) errors were calculated between Figures 1 and 3, Figures 1 and 4, and Figures 3 and 4. For Figures 1 and 3, it is 0.47; between Figures 1 and 4 it increases to 0.51; but it is only 0.24 between Figures 3 and 4. This is about as expected. The largest difference is between the conventional rates and the MCA rates with full interactions. The difference between MCA with full interactions and without is by far the least of the differences. Given an average trip rate of around 1.45, the differences between the conventional method and the MCA methods are on the order of one-third of the average trip rate.

Chi-square contingency tests between values close to 1.0 are notoriously misleading because the value of chi-square is necessarily small in such a case. This case is no exception, with the three comparisons producing chi-squares of 1.88, 4.22, and 1.30, each with 21 degrees of freedom. These values would not be considered significant. However, if the rates are multiplied by the number of households in the sample (Figure 2), the chi-square test would be for differences in the numbers of trips produced for work. In this case the chi-squares are 55.5, 19.0, and 41.4, respectively. The degrees of freedom are the same as before, and all values except the second one are significant beyond 95 percent. The low chi-square between Figures 1 and 4 appears to arise purely by chance, where two of the larger groups of households are associated with a small difference in trip rates, fortuitously. It is not clear whether this result should lead to a conclusion of no significant difference in trip rates between the two cases. Thus these results indicate some real differences in trip rates that are likely to lead to significant differences in forecasts.

#### CONCLUSIONS

The two case studies presented in this paper serve to illustrate the potentials provided by the MCA method and ANOVA from which it stems. This procedure overcomes a number of the criticisms that have been made before about cross-classification models. Specifically, the method permits a statistically based selection of variables for the cross-classification model, and also allows comparisons to be made between alternative groupings of any given variable. From this it is possible to provide a model structure that has both conceptual and statistical merit, rather than relying only on a conceptual selection.

Second, the method provides a statistically sound procedure for estimating cell means, which reduces the inherent variability of rates computed from different size samples of households and is capable of providing estimates for some cells where data may be lacking in the base data set (although the use of this capability does reduce some of the available statistical information). Third, there are good-

ness-of-fit statistics from all of these steps in the process that permit more specific comparisons to be made, good hypothesis-testing procedures to be followed, and results to be assessed in terms of the amount of the variability of the dependent variable that is captured in the model. Finally, and most important, the method takes into account the interactions among the alternative independent variables, which have never been taken into account in standard cross-classification models.

It should be noted that similar models have been developed for predicting vehicle availability, as well as for trip productions by a variety of purposes. There is no reason why such cross-classification models should not be built for any other phenomenon that is appropriately modeled by this procedure. Principally, any phenomenon that has a nonlinear, and possibly discontinuous, functional form, and that is most readily related to variables that are categorical in nature, would be a prime candidate for the method.

#### REFERENCES

1. Guidelines for Trip Generation Analysis. FHWA, U.S. Department of Transportation, 1967.
2. G.M. McCarthy. Multiple-Regression Analysis of Household Trip Generation--A Critique. HRB, Highway Research Record 297, 1969, pp. 31-43.
3. H. Kassoff and H.D. Deutschman. Trip Generation: A Critical Appraisal. HRB, Highway Research Record 297, 1969, pp. 15-30.
4. R. Lane, T.J. Powell, and P. Prestwood-Smith. Analytical Transport Planning. Halsted Press, Wiley, New York, 1973.
5. P.R. Stopher and A.H. Meyburg. Urban Transportation Modeling and Planning. D.C. Heath and Co., Lexington Books, Lexington, Mass., 1975.
6. N.L. Johnson and F.C. Leone. Statistics and Experimental Design: Volume II. Wiley, New York, 1964.
7. D. Segal. Discrete Multivariate Model of Work-Trip Mode Choice. TRB, Transportation Research Record 728, 1979, pp. 30-35.
8. P.R. Stopher. Goodness-of-Fit Measures for Probabilistic Travel Demand Models. Transportation, Vol. 4, 1975, pp. 67-83.
9. N.H. Nie, C.H. Hull, J.G. Jenkins, K. Steinbrenner, and H.D. Bent. Statistical Package for the Social Sciences, 2nd ed. McGraw-Hill, New York, 1975.
10. P.M. Allaman, T.J. Tardiff, and F.C. Dunbar. New Approaches to Understanding Travel Behavior. NCHRP, Rept. 250, Sept. 1982, 147 pp.
11. Schimpeler-Corradino Associates. Review and Refinement of Standard Trip Generation Model. Florida Department of Transportation, Tallahassee, Final Report Task B, June 1980.
12. Schimpeler-Corradino Associates. Urban Transportation Planning Models. Department of Transportation and Public Works, Commonwealth of Puerto Rico, San Juan, Final Rept., July 1982.

*Publication of this paper sponsored by Committee on Traveler Behavior and Values.*