Comparison of Alternative Methods for Estimating Household Trip Rates of Cross-Classification Cells With Inadequate Data

by Judith L. Mwakalonge and Daniel A. Badoe

This paper investigates the forecast performance of a traditional cross-classification model and alternative models that seek to address the shortcomings of traditional cross-classification analysis, specifically when it has cells with inadequate data. The study uses five cross-sectional datasets collected in the San Francisco Bay Area in 1965, 1981, 1990, 1996, and 2000. Alternative models, estimated with travel data collected in the base year, were assessed for their ability to replicate the number of trips made by households in each cell of a cross-classification matrix and at the traffic zone level, respectively, in each of the five years. The results showed that the traditional cross-classification analysis (CCA) model, notwithstanding having a few unreliable cells provided more consistent predictions of travel than any of the alternative methods. They also show that it is better to synthesize trip rates for only those cells of the cross-classification matrix with inadequate data rather than to adjust the entire trip-rate matrix as is currently the practice.

INTRODUCTION

The four-step Urban Transportation Modeling System (UTMS) continues to be the method adopted by the majority of metropolitan planning organizations for simulating traffic volumes using the links of urban transportation networks (TRB 2007). This paper focuses on trip generation, the first step of the four-step UTMS. Given the sequential nature of UTMS, improved forecast accuracy at the trip generation stage is important to reducing errors in the forecasts emanating from the final step of the process.

A number of methods for accomplishing trip generation are documented in the travel demand modeling literature. These include multiple linear regression (Cotrus et al. 2003, Ewing et al. 1996), cross-classification analysis (Walker and Peng 1991, Rengaraju and Satyakumar 1995), discrete choice models (Zhao 2000), fuzzy logic models, and artificial neural networks (Huisken 2000). However, of these methods, cross-classification analysis (CCA) is the most widely used in practice (Rengaraju and Satyakumar 1994).

Cross-classification analysis involves the use of trip rates (i.e., trips per person or trips per household) to compute regional travel demand. Recognizing the heterogeneity in regional populations, the approach first divides the population into relatively homogeneous groups or categories based on two or three household attributes. Thereafter, a trip rate is calculated for each relatively homogeneous group. The technique is non-parametric in that it does not assume any probabilistic distributional relationship between the dependent and explanatory variables. Furthermore, the method makes use of the raw data obtained from a household travel behavior survey directly, and its simplicity has made it attractive to practitioners (Rengaraju and Satyakumar 1994). The method, however, has its shortcomings.

First, given the typical size of travel survey samples that most planning agencies have available for travel demand model development, cross classifying the sample into a large number of relatively homogeneous categories leaves some cells with few or no observations for the computation of trip rates. These problematic cells typically exist at the extreme ends of the cross-classification matrix. As an example, the proportion of households in an urban area with a single person and owning three

or more vehicles is likely to be very small. A simply drawn random sample of households from the regional population may include few or no households with such characteristics. Therefore, cross classifying the travel data could result in such a cell being empty, making it impossible to estimate directly a trip rate for it.

Second, the estimated trip rates of the cross-classification matrix suffer from differential reliability resulting from the differences in the numbers of households in each cell for trip-rate computation. Trip rate is the expected number of trips a household makes per day. This difference in reliability could result in counterintuitive trip-rate progressions in the trip-rate matrix. These two shortcomings among others documented in the literature have spurred researchers to investigate new techniques for improving upon the basic model. Examples of these studies include those by Rengaraju and Satyakumar (1994), Kikuchi and Rhee (2003), and Stopher and McDonald (1983). The most known of these methods, proposed by Stopher and McDonald (1983), makes use of multiple classification analysis (MCA). However, its implementation also raises concerns. First, it modifies all the trip rates obtained using the CCA procedure, notwithstanding several cells in the matrix having adequate data for computation of reliable trip rates. Second, sometimes implementation of the MCA procedure results in the computed trip rate for some of the cells of the classification matrix having a negative sign, which is not meaningful. The analyst addresses the resultant negative trip-rates problem by assigning a zero trip rate to such a cell (Ortuzar and Willumsen 2001). Assigning zeros to cells that either had values earlier or were empty in CCA is unrealistic.

Kikuchi and Rhee (2003) applied a fuzzy optimization method to synthesize missing cell values and adjust cell values with abnormal behavior when compared to neighboring cells. However, the fuzzy optimization method, like the MCA, changes the cell values of the entire classification matrix instead of the cells with inadequate data. Additionally, the fuzzy optimization technique requires knowledge of a programming language and is therefore not readily accessible to transportation planners, which limits its use by practitioners.

Thus, while these attempts to remedy the weaknesses of CCA are recognized, the problem of adjusting the trip rates that are derived from the observed sample persists. Additionally, it appears that no study has investigated both the short-term and long-term forecast performance of the methods proposed to remedy the shortcomings of CCA. Guevara and Thomas (2007) recommended not using the MCA method proposed by Stopher and McDonald (1983). However, their recommendation was based in part on analysis done using a single origin-destination survey data. Further, they conducted their model evaluation using forecasted land use scenarios and not observed land use and travel characteristics. The above discussion motivates an investigation into alternative methods or modifying existing methods for synthesizing trip rates for cross-classification cells with no data that do not require the modification of trip-rate values for cells with adequate data.

Specific objectives of the paper are, first, to develop trip generation models using CCA and MCA, respectively, and to compare how the models perform in the prediction of travel in the base year. The second objective is to compare the performance of both CCA and MCA models in short-term and long-term forecast applications. The third is to present alternative methods for addressing the shortcomings of CCA and to compare the forecast performance of these alternative methods against the models developed using CCA and MCA, respectively.

The rest of the paper is organized as follows. In the second section, the theory underlying the existing and proposed methods for estimating a trip rate for a cross-classification cell with no data are presented. The third section presents the descriptive analysis of the travel data used in the research. The fourth section presents the model estimation results and results from applying the alternative methods in predicting travel. Finally, the last section presents a summary and conclusions drawn from the study.

ALTERNATIVE MODELS FOR SYNTHESIZING TRIP RATES FOR CROSS-CLASSIFICATION CELLS WITH NO DATA

This section presents a brief description of the theory underlying the alternative models investigated in this study. The existing models considered in this study include CCA and MCA models. The current practice is to employ the MCA technique that modifies the whole cross-classification trip-rate matrix. However, MCA can also be used to estimate trip rates for empty cells and unreliable cells. Therefore, this study makes use of MCA models and techniques employed in estimating missing values to compute trip rates for empty and less reliable cells. The techniques for estimating missing values investigated in this research are Multiple Imputation (MI) and K-Nearest Neighbor (KNN). The theory of each of these methods is discussed in turn below.

Cross-Classification Analysis

As discussed in the introduction, CCA involves the computation of trip rates typically at the household level. However, recognizing the heterogeneity in travel behavior that exists among households in an urban region, households are grouped according to two or more characteristics that are strongly associated with trip-making behavior. Households belonging to each defined group are therefore assumed relatively similar in trip-making behavior. The model's basic assumption is that household trip rates remain stable over time for defined household stratifications. It should be noted that the model could be developed for each trip purpose. However, in this research, we consider trips made across all trip purposes by a household and two-household attributes for defining groups of similar travel behavior. The household trip rate for each defined group is calculated as:

$$(1) \qquad \bar{y}_{mn} = \frac{\sum_{h=1}^{H_{mn}} y_{mn}^{h}}{H_{mn}}$$

Where

m, n = values of two-household attributes used in defining homogeneous groups (cells)

$\bar{y}_{mn}$ = trip rate for cell of cross-classification matrix with household attribute values *mn*

y_{mn}^{h} = trips made by household *h* in cell *mn*

H_{mn} = total number of households in cell *mn*

Multiple Classification Analysis

MCA is similar to multiple regression analysis with dummy variables. The approach is applicable where the dependent variable is quantitative and the explanatory variables are categorical, represented by dummy variables. Therefore, MCA with one categorical variable is equivalent to one-way Analysis of Variance (ANOVA), similarly MCA with two categorical variables correspond to two-way ANOVA (Retherford and Choe 1993). Stopher and McDonald (1983), as a remedy to the shortcomings of CCA, were the first to apply the technique in trip generation analysis. Thereafter, several researchers (Ortuzar and Willumsen 2001, Wardman and Preston 2001, Abdel-Aal 2004) applied the method. However, none of the mentioned studies used MCA to estimate trip rates for empty and/or unreliable cells only. Rather, they employed it to modify the whole trip-rate matrix. The general mathematical form of the MCA model is expressed as:

$$(2) \qquad \bar{y}_{mn} = G_{\mu} + \alpha_m + \beta_n + \varepsilon_{mn}$$

Where

$\bar{y}_{mn}$ = the trip rate for a cell in a cross-classification matrix with household attribute values *mn*
G_{μ} = the grand mean of trips made by the households in the dataset
α_{m} = the column-effect for column *m* of a cross-classification matrix
β_{n} = the row-effect for row *n* of a cross-classification matrix
ε_{mn} = error term

For comparison purposes, this study reviews and investigates three MCA models designated as MCA1, MCA2, and MCA3. The first, MCA1, takes the following form (Guevara and Thomas 2007).

(3) $\bar{y}_{mn} = \mathrm{G}_{\mu} + \alpha_{\mathrm{m}} + \beta_{\mathrm{n}} \quad \{ \begin{matrix} \forall m \in M \\ \forall n \in N \end{matrix}$

Where

(4) $G_{\mu} = \frac{\sum_{h=1}^{H} y^{h}}{H}$

(5) $\alpha_{m} = \frac{\sum_{n \in N} y_{mn}^{h}}{\sum_{n \in N} H_{mn}} - G_{\mu}$

(6) $\beta_{n} = \frac{\sum_{m \in M} y_{mn}^{h}}{\sum_{m \in M} H_{mn}} - G_{\mu}$

N, M = the respective number of classes for the two stratification variables
n, m = the values of two household attributes used in defining homogeneous groups (cells)
H = the total number of households
y^{h} = the trips made by household *h*
G_{μ} = the grand mean of trips made by the households in the dataset
α_{m} = column effect for column *m* of a cross-classification matrix
β_{n} = row effect for row *n* of a cross-classification matrix
ε_{mn} = error term

The second MCA model, MCA2, takes the same mathematical form as the first one except the row and column effects are calculated as weighted means, which therefore takes into consideration the unequal number of observations in the cells of the cross-classification matrix (Stopher and McDonald 1983, Guevara and Thomas 2007).

(7) $\alpha_{m} = \left(\sum_{n \in N} w_{mn} \bar{y}_{mn} / \sum_{n \in N} w_{mn} \right) - G_{\mu}$

(8) $\beta_{n} = \left(\sum_{m \in M} w_{mn} \bar{y}_{mn} / \sum_{m \in M} w_{mn} \right) - G_{\mu}$

Where

w_{mn} = weighting factor for cell *mn*
$\bar{y}_{mn}$ = trip rate for a cell in a cross-classification matrix with household attribute values *mn*
G_{μ} = overall mean that is average number of trips per household
β_n = row effect for row *n* of a cross-classification matrix
α_m = column effect for column *m* of a cross-classification matrix

The third, MCA3, is from an MCA regression of household trips on all classification variables. However, the model is slightly different from ordinary least squares in that when calculating the marginal effect of an explanatory variable, the other explanatory variables are held constant at their mean values in the entire sample (Retherford and Choe 1993). The model's mathematical form is

$$(9)\quad \bar{y}_{mn} = a + \sum_{n \in N} \beta_n X_n + \sum_{m \in M} \alpha_m X_m$$

Then the trip rates for the categories of variable X_n are calculated as:

$$(10)\quad \bar{y}_{mn} = a + \sum_{n \in N} \beta_n X_n + \sum_{m \in M} \alpha_m \bar{X}_m$$

Where

X_n, X_m = 1 if the *n*th or m^{th} element of X is observed, and equals a zero otherwise.
$\bar{y}_{mn}$ = trip rate for a cell in a cross-classification matrix with household attribute values *mn*
β_n = row effect for row *n* of a cross-classification matrix
α_m = column effect for column *m* of a cross-classification matrix

n and *m* are initial classes that are considered as reference classes, hence a constant *a* to be estimated is added.

Multiple Imputations (MI)

MI is a three-step approach that employs regression analysis to impute missing values (Rubin 1976). The first step is to estimate a model using observations with complete data and, thereafter, use the estimated model to fill in the missing values. The second step is to estimate a model using a complete data set with both observed and imputed values. For this case, the analyst substitutes predicted values for the missing values to create imputed datasets. The procedure is repeated until the analyst has the desired number of imputed datasets. Usually, three to ten imputed datasets are desirable (Wayman 2003). Finally, the estimates from steps one and two are combined to account for the uncertainty regarding the imputation. In mathematical form, the joint distribution is a function of the marginal and conditional distribution and it is represented as (Horton and Kleinman 2007):

$$(11)\quad f(Y_h, X_h) = f\left(Y_h^{miss}, Y_h^{obs} \mid X_h, \boldsymbol{\beta}\right) P(X_h)$$

Where

Y^{obs} = observed dependent variable (trip rates)
Y^{miss} = missing dependent variable
X_h = vector of explanatory variables (two household attributes used in defining homogeneous groups (cells))
β = vector of parameters
$f\left(Y_h^{miss}, Y_h^{obs} \mid X_h, \boldsymbol{\beta}\right)$ = Conditional probability distribution
$P(X_h)$ = Marginal probability distribution

The final imputed estimate is the combined estimate that follows Rubin's procedure (Rubin 1976), which is a simple average of individual estimates from the observed and imputed datasets. Mathematically this is,

$$(12) \quad \bar{y}^{h} = (1/K)\sum_{i=1}^{K} \hat{y}_i$$

$$(13) \quad \bar{y}_{mn} = \sum_{h=1}^{H_{mn}} \bar{y}_{mn}^{h} / H_{mn}$$

Where
K= number of imputed full datasets
All other variables are as defined earlier.

K-Nearest Neighbor (KNN)

KNN is a technique for estimating unobserved data based on the characteristics and values of the observed nearest data. KNN technique has been widely applied in medical research and geosciences (Muhammad et al. 2004) but less so in transportation. The simplicity of the KNN method motivated its application in estimating empty cells in the trip-rate matrix. Selection of nearest cells is determined based on similarity in characteristics between the filled nearest cells and the empty cell. For example, a missing trip rate for a single-person household with four or more vehicles may have similar characteristics to a single-person household with three vehicles, since both households have surplus vehicle supply. Therefore, a missing cell value is computed by weighting the predetermined nearest cell values as follows,

$$(14) \quad \bar{y}_{mn} = \sum_{\substack{n\in N \\ m\in M}} w_{mn}\, \hat{y}_{mn} \Big/ \sum_{\substack{n\in N \\ m\in M}} w_{mn}$$

$$(15) \quad w_{mn} = o_{mn} / \sigma_{mn}^{2}$$

Where
σ_{mn}^{2} = variance estimate for the mn^{th} nearest cell
o_{mn} = number of observations in the mn^{th} nearest cell

All other variables are as defined earlier.

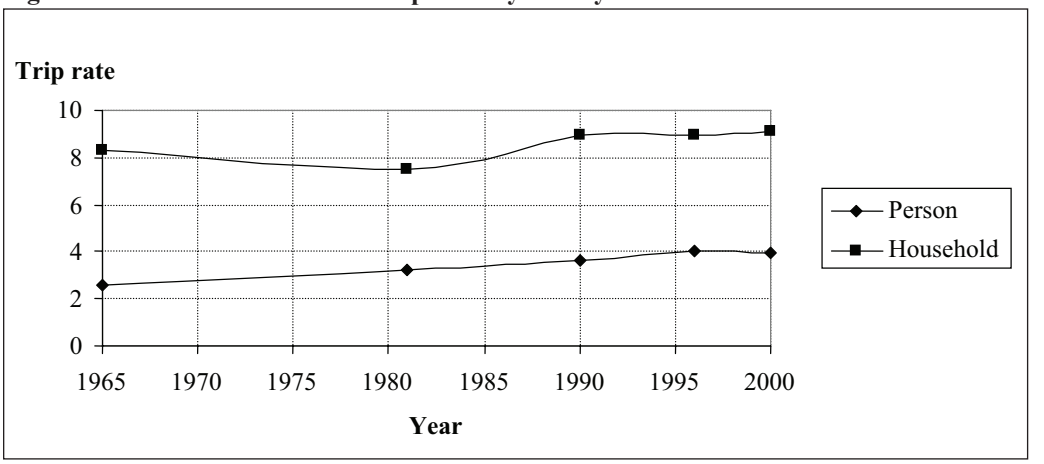
DATA

The research uses five cross-sectional datasets collected in different years (1965, 1981, 1990, 1996, and 2000) in the San Francisco Bay area. The 1965 dataset has information on more than 20,000 households, while the 1981 dataset has information on more than 7,000 households. The 1990 dataset has information on more than 9,000 households, while the 1996 dataset is the smallest sample with information on a little more than 3,600 households. Finally, the 2000 dataset has information on more than 15,000 households. The analysis presented below uses the sample data and unlinked trips. Information on linked trips and the trip-linking procedure are in MTC (2003). The five datasets are comparable since the region has remained relatively stable in terms of geographic area. However, the survey instrument changed from home interview to telephone interview (1981 onward), and trip recall to activity diary (1996 and 2000 are activity-based surveys). In the context of how the alternative modeling methods are to be assessed in this study, the differences in instruments are unlikely to pose any problems.

Trip Rate Distribution

With the exception of 1981, Figure 1 shows that the household trip rate in the Bay Area remained relatively stable. There is a noticeable decrease in household trip rates in 1965 compared with 1981. Purvis (1994) reported that other major cities, namely Dallas and Denver, exhibited the same pattern in trip-making behavior and noted this decrease in trip rate. However, at the individual level, there is a progressive increase in trip rate from 1965 to 1996, and thereafter it remained stable.

Figure 1: Household and Person Trip Rate by Survey Year



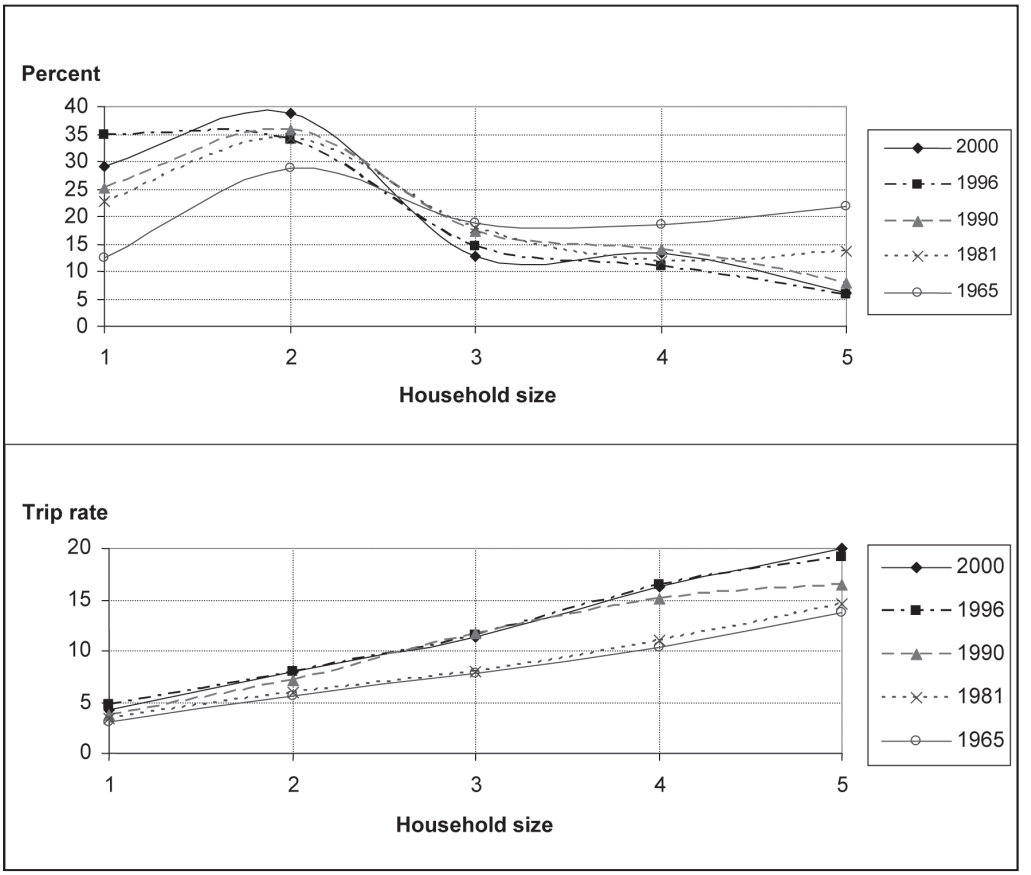
Household Size

Household size affects travel demand; on average, the larger a household, the greater its activity needs and the number of trips made. Figure 2(a) shows household size distribution across the analysis years. Generally, there is an increase in single person households and a decrease in four or more person households from 1965 to 2000. Although the trip rate increases with household size, it increases at different rates over the analysis years across different household groups. For example, there is a dramatic increase in travel demand from 1981 to 1990 for households with three to four persons. This increase is partly explained by a more than 10% increase in the working age group (age 36 to 55), a small trip-rate increase of 0.43 trips for single-person households and an increase of 1.30 trips for two-person households. All else being equal, travel behavior was stable from 1965 to 1981 and from 1996 to 2000 as shown in Figure 2 (b).

Vehicle Ownership

People purchase vehicles with the aim of increasing their mobility and activity participation. On average, the greater the number of vehicles owned by a household, the greater the number of trips they are likely to make by vehicle. As observed, the percentage of households with no vehicle was higher in 1981 than in 1965. With the exception of the 1990 household trip rates, there is a consistent, although minor increase in trip rate for zero-vehicle households from 1965 to 2000, and a stable trip rate for households with one or more vehicles. Households with three or more vehicles had a much higher trip rate in 1990 than in any of the other years. Figure 3 is a graphical summary of these details.

Figure 2: (a) Household Size Distribution by Survey Year
(b) Trip Rate by Household Size for Each Survey Year



Variable Selection

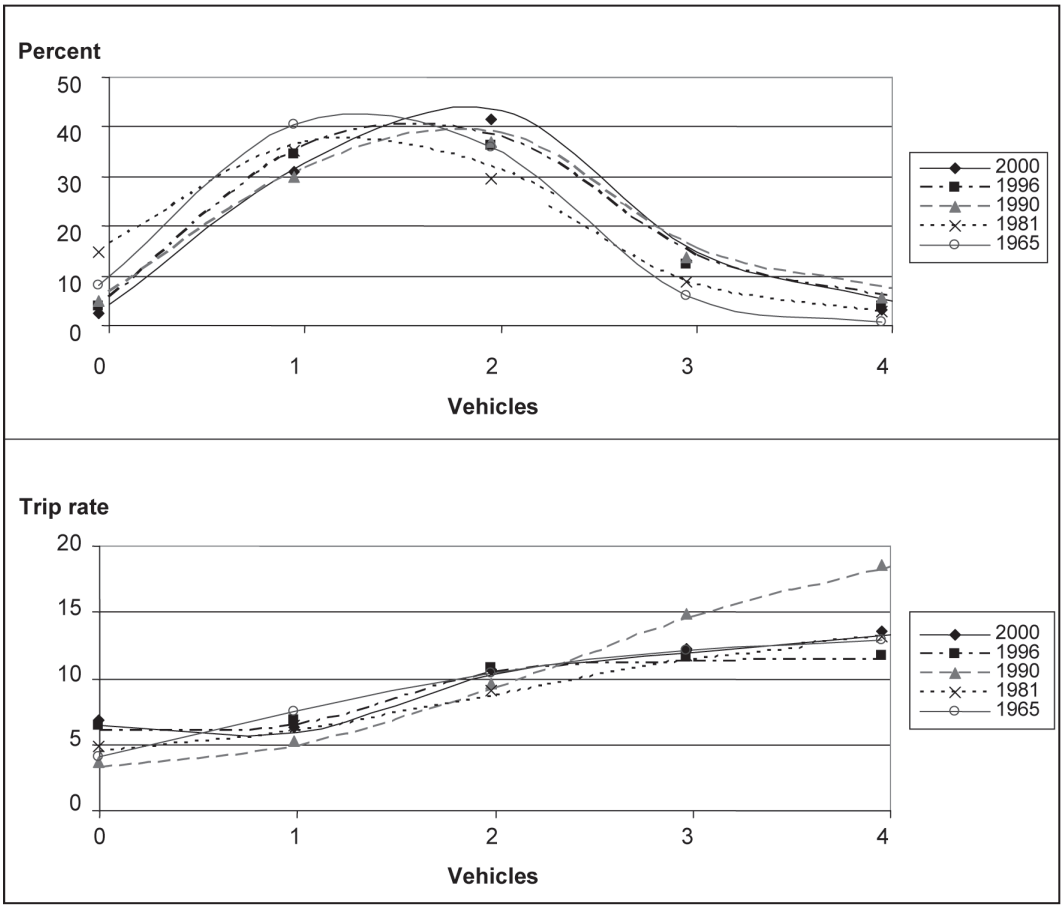
The 1965 dataset has eight potential explanatory variables. The objective was to select two or three that could capture most of the variation in household trips. In accomplishing this objective, the study uses analysis of variance procedure (ANOVA), and the results are in Appendix A. At the 5% level of significance, the results show that house tenure (own or rent) and dwelling type, with respective probabilities of 0.2942 and 0.2556, were not statistically significant. Variables that were statistically significant are household size, number of household members with drivers licenses and household income, the number of motorcycles owned by a household and vehicles owned by a household. Appendix A shows that the number of households with drivers licenses correlates moderately with the number of vehicles owned by a household. Consequently, the analysis uses the number of vehicles owned by a household and household size as the stratification variables.

EMPIRICAL TEST

Test Procedure

The assessment of the performance of the alternative methods for developing a cross-classification model for trip generation involved five steps. In the first step, the study estimates the CCA model and the three MCA models with the 1965 data using the household as the modeling unit. The

Figure 3: (a) Vehicle Ownership Distribution by Survey Year
(b) Trip Rate by Vehicle Ownership for Each Survey Year



second step uses each of the four models to predict travel collectively made by households in each classification cell and by households in each traffic analysis zone in 1965, respectively. The latter assessment of model performance at the traffic zone level is important because trip distribution, a step in the four-step UTMS, requires as input trip productions and trip attractions at the traffic zone level. In the third step, the study uses the four models in step one to predict household travel in each cross-classification cell in 1981, 1990, 1996, and 2000, respectively. In the fourth step, each of the methods proposed for synthesizing trip rates for cross-classification cells with little or no data was applied to predict the trip rate for only those cells of the traditional cross-classification matrix considered unreliable. The household trip rates from the traditional CCA were preserved for the cells with enough observations. Finally, the fifth step uses the cross-classification matrices from the fourth step to predict household travel in all the years for which data were available.

RESULTS AND DISCUSSION

Estimated Models

The results of the model estimation in the first step are in Table 1. They show that each cell has a different number of observations. For example, for single-person households, the sample size for

those with no vehicle is 1,062, whereas for those with four or more vehicles it is 11. Given that the reliability of the trip rate for each cell is a function of the number of observations in the cell, it is apparent that there are differences in cell reliability in the CCA model due to each cell having a different number of observations.

In the descriptive analysis presented earlier, there was a monotonically increasing relationship between trip rate and household size (Figure 2b). A similar relationship was observed between trip rate and vehicle ownership (Figure 3b). However, this increasing trend is not consistently observed when one examines how trip rates that are conditional on a specific household size vary with vehicle ownership or how trip rates that are conditional on a specific vehicle ownership level vary with household size in the CCA matrix. As an example, Figure 2b shows that for a household size of one (single-person household), trip-rate increases with increasing vehicle ownership until a household vehicle ownership level of three when it drops, and then increases thereafter for those single person households that have four or more vehicles. A similar observation in Figure 3b regards the relationship between trip rate and household vehicle ownership for two-person households.

As expected, Table 1 shows a counterintuitive progression in trip rates in the less reliable cells (e.g., single-person household with three vehicles). To address this problem, the practice is to employ MCA; and the results of doing so for the different methods are in Table 1. MCA2 yielded a trip rate for single-person households with no vehicle that has a negative sign. Since a negative trip rate is unrealistic, the practice is to set it to zero (Ortuzar and Willumsen 2001). However, for illustration, Table 1 preserves this negative-valued trip rate although in the forecasting analysis done later it is set to zero. MCA1 and MCA3 yielded household trip rates with trends that are consistent with expectation; that is, higher household trip rates for higher values of vehicle ownership and household size, respectively.

Prediction of Travel at the Household and Traffic Zone Level In 1965

Travel demand models need to provide accurate predictions to guide decisionmakers in infrastructure investment decisions. Therefore, the four models were applied in turn to predict travel in the base year at the disaggregate household level, and their performance was judged based on the coefficient of determination (R^2) and the percent mean absolute error (PMAE) calculated as,

$$(16) \quad PMAE = \left(\sum_{\substack{m \in M \\ n \in N}} \left(\frac{y_{mn}^{pred} - y_{mn}^{obs}}{y_{mn}^{obs}} \right) * 100 \right) / (N * M)$$

Where

y_{mn}^{pred} = predicted number of trips made by households in cell *mn*

y_{mn}^{obs} = observed number of trips made by households in cell *mn*

N, M = the respective number of classes for the two stratification variables

Table 2 shows an assessment of the accuracy of each of the four models in predicting the trips made by each household in 1965. Of the three MCA models, MCA1 has the smallest PMAE while MCA2 has the largest error value. Contributing to the high PMAE value of MCA2 was its complete failure to predict any trips made by single person households with no vehicles. (Column three of row four in Table 1 has a negative trip rate that is set to zero in forecasting). Also shown in columns two and three are the results of regressing the observed number of daily trips made by the households in each cross-classification cell against the number of daily trips to be made by the households in each cross-classification cell predicted by each model (CCA, MCA1, MCA2, or MCA3). They indicate that for CCA, MCA1, and MCA3, the estimated slope coefficients are almost one while the slope coefficient of MCA2 is about 0.92. Based on the coefficients of determination, MCA1, MCA2, and MCA3 explain the variation in household trips in the 1965 data very well; MCA1 and MCA3

Table 1: Estimated CCA and MCA Models Using 1965 Data

Household Size	Model	Number of Vehicles					
		0	1	2	3	4+	Total
	CCA	2.201	3.644	4.756	3.941	5.000	3.084
	MCA1	1.091	3.108	4.567	5.290	5.486	3.908
	MCA2	-1.675	1.827	4.757	6.586	7.314	3.084
	MCA3	2.028	3.710	5.182	6.404	6.571	4.376
1	No. of obs. in cell	1062	1367	82	17	11	2539
	CCA	3.658	5.417	6.216	7.161	6.889	5.603
	MCA1	3.051	5.068	6.527	7.250	7.445	5.868
	MCA2	0.844	4.346	7.276	9.105	9.833	5.603
	MCA3	3.446	5.128	6.601	7.823	7.989	5.794
2	No. of obs. in cell	556	3016	2081	193	54	5900
	CCA	5.406	7.240	8.159	8.940	8.978	7.773
	MCA1	4.927	6.944	8.403	9.126	9.322	7.744
	MCA2	3.014	6.516	9.446	11.275	12.003	7.773
	MCA3	5.181	6.863	8.335	9.558	9.724	7.529
3	No. of obs. in cell	175	1543	1575	448	89	3830
	CCA	6.774	9.111	10.999	11.797	12.046	10.362
	MCA1	7.328	9.345	10.804	11.527	11.723	10.145
	MCA2	5.603	9.105	12.034	13.864	14.592	10.362
	MCA3	7.613	9.295	10.768	11.99	12.156	9.961
4	No. of obs. in cell	106	1283	1818	424	130	3761
	CCA	9.173	11.883	14.46	16.367	16.270	13.769
	MCA1	10.813	12.830	14.289	15.012	15.208	13.630
	MCA2	9.010	12.512	15.441	17.271	17.999	13.769
	MCA3	10.981	12.663	14.135	15.358	15.524	13.329
5+	No. of obs. in cell	139	1444	2137	523	211	4454
	CCA	3.587	7.089	10.018	11.848	12.576	8.346
	MCA1	5.442	7.459	8.918	9.641	9.837	8.259
	MCA2	3.587	7.089	10.018	11.848	12.576	8.346
	MCA3	5.998	7.680	9.153	10.375	10.541	8.346
Total	No. of obs. in cell	2038	8653	7693	1605	495	20484

Note: CCA-Cross Classification Analysis; MCA1-Multiple Classification Analysis Model 1; MCA2-Multiple Classification Analysis Model 2; MCA3-Multiple Classification Analysis Model 3.

explain more than 99% of the variation in household trips in the 1965 dataset, while MCA2 explains about 97.64% of the variation.

In the conventional four-step UTMS modeling approach, regardless of the unit employed in trip generation, the predicted trips by households are aggregated to traffic zone levels for input into the trip distribution or modal choice step. Consistent with this procedure, the study combines the trips predicted for households by each of the four models to traffic zone levels. Afterwards, the observed zonal trips were regressed against the predicted zonal trip productions yielded by each of the four models. A summary of the results are in the bottom half of Table 2. Based on the coefficient of

determination (R^2) CCA, MCA1, and MCA3 explain over 96% of the variance in the observed trips at the traffic zone level while MCA2 explains slightly over 95% of this variance. Using the mean absolute error measure (PMAE), the CCA model yields the lowest error measure of 12.373. The MCA models, which were supposed to address the shortcomings of CCA, yield zonal predictions of trips that have greater error compared with the CCA model. The results from regressing the observed zonal trips against the predicted zonal trips using the four models are presented in columns two and three of the bottom half of Table 2.

Table 2: Performance of 1965 CCA and MCA Models in Predicting Trips at Household and Traffic Zone Levels in 1965

Household Level				
Model	**Intercept**	**Slope**	**R2**	**PMAE**
CCA	0.00	1.0000	1.0000	0.000
MCA1	140.68	0.9939	0.9957	9.264
Standard Error	140.00	0.0136		
t-value	1.00	73.0700		
MCA2	466.63	0.9222	0.9764	26.887
Standard Error	323	0.0299		
t-value	1.44	30.8400		
MCA3	-61.52	1.0090	0.9972	9.605
Standard Error	114.00	0.0111		
t-value	-0.54	90.6100		
Traffic Zone Level				
Model	**Intercept**	**Slope**	**R2**	**PMAE**
CCA	-15.97	1.0266	0.9661	12.373
Standard Error	8.39	0.0115		
t-value	-1.90	89.5900		
MCA1	-10.07	1.0318	0.9633	12.895
Standard Error	8.68	0.0120		
t-value	-1.16	86.0600		
MCA2	8.04	0.9968	0.9556	14.730
Standard Error	9.38	0.0128		
t-value	0.86	77.9100		
MCA3	-16.80	1.0280	0.9654	12.758
Standard Error	8.48	0.0116		
t-value	-1.98	88.7100		

Note: CCA-Cross-Classification Analysis; MCA1-Multiple Classification Analysis Model 1; MCA2-Multiple Classification Analysis Model 2; MCA3-Multiple Classification Analysis Model 3.

Forecast Performance of Alternative Models

The models estimated in the first step were then used to forecast travel in the years for which data were available. The time lag between 1965 and 1981, 1990, 1996, and 2000 provided for an assessment of the medium- to long-term forecast performance of these models. The results of the analyses are in Table 3. The regression results of the observed number of daily trips made by households in each cross-classification cell against the corresponding number of daily trips predicted by the models (CCA, MCA1, MCA2, or MCA3) for each cross-classification cell are in columns three to nine of this table. Examining the 1981 results, with the exception of MCA2 all the models explained in excess of 98% of the variation in the observed trips made by households. Values of the percent mean absolute error measure in column 10 of Table 3 for all the models were smallest for 1981 compared with those in any of the other years. Focusing on 1981, CCA had the smallest percent mean absolute error measure value of 11.69% as shown in the tenth column of the second row of Table 3.

For 1990, the models explain slightly more than 91% of the variation in household trips, which is about 7% less than the explained variation using the 1981 dataset for CCA, MCA1, and MCA3 models. Additionally, the error measures for CCA, MCA1, and MCA3 in 1990 is approximately double their corresponding values in 1981, while that for the MCA2 model declines. The CCA model performs better than the MCA models for the 1990 application. The three models, CCA, MCA1, and MCA3, explain trip variation at the household level in excess of 96% using the household trip data in 1996. MCA2, on the other hand, explains 85% of the variation in the trip data, which is about 11% less than that for the other three models. In terms of the error measure, the CCA model ranks first – it has the lowest percent average error, followed by MCA1 and then MCA3. The MCA2 model yields the highest percent average error and therefore ranks fourth.

The application of the models to generate long-term forecasts for 2000 yielded results similar to those obtained for 1990 and 1996. In terms of explaining trip variation at the household level, all the models performed well by explaining more than 96% of the total variation in the trips made by households. In general, the CCA model yields travel forecasts with lower error values compared with error values obtained with forecasts by the MCA models in all the applications.

Prediction of Household Trip Rates for Cross-Classification Cell with Inadequate Data

As discussed earlier, a challenge in the use of CCA is the possibility of having a number of cells of the cross-classification matrix having few or no observations. The primary concern under such circumstances should be with the problematic cells only; that is, those with little or no data and not the entire trip-rate matrix. However, current planning practice calls for modifying the entire household trip-rate matrix obtained by CCA (Ortuzar and Willumsen 2001) rather than just the problematic cells. This study applies the same MCA models to estimate a household trip rate for each of the problematic cells only, while preserving the household trip rates obtained from ordinary CCA for the remaining cells. In addition to the MCA models, two other techniques for estimating missing cell values, namely KNN and MI, are employed for predicting household trip rates for only those empty and/or unreliable cells, and after the forecast performance of all these models are assessed. It is noted that no study was found in the literature that employed MI or KNN to synthesize household trip rates for cross-classification cells with inadequate data.

From the ordinary cross-classification analysis results using the 1965 data presented in Table 1, each cell of the cross-classification matrix had observations. However, based on the threshold number of observations required for statistical reliability reported in Ortuzar and Willumsen (2001) the number of observations for three of the cells was low. The defining characteristics of these cells are: (1) single-person households owning three vehicles, (2) single-person households owning four or more vehicles, and (3) two-person households owning four or more vehicles. Therefore, the three

Table 3: Performance of 1965 Models in Predicting Trips by Households in Each Cross-Classification Cell in Years 1981, 1990, 1996, and 2000

Year	Model	Intercept			Slope				
		Coefficient	Standard Error	t-value	Coefficient	Standard Error	t-value	R^2	PMAE
1981	CCA	39.89	37	1.09	1.0188	0.0138	73.84	0.9960	11.69
	MCA1	112.47	78	1.43	1.0156	0.0304	33.37	0.9800	15.98
	MCA2	365.71	181	2.02	0.8775	0.0668	13.14	0.8820	33.18
	MCA3	21.98	52	0.42	1.0247	0.0200	51.3	0.9910	16.00
1990	CCA	18.15	265	0.07	1.2393	0.0719	17.24	0.9280	27.12
	MCA1	240.30	287	0.83	1.2377	0.0811	15.26	0.9100	30.14
	MCA2	287.76	268	1.07	1.0909	0.0670	16.27	0.9200	32.40
	MCA3	42.40	249	0.17	1.2059	0.0659	18.31	0.9360	31.17
1996	CCA	32.93	53	0.62	1.3164	0.0385	34.23	0.9807	27.62
	MCA1	58.80	68	0.86	1.3191	0.0503	26.21	0.9676	29.87
	MCA2	145.59	145	1.00	1.1724	0.1015	11.55	0.8529	34.54
	MCA3	46.13	70	0.66	1.2681	0.0490	25.88	0.9668	28.96
2000	CCA	173.75	221	0.79	1.2776	0.0337	37.95	0.9840	28.76
	MCA1	354.12	247	1.43	1.2588	0.0377	33.39	0.9800	31.47
	MCA2	467.17	334	1.40	1.1201	0.0460	24.36	0.9630	31.91
	MCA3	285.78	281	1.02	1.2229	1.2229	29.41	0.9740	31.45

Note: CCA-Cross Classification Analysis; MCA1-Multiple Classification Analysis Model 1; MCA2-Multiple Classification Analysis Model 2; MCA3-Multiple Classification Analysis Model 3.

MCA models, and KNN and MI in turn, were used to synthesize household trip rates for just these three cells that would otherwise have unreliable household trip rates. The remaining cells of the matrix retained their household trip rates obtained from the ordinary cross-classification analysis (CCA). The estimated household trip rates for these cells obtained by the CCA, MCA, KNN, and MI are in Table 4. For each of the three cells, the estimated household trip rate by MCA1, MCA2, MCA3, KNN, or MI exceeds the corresponding household trip rate obtained by CCA. Further, replacing the household trip rates obtained by CCA for the three problematic cells with those yielded by any of the models results in the expected increasing relationship between household trip rates and increasing household size or increasing vehicle ownership respectively.

Forecast Performance of Household Trip-Rate Matrices Developed

Table 5 presents the values of the measures for evaluating the accuracy of household trip predictions given by the five alternative models, respectively. The measures are evaluated using the observed trips and the predicted trips made by households in each cross-classification cell.

Evaluation of the accuracy of predictions of travel in 1965. In the base year (1965), MCA1 had the lowest PMAE value and therefore the best performance in predicting travel based on this measure. It is followed by MI. MCA2 had the worst performance in predicting travel, reflected by it having the highest PMAE value. The coefficient of determination is one for all the models, indicating that each explained all the variation in household trips in the base year.

Table 4: Predicted Household Trip Rates for Cells of 1965 Trip-Rate Matrix with Inadequate Data Given by Alternative Models

Model	H^1=1, V^2=3	H^1=1, V^2=4+	H^1=2, V^2=4+
Cross Classification Analysis	3.941	5.000	6.889
Multiple Classification Analysis Model 1	5.290	5.486	7.445
Multiple Classification Analysis Model 2	6.586	7.314	9.833
Multiple Classification Analysis Model 3	6.404	6.571	7.989
Multiple Imputation	5.351	6.585	8.912
K-Nearest Neighbor	6.186	7.139	8.344

1. H = Size of the household
2. V = Number of vehicles available to the household

Evaluation of the accuracy of predictions of travel in 1981. Based on the coefficient of determination, all the models explain in excess of 99% of the variation in household trips in 1981. CCA yielded the lowest PMAE value of 11.689, indicating it had a travel forecast accuracy superior to that of the other models. MCA1 had the next lowest PMAE value followed by MI, then KNN, and then MCA3. MCA2 had the highest PMAE value due to the rather large household trip rate estimates it gives for the three cells with inadequate data (see Table 4). For each model, the PMAE value in 1981 is higher than the corresponding value in 1965.

Evaluation of the accuracy of predictions of travel in 1990. The coefficient of determination using the predictions of household travel by each of the models ranges from 0.924 to 0.928, indicating that the models are able to explain in excess of 92.4% of the variation in household trips. The values of PMAE range from 27.117 for CCA to 29.392 for KNN. MCA1 has the second lowest PMAE value (27.260). This indicates that based on this measure (PMAE) the CCA model of household trip rates, notwithstanding three of the cells having inadequate data, gives more accurate household travel forecasts than those given by the other models. Immediately following this is MCA1. Again, for each model, the PMAE value in 1990 is higher than the corresponding value in 1981.

Evaluation of the accuracy of predictions of travel in 1996. The coefficient of determination evaluated using the predictions of household travel by the six models ranges from 0.975 for MCA2 to 0.981 for CCA. This indicates that the models explain in excess of 97.5% of the variation in household trips. PMAE is highest for MCA2 (31.972), indicating the worst forecast performance of household travel based on this measure. CCA has the lowest PMAE value of 27.619, indicating the best forecast performance of travel based on this measure. Immediately following it is MCA1, which has a PMAE value of 28.215. MI, with a PMAE value of 29.453, has the third best forecast performance of travel. Again, for each model, the PMAE value in 1996 is higher than the corresponding value in 1990.

Evaluation of the accuracy of predictions of travel in 2000. The coefficient of determination based on the predictions of household travel by each of the models is 0.983. This indicates that all the models are able to explain 98.3% of the variation in household trips. KNN has the highest PMAE value of 32.686, indicating the worst forecast performance of household travel based on this measure, while CCA with a PMAE value of 28.756 has the best forecast performance of household travel. MCA1, with a PMAE value of 30.147, has the next best forecast performance of household travel.

Table 5: Performance of 1965 Alternative Models in Predicting Trips by Households in Each Cross-Classification Cell in Years 1965, 1981, 1990, 1996, and 2000

Year	Model	Intercept			Slope				
		Coefficient	Standard Error	t-value	Coefficient	Standard Error	t-value	R^2	PMAE
	MCA1	-4.46	2	-2.14	1.000	0.0002	4868	1.000	2.287
	MCA2	-15.90	9	-1.83	1.001	0.0008	1193	1.000	6.245
	MCA3	-8.28	4	-2.19	1.001	0.0004	2742	1.000	4.395
	KNN	-9.78	5	-2.13	1.001	0.0004	2253	1.000	4.834
1965	MI	-9.64	5	-1.79	1.001	0.0005	1919	1.000	3.624
	CCA	39.89	37	1.09	1.0188	0.0138	74	0.996	11.689
	MCA1	35.97	36	0.98	1.020	0.0139	73	0.996	13.521
	MCA2	25.57	38	0.67	1.023	0.0145	71	0.995	16.554
	MCA3	32.38	37	0.87	1.021	72.690	0	0.996	15.279
	KNN	31.35	37	0.84	1.021	72.470	0	0.996	15.194
1981	MI	31.28	37	0.84	1.021	0.0141	72	0.995	14.208
	CCA	18.15	265	0.07	1.2393	0.0719	17	0.928	27.117
	MCA1	1.09	265	0.00	1.242	0.0720	17	0.925	27.260
	MCA2	-39.36	268	-0.15	1.250	0.0727	17	0.924	29.421
	MCA3	-15.43	266	-0.06	1.245	0.0723	17	0.925	29.126
	KNN	-19.74	267	-0.00	1.246	0.0723	17	0.925	29.392
1990	MI	-16.78	266	-0.06	1.245	0.0722	17	0.925	27.787
	CCA	32.93	53	0.62	1.3164	0.0385	34	0.981	27.619
	MCA1	23.17	54	0.42	1.320	0.0395	33	0.979	28.215
	MCA2	4.37	59	0.07	1.325	0.0432	31	0.975	31.972
	MCA3	14.02	56	0.25	1.323	0.0408	32	0.978	30.165
	KNN	12.39	56	0.22	1.323	0.0410	32	0.977	30.679
1996	MI	15.77	56	0.28	1.322	0.0408	32	0.978	29.453
	CCA	173.75	221	0.79	1.2776	0.0337	38	0.984	28.756
	MCA1	156.89	222	0.71	1.279	0.0340	37	0.983	30.147
	MCA2	127.22	227	0.56	1.281	0.0350	36	0.983	30.852
	MCA3	140.90	223	0.63	1.280	0.0342	37	0.983	32.142
	KNN	136.78	224	0.61	1.281	0.0343	37	0.983	32.686
2000	MI	138.87	224	0.62	1.280	0.0343	37	0.983	31.326

Note: CCA-Cross-Classification Analysis; MCA1-Multiple Classification Analysis Model 1; MCA2-Multiple Classification Analysis Model 2; MCA3-Multiple Classification Analysis Model 3.

SUMMARY AND CONCLUSIONS

This paper investigated the forecast performance of trip generation models based on cross-classification (CCA) and multiple classification analysis (MCA). In addition, it examined the replacement of household trip rates in unreliable cross-classification cells with values estimated by three MCA models and two methods for estimating missing values namely Multiple Imputation (MI) and K-Nearest Neighborhood (KNN). The results of the study lead to the following conclusions.

First, the methods that call for modifying the entire household trip rate matrix obtained from ordinary cross-classification analysis give a performance in prediction of household travel that is worse than that given by the methods that call for synthesizing household trip rates for cells with inadequate data only while preserving the other household trip rates obtained from ordinary cross-classification analysis. This result is evident by comparing the upper part of Table 2 to the upper part of Table 5. Thus, it is concluded that adjusting all the trip rates of a CCA matrix using MCA, the current industry standard, results in a forecasting model that is inferior to CCA and hence should be avoided by practitioners. Whenever cells with inadequate data exist in a CCA matrix, the substitution of the trip rates of these unreliable cells only with trip rates obtained from the MCA models, the MI method, or KNN results in more accurate forecasts compared with adjusting the trip rates for all the cells.

Second, even though three of the cells of the ordinary cross-classification matrix had inadequate data, the model surprisingly and consistently gave the best performance in the prediction of household travel in both the medium and the long term (see column 10 of Table 3). Thus, the basic CCA model is robust and practitioners can use it to provide credible forecasts of travel if few of the cells of the CCA matrix are unreliable. It may also indicate that the recommended minimum number of observations for a cell can perhaps be reduced and still lead to the development of reliable cross-classification models. It is for future research to determine the appropriate minimum number of observations for a cell.

Third, replacing the unreliable household trip rates of an ordinary CCA matrix with household trip rates estimated using the MCA models, KNN and MI did improve upon the performance of the cross-classification model compared with adjusting all the trip rates of the CCA matrix. Among these methods for synthesizing a household trip rate, on average, MCA1 and MI have the lowest error values (column 10 of Table 5). However, since MCA1 is subject to biases (Guevara and Thomas 2007), the MI model may be preferred over MCA1 even though MCA1 may be a simpler model compared with the MI model.

Finally, the forecast performance of cross-sectional models declines with time. For example, the corresponding PMAE values associated with each model increased with the time interval between the base and application years (see column 10 of Table 5). This certainly is logical because of the greater changes expected to occur in land use patterns, socio-demographic characteristics and attitudes of the population, transportation system characteristics, and technology with time elapsed from the base year. Thus, irrespective of the method used to synthesize household trip rates for unreliable cells, the further out the application year the greater the inaccuracy of travel forecasts. The prime limitation of this study is with the single region source of the dataset used. Clearly, to generalize, the conclusions tests have to be done on data from several other regions.

APPENDIX A: Results of Analysis of Variance and Correlation Analysis Respectively

Analysis of Variance

Source	Partial Sum of Squares	Degrees of Freedom	Mean Square	F value	P value
Model	309730	82	3777	104.73	0.0000
Household Size	127687	18	7094	196.68	0.0000
Number of Motorcycles	623	3	208	5.76	0.0006
Number of Drivers	9537	8	1192	33.05	0.0000
Tenure	346	8	43	1.20	0.2942
Household Income	11669	14	833	23.11	0.0000
Dwell Type	491	11	45	1.24	0.2556
Number of Vehicles	1551	20	77	2.15	0.0021
Residual	682363	18919	36		
Total	992093	19001	52		
Number of Observations	19002				
R-squared	0.3122				

Correlation Matrix

	Household Size	Number of Motorcycles	Number of Drivers	Number of Vehicles	Household Income
Household Size	1				
Number of Motorcycles	0.0466	1			
Number of Drivers	0.4717	0.0955	1		
Number of Vehicles	0.2898	0.0460	0.5294	1	
Household Income	0.1426	0.0169	0.2920	0.2805	1

Acknowledgements

The authors would like to thank Charles Purvis of the San Francisco Bay Area Metropolitan Transportation Commission for providing data used in this study. Additionally, the first author conducted this research while at Tennessee Technological University.

References

Abdel-Aal, M.M. “Cross Classification Trip Production Model for the City of Alexandria.” *Alexandria Engineering Journal* 43 (2), (2004): 177-189.

Cotrus, A., J.N. Prashker, and Y. Shiftan. “Spatial and Temporal Transferability of Trip Generation Demand Models in Israel.” *Journal of Transportation and Statistics* 8 (1), (2003): 37-56.

Ewing, R., M. Deanna, and S. Li. “Land Use Impacts on Trip Generation Rates.” *Journal of the Transportation Research Board* 1518, (1996): 1–6.

Guevara, C. A. and A. Thomas. “Multiple Classification Analysis in Trip Production Models.” *Transport Policy* 14, (2007): 514-522.

Horton N.J. and K.P. Kleinman. "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models." *American Statistical Association* 61(1), (2007): 79-90.

Huisken, G. "Neural Networks and Fuzzy Logic to Improve Trip Generation Modeling." Paper presented at the 79th Annual Meeting of the Transportation Research Board, Washington, D.C., 2000.

Kikuchi, S. and J. Rhee. "Adjusting Trip Rate in the Cross-Classification Table by Using the Fuzzy Optimization Method." *Journal of the Transportation Research Board* 1836, (2003): 76–82.

Metropolitan Transportation Commission (MTC). "Trip Linking Procedures: 1990 Bay Area Household Travel Survey." Working Paper 2. Oakland, CA, Revised June 2003.

Muhammad S.B., I.G. Sehgal, and D. Laurence. "K-Ranked Covariance Based Missing Values Estimation for Microarray Data Classification." Proceedings of the 4th International Conference on Hybrid Intelligent Systems, Kitakyushu, Japan, December 5-8, 2004.

Ortuzar, J.D. and L.G.Willumsen. *Modelling Transport*. 3rd edition. John Willey and Sons, Inc., Chichester, England, 2001.

Purvis, C.L. "Changes in Regional Travel Characteristics and Travel Time Expenditures in the San Francisco Bay Area: 1960 – 1990." *Journal of the Transportation Research Board* 1466, (1994): 99–109.

Rengaraju, V. and M. Satyakumar. "Structuring Category Analysis Using Statistical Technique." *Journal of Transportation Engineering* 20 (6), (1994): 931-939.

Rengaraju, V. and M. Satyakumar. "Three-Dimensional Category Analysis Using Probabilistic Approach." *Journal of Transportation Engineering* 121(6), (1995): 538-543.

Retherford, R. D. and M.K. Choe. *Statistical Model for Causal Analysis.* John Wiley & Sons, New York, Inc., NY, 1993.

Rubin D.B. "Inference and Missing Data." *Biometrika* 63, (1976): 581–590.

Stopher, P.R. and K.G. McDonald. "Trip Generation by Cross-Classification: An Alternative Methodology." *Journal of the Transportation Research Board* 944, (1983): 84–91.

Transportation Research Board (TRB). "Metropolitan Travel Forecasting: Current Practice and Future Direction." Special Report 288, Washington, DC, 2007.

Walker, T. and H. Peng. "Long-Range Temporal Stability of Trip Generation Rates Based on Selected Cross-Classification Models in the Delaware Valley Region." *Journal of the Transportation Research Board* 1305, (1991): 61–71.

Wardman, M.R. and J.M. Preston. "Developing National Multi-modal Travel Models: A Case Study of the Journey to Work." Paper presented at the 9th World Conference on Transport Research, Seoul, South Korea, 2001.

Wayman J.C. "Multiple Imputation For Missing Data: What Is It And How Can I Use It?" Paper presented at the *2003 Annual Meeting of the American Educational Research Association*, Chicago, Illinois, 2003.

Zhao, H. "Comparison of Two Alternatives for Trip Generation." Paper presented at the 79th Annual Meeting of the Transportation Research Board, Washington, D.C., 2000.

Judith L. Mwakalonge *is an assistant professor in the civil and mechanical engineering technology department at South Carolina State University. Mwakalonge teaches and conducts research in transportation. Her primary research interests include travel demand modeling, model transferability, and traffic operations. Mwakalonge's current work explores the use of Radio Frequency Identification (RFID) and Bluetooth technologies in transportation and climate change as it relate to transportation. She received her Ph.D. degree in civil engineering from Tennessee Technological University.*

Daniel A. Badoe *is a professor in the department of civil and environmental engineering at Tennessee Technological University. He teaches undergraduate and graduate courses in transportation engineering and planning, and conducts research in the area of transportation data collection, travel behavior analysis, and travel demand modeling. He received his Ph.D. degree in civil engineering from the University of Toronto, Canada.*