UTC-WP-08-02

Trip Generation Revisited: Estimation of Trip Generation Rates from Small-size Household Travel Surveys

by

Paul Metaxatos
Research Assistant Professor
Urban Transportation Center (M/C 357)
University of Illinois at Chicago
412 South Peoria Street
Suite 340
Chicago, IL 60607-7036
Tel: (312) 996-4820
Fax: (312) 413-0006
E-mail: pavlos@uic.edu

and

Rita Morocoima-Black
Transportation Manager
Champaign County Regional Planning Commission
1776 E. Washington St. P.O. Box 17760
Urbana, IL 61803-7760
Phone: (217) 328-3313
Fax: (217) 328-2426
E-Mail: rblack@ccrpc.org

FEBRUARY 2008

UIC UNIVERSITY OF ILLINOIS AT CHICAGO Urban Transportation Center
COLLEGE OF URBAN PLANNING & PUBLIC AFFAIRS

ABSTRACT

We revisit, in this paper, issues facing transportation planners estimating and validating household trip generation rates from small-scale household travel surveys. Three problems are addressed: (a) unusual observations, (b) small number of observations, and (c) no observations. Unusual observations are identified using traditional methods. Classification and regression tree (CART) analysis is proposed for the second problem. Finally, the third problem is addressed using row-column decomposition analysis. The methods are demonstrated using a small-scale household travel survey and are simple enough to be implemented with the resources available to transportation analysts, especially in smaller Metropolitan Planning Organizations.

Keywords: small samples, household travel surveys, smaller MPOs, trip generation rate reliability, row-column decomposition analysis, classification and regression tree analysis, imputation.

1. INTRODUCTION

Within the context of the Federal transportation program states and local agencies must have technical planning processes in place to be able to program federal funds. As a result, Metropolitan Planning Organizations (MPOs) must develop and maintain transportation models to support various transportation and land use policies. Such requirements raise concerns, particularly among smaller MPOs that lack the resources to undertake expensive primary data collection such as household travel surveys. Frequently, transportation professionals are called to conduct complicated modeling tasks based on small samples. In such cases the investigation of data quality issues during the model development stage becomes critical.

Although several new approaches in travel modeling methods have emerged in recent years, the traditional four-step procedures are still in widespread use. Perhaps because the modeling community's focus has somewhat shifted to newer models, some of the old questions about the reliability and stability of trip generation rates, particularly when sample sizes are small, remain unanswered. This paper revisits the trip generation step of the conventional four-step modeling process and focuses on problems that are particularly troublesome when relatively small samples (of the order of a few hundred observations in total) are used for data collection. We also focus only on categorical trip generation models – which are the ones used most often in practice – and for good statistical reason. Future papers will address other steps of the four-step procedure.

Difficulties encountered when using small surveys and methods proposed here to alleviate them are illustrated using a small-scale household travel survey from Champaign-Urbana, Illinois. In the sections below three specific issues are addressed: identification of outliers, improving the reliability of small samples, and imputation issues in cases of missing information.

Identification of outliers: When sample sizes are large, estimates are generally not affected too much by a suspect observation. However, with small sample sizes, a 'bad' observation (due to a mistake in recording data or a respondent giving misleading information) can play havoc with the quality of estimates and have a profound effect on forecasts. Such bad estimates need be of concern only when they are influential. Influential observations that also are typically outliers are discussed in a Section 4.

Reliability of small samples: Table 1 shows a typical (categorical) trip generation table. The total number of observations is 360 with the consequence that cells representing several categories of household sizes and numbers of workers have only very few observations. In such a case, reliability can be enhanced combining table cells with like trip rates. A method for doing so is classification and regression tree (CART) analysis. The procedure is well known in the statistical literature but has not been used much in transportation. It is discussed in Section 5.

Imputation: Finally, some cells might not receive an observation at all – or the planner might find the observations too few or too unreliable. In that case, cell estimates might be imputed from observations in other cells. A procedure we have called row-column decomposition analysis is shown in Section 6. Imputation can also be used in lieu of combining of cells using CART.

While the focus of this paper is on obtaining estimates from small surveys, some of the discussion is germane to all trip generation efforts. Notice that the concern is not so much on the overall size of the sample but rather the size of the sample in each cell of the cross classification. Most applications of the four-step procedure are typically implemented using two-way classifications, e.g., number of adults in household and number of vehicles. Improvements can be made increasing the number of variables in the classification under consideration. This has not been done because then the observations per cell would decline. The methods given in this paper would be used to address this issue.

2. STUDY AREA AND DATA

The Champaign-Urbana-Savoy urbanized area has a 2000 population of almost 123,900 people (including University of Illinois students) who live in an area of more than 40 square miles (http://www.ccrpc.org/planning/transportation/lrtp/pdf/2%20CUSB%20Urbanized%20Area,%2011-22-04.pdf). The area is located in east central Illinois, approximately 135 miles south of Chicago.

The data used for the analysis in this paper are based on the Champaign-Urbana-Savoy, 2002 household travel survey conducted by the local MPO, Champaign Urbana Urbanized Area Transportation Study (CUUATS) in Illinois (Morocoima-Black and Kang, 2003). The survey was conducted to facilitate the development of a transportation model in support of the Long Range Transportation Plan for the urbanized area.

The survey, a one-day home-interview survey, collected data on the average weekday, local and regional personal travel made by residents of the Champaign-Urbana-Savoy urbanized area and was conducted from summer of 2002 to spring 2003. It had a 15% response rate and contained a sample of 362 usable households or, approximately, 0.7% of households in the study area. For comparison purposes and given that sampling errors are dependent on absolute sizes of samples and not on percentages, the 1990 Chicago Area Transportation Study's 1990 Household Travel Survey obtained a sample of more than 19,000 usable households or about 0.7% sample of households.

3. SMALL SAMPLES IN TRIP GENERATION

Small samples in trip generation occur for two reasons. The first reason is that simply the size of the overall survey sample is small. This situation may occur, for example, at smaller MPOs where there aren't enough resources (e.g., staff, budget, etc.) available to obtain a sample size suitable to meet the survey objectives. As a result, population parameters (e.g., mean trip rates per unit, modal proportions, etc.) that will be estimated based on the survey will be subject to greater sampling variability. The paper offers later a discussion to address this problem in Sections 5 and 6.

On the other hand, quite often the statistical analysis of survey data during trip generation modeling results in small samples. For example, in cross-classification analysis during model development, a table of, say, family size by the number of workers in the household may have cells that have small samples of responding units (households, persons, etc.) or no samples at all. Moreover, higher dimension classifications would result in cells with even 'thinner' samples.

To illustrate the previous point, let us consider the categorical (cross-classification) trip generation model in Table 1. This model estimates the average number of trips (trip rate) made by households using a two-way classification of the factors family size and number of workers, each with several categories (e.g., one-member households, two-member households, etc.; zero-worker households, one-worker households, etc). Notice that several of the categories (cells) in Table 1 have a very small number of responding households.

Table 1. Average Number of Trips per Household
(number of households in parenthesis)

Number of Workers	Household Size			
	1	2	3	4+
0	6.28 (53)	9.22 (40)	16.00 (3)	8.00 (3)
1	6.09 (87)	10.42 (45)	10.94 (18)	9.06 (15)
2	- (-)	9.56 (46)	10.81 (11)	15.46 (28)
3	- (-)	- (-)	11.00 (2)	11.83 (6)
4	- (-)	- (-)	- (-)	10.66 (3)

-: indicates category not possible in this classification.

The problem is amplified when we expand a cross-classification to include additional categories or independent variables. For example in a given travel survey, a four-by-five classification of family size and number of workers would result in smaller samples than a three-by-four classification of the same variables. Similarly, adding more variables would result in 'thinner' samples. For example, a three-by-four-by four classification of family size, number of workers and vehicle availability would have smaller number of samples than a two-way classification of the first two variables.

For illustration purposes in Table 2, let us consider, in addition to the variables in Table 1, four levels of auto availability ('auto' is a generic term used to include cars, vans and pick-up trucks) as a third classification variable, also from the CUUATS 2002 household travel survey. Clearly, there are several instances in this three-way classification with no households and several more with a very small number of households. Indeed in Table 2 we have 56 possible cells (five worker categories times four household categories times four vehicle categories resulting in 80 cells with 24 cells not possible), which strains the ability to get reasonable numbers of observations within each cell.

Table 2. Average Number of Trips per Household
(number of households in parenthesis)

Workers	Vehicle Availability	Household Size			
		1	2	3	4+
0	0	5.85 (14)	ND (0)	ND (0)	ND (0)
0	1	6.60 (35)	10.14 (14)	8.00 (1)	9.00 (2)
0	2	5.00 (4)	8.66 (24)	20.00 (2)	6.00 (1)
0	3+	ND (0)	9.50 (2)	ND (0)	ND (0)
1	0	6.09 (22)	ND (0)	ND (0)	ND (0)
1	1	6.05 (59)	10.44 (18)	9.36 (11)	8.88 (9)
1	2	6.60 (5)	10.45 (24)	13.60 (5)	10.00 (5)
1	3+	6.00 (1)	10.00 (3)	13.00 (2)	6.00 (1)
2	0	- (-)	8.00 (1)	ND (0)	ND (0)
2	1	- (-)	9.00 (12)	12.66 (3)	15.00 (4)
2	2	- (-)	9.92 (27)	9.25 (4)	15.85 (20)
2	3+	- (-)	9.33 (6)	11.00 (4)	14.00 (4)
3	0	- (-)	- (-)	ND (0)	ND (0)
3	1	- (-)	- (-)	ND (0)	ND (0)
3	2	- (-)	- (-)	19.00 (1)	9.00 (2)
3	3+	- (-)	- (-)	3.00 (1)	13.25 (4)
4	0	- (-)	- (-)	- (-)	ND (0)
4	1	- (-)	- (-)	- (-)	ND (0)
4	2	- (-)	- (-)	- (-)	ND (0)
4	3+	- (-)	- (-)	- (-)	10.66 (3)

-: indicates category not possible in this classification; ND: no data.

4. IDENTIFICATION OF OUTLIERS

The first problem addressed in this paper is the identification of outliers that emerges when survey responses (observations) that involve trip-making activity stand out as unusual. This problem is particularly vexing when the observation is the only one or one of very few in a cell (in a cross-classification table) and has an inordinate influence on trip rates. To counteract this, large numbers of observations would be required, especially in higher-dimension classifications. However, this is often not feasible in the smaller sample sizes typically employed in smaller towns.

To identify outliers in the CUUATS 2002 household travel survey we employed traditional methods found in standard statistical textbooks (see for example, Sen and Srivastava, 1990). If we consider the trip generation model in Table 1 as a regression problem (see Thakuriah et al., 1993 for a discussion about the equivalency between categorical and regression models) with trip rates as the dependent variable and the number of workers and family size as the independent variables, then it is reasonable to assume that outliers would exhibit numerically large residuals. However, not all influential points would have large residuals. Therefore, we focused our attention to households with a large residual or located far away from other households in the space of the independent variables (number of workers and family size in this case).

There are a great number of measures available in the literature for identifying outliers (Sen and Srivastava, 1990, p.161). In this paper, we examined the Studentized residuals (see Sen and Srivastava, 1990, p.156 for an exact definition) resulting from the regression above. It can be proved that this quantity has a *t* distribution when the errors are Gaussian and has a near *t* distribution under a wide range of circumstances. A Studentized residual is a standardized measure of the distance between a case (an observation) and the model estimated on the remaining cases. Therefore, it can be used as a test statistic to determine whether a case belongs to the model.

For a search of influential points we examined the measure DFFITS (Sen and Srivastava, 1990, Section 8.5). The statistic measures (in the previous regression) how much the predicted value of the dependent variable (trip rates) would be affected at a particular point (observation) if that observation were to be deleted. It can be proved that DFFITS is a functionally related with the Studentized residuals. If the latter increases then DFFITS increases too. Both Studentized residuals and DFFITS measures are available in the regression output of software packages for statistical analysis.

Using both measures above, we found a list of unusual cases that is shown in Table 3. Case #51, for example, shows a household with one member and no employees taking 24 trips when the average trip rate for the group is 6.28 (Table 1). The studentized residual for this case is 3.93 substantially higher than a cutoff point of 2. As an aside, a cutoff point of 2 would mark 5 percent of the observations as unusual because under normality assumptions studentized residuals follow approximately a t distribution.

We also applied the criterion $2\left[(14+1)/360\right]^{1/2} = 0.408$ as a cutoff point for DFFITS (Sen and Srivavstava, p. 160) given 360 observations and 14 independent variables (the indicator variables corresponding to each trip rate category in Table 1). Not surprisingly, almost all of the cases with large residuals also had large DFFITS values. An inordinate amount of influence can be seen in Table 3 in cases stemming from small samples (e.g., case numbers 273, 304, 305 and 352). Also, case number 349 appears to be quite unusual and influential.

Table 3. Identified Outliers and Influential Observations

Case Number	Studentized Residual	DFFITS Value	Number of Employees	Household Size	Number of Samples	Number of Trips	Average Trip Rate in Group
51	3.93	0.55	0	1	53	24	6.28
91	2.41	0.26	1	1	87	17	6.09
186	3.01	0.46	1	2	45	24	10.42
263	2.32	0.34	2	2	46	20	9.56
273	-2.15	-1.53	0	3	3	8	16.00
275	2.50	0.61	1	3	18	22	10.94
281	2.72	0.66	1	3	18	23	10.94
288	-2.02	-0.49	1	3	18	2	10.94
291	2.27	0.55	1	3	18	21	10.94
300	2.57	0.82	2	3	11	22	10.81
304	-2.48	-2.50	3	3	2	3	11.00
305	2.48	2.50	3	3	2	19	11.00
327	-2.56	-0.49	2	4+	28	4	15.46
340	2.80	0.54	2	4+	28	28	15.46
345	2.58	0.50	2	4+	28	27	15.46
349	6.61	1.35	2	4+	28	45	15.46
350	2.35	0.45	2	4+	28	26	15.46
352	2.44	1.10	3	4+	6	22	11.83

The unusual cases in Table 3 would normally invite further scrutiny. It would be important to understand the dynamics of household formation that gives rise to such trip patterns. For example, case number 273 belongs to a category (cell) in which two of the three households surveyed were retired families with high incomes, and the third one was a family with several children headed by a student with a part-time job. In another example, in case number 349 one of the four household members recorded a rather high number of 20 short trips.

The analyst at this point needs to understand what behavior is truly unusual and influential or simply the result of a small-scale survey. If the former is true one could drop such observations from further analysis because there is enough evidence that such observations do not belong in the same model. If the latter were true then a course of action discussed in the next section would improve the reliability of trip rates.

5. IMPROVING THE RELIABILITY OF TRIP GENERATION RATES WITH CART PROCEDURES

To improve the reliability of trip rates in the presence of small samples we propose the use of classification and regression tree (CART) analysis. CART analysis is based on the binary decision tree algorithm (Breiman *et al.,* 1984). The method is also documented in Ripley (1996), and Venables and Ripley (1997). In this section we will demonstrate the method using a two- and a three-independent variables examples.

Background Information

As a non-parametric method of estimating conditional distributions, CART models have some potential advantages over parametric models (Reiter, 2003). First, CART modeling may be more easily applied than parametric modeling, particularly for continuous data that are truncated or not smooth. Second, CART models can capture nonlinear relationships and interaction effects that may not be easily revealed in the process of fitting parametric models. Third, CART provides a semi-automatic way to fit the most important relationships in the data, which can be a substantial advantage when there are many potential predictors. Primary disadvantages of CART models relative to parametric models include difficulty of interpretation, discontinuity at partition boundaries, and decreased effectiveness when the data follow relationships easily captured by parametric models (Friedman, 1991).

The CART algorithm is a binary (nodes are always split into two) recursive partitioning algorithm. The original version uses the Gini index of diversity as the default splitting criterion (Breiman *et al.,* 1984). In this paper we used Clark and Pregibon's (1992) variation implemented in the S language that uses the deviance as the splitting criterion. Clark and Pregibon (1992) define the deviance of a node as the sum of the deviances of all observations in the node (Equation 1). This method is based on an impurity index known as entropy (Ripley, 1996, p. 12). If the deviance of a node is not zero and there are sufficient observations in the node, splitting proceeds by comparing the deviance of the node to that of two possible subnodes. The split that maximizes the change in deviance is chosen.

$$D = -2\sum_{j=1}^{J}(n_{jL}\log p_{jL} + n_{jR}\log p_{jR}) \qquad (1)$$

where, *D,* is the deviance, n_{jL} (n_{jR}) denote the number and p_{jL} (p_{jR}) the proportion of observations in the left (right) node in level *J*.

CART-type analysis has been growing in popularity since a variety of algorithmic approaches have been implemented in major statistical packages. Some implementations are licensed as add-ons requiring additional license fees that can be substantial during the licensing period. A number of implementations are disseminated as free software and give the best chance to practitioners to start experimenting with such methods. Examples of free software (available under the Open Source license agreement) that we have had a limited experience with include: the GUIDE Regression Tree software (http://www.stat.wisc.edu/~loh/guide.html), the WinMine

Toolkit (http://research.microsoft.com/%7Edmax/WinMine/tooldoc.htm), and Classification Tree in Excel (http://www.geocities.com/adotsaha/CTree/CtreeinExcel.html) (the last assumes that Microsoft's Excel is installed).

CART Demonstration Using Two Independent Variables

Let us start the CART analysis with the trip generation rates in Table 1. Running the same data through the CART algorithm produced the classification shown in Table 4. Notice that the (workers, household size) categories (0, 3) and (1, 3) are now one category with 21 households and 245 trips (the sum of the two categories it replaced as seen in Table 1). Similarly, categories (2, 3) and (3, 3), have formed a new category with 13 households and 141 trips. Also, the category formed by categories (0, 4+) and (1, 4+) has 18 households and 160 trips. Finally, categories (3, 4+) and (4, 4+) have been combined into one category with 9 households and 103 trips.

Table 4. Average Trip Rate per Household
(after CART classification)

Workers	Household Size			
	1	2	3	4+
0 # of households	6.28 53	9.22 40	11.67	8.89
1 # of households	6.09 87	10.42 45	21	18
2 # of households	- -	9.56 46	10.85	15.46 28
3 # of households	- -	- -	13	11.44
4 # of households	- -	- -	- -	9

-: indicates category not possible in this classification.

It is not uncommon in practice to estimate trip production rates by traffic analysis zone (TAZ) using estimated trip generation rates from a household travel survey based on a cross-classification of trip purpose by household size. For example, for each TAZ, CUUATS uses the formula

$$T_{pk} = \sum_i x_{ik} t_{ik} t_{ipk} \qquad (2)$$

where, T_{pk}, is the total number of trips by purpose p in TAZ k, x_{ik} is the number of households in TAZ k in (household size) category i; t_{ik} is the trip rate for households in category i in TAZ k; and t_{ipk} is the percent of trips for households in TAZ k in category i and trip purpose p (CUUATS, Long Range Transportation Plan – Appendix3: Transportation Model Report, p.17, 2004). The trip rates for this particular classification are given in Table 5. The number of cases that

contribute to a particular trip purpose and household size combination is shown under the respective trip rate (generally, households contribute to multipurpose trips). For this particular demonstration we have removed trips with either end outside the Champaign-Urbana urbanized area.

Table 5. Trip Rates by Household Size and Trip Purpose

Trip Purpose	Household Size			
	1	2	3	4+
Home-based Work	1.90	2.43	2.60	2.64
Number of cases	71	70	25	36
Home-based Shop	1.64	2.26	2.50	2.81
Number of cases	42	76	20	32
Home-based School	2.19	2.26	1.22	2.72
Number of cases	58	23	9	11
Home-based Other	2.44	4.00	3.97	6.14
Number of cases	93	105	29	47
Non-home Based	2.82	3.61	4.48	4.61
Number of cases	99	107	29	41

An alternative way to compute trip rates by household size and trip purpose is now proposed using the CART method. A CART analysis of Table 5 produced the trip rates shown in Table 6. The top number in each category is the trip rate and the bottom number is the number of households in the same category. The previous 20 categories in Table 5 have been fused into 10 categories in Table 6 with much larger sample sizes (it can readily be verified that, notwithstanding rounding errors, Tables 5 and 6 give the same total number of trips). In this regard, Table 6 is a reasonable alternative to Table 5 and offers the additional advantage that the estimation of trip rates are based on more samples.

The results from the CART analysis in Table 6 show that small-size (single-person), medium-size (two- and three-person), and larger-size (four-or-more person) households appear to have distinctive trip-making profiles in regard to relatively less discretionary (or more mandatory) trips (home-based work, home-based shop, and home-based school) as opposed to relatively more discretionary trips (home-based other and non-home based). If this observation can be validated from other reliable sources, it could have important implications in the design of household travel surveys.

Table 6. Trip Rates by Household Size and Trip Purpose (after CART Analysis)

Trip Purpose	Household Size			
	1	2	3	4+
Home-based Work				
Number of cases				
Home-based Shop	1.94	2.33		2.72
Number of cases	171	223		79
Home-based School				
Number of cases				

Home-based Other Number of cases	2.44 93	4.00 105	4.22 58	6.15 47
Non-home Based Number of cases	2.83 99	3.62 107		4.61 41

Comparison between CUUATS and CART models

Using the trip rates form Table 6, we could estimate the total number of trips from the formula

$$T_{pk}^{CART} = \sum_i x_{ik}^{CART} t_{ipk}^{CART} \quad (3)$$

where, T_{pk}^{CART}, is the total number of trips in TAZ *k* by purpose *p* using the CART procedure, x_{ik} is the number of households and trip purpose combinations in (household size) category *i* and TAZ *k* as determined by the CART procedure; and t_{ipk}^{CART} is the trip rate for households in category *i* and trip purpose *p* in TAZ *k*, also as determined by the CART procedure.

Clearly, the models in Equations (2) and (3) are different algebraically. However, they both give very similar total number of trips by purpose as shown in Table 7. As a result, the CART approach appears to be a reasonable alternative to more traditional procedures for trip generation estimation. The added value for the CART approach, however, is that it is based on trip rates that have been estimated from a richer sample of households (in this example, household-trip purpose combinations to be more accurate).

Table 7. Comparison between CUUATS and CART Models

Trip Purpose	1-person hhlds		2-person hhlds		3-person hhlds		4+ person hhlds		Total Number of Trips in Survey	
	CUUATS	CART	CUUATS	CART	CUUATS	CART	CUUATS	CART	CUUATS	CART
HBWork	137.7	136.6	163.1	166.9	58.3	58.1	97.9	97.8	457.0	459.5
HBShopping	81.5	76.9	177.1	179.7	46.6	46.5	87.0	83.8	392.2	386.9
HBSchool	112.5	111.0	53.6	51.4	21.0	19.4	29.9	27.9	217.0	209.7
HBOther	226.9	230.6	420.0	423.7	122.4	120.2	289.1	286.4	1058.4	1060.8
NHBased	280.2	281.8	387.3	385.1	122.4	124.0	189.0	195.6	978.9	986.6
Total # of trips	838.8	836.9	1201.1	1206.8	370.6	368.2	692.9	691.5	3103.5	3103.4

CART Demonstration Using Three Independent Variables

Using the same data as before and considering trip purpose as an additional independent variable, we present average trip rates per household for trips internal to the study area in Table 8. According to the survey documentation (Morocoima-Black and Kang, 2003), home-based work trips include trips from home to work and work-related business, and return to home. Home-based shopping trips include any kind of trips for shopping, and return to home. Home-

based school trips are trips from home to school and back home. Home-based other is a category for the remainder of home-based trips. Non-home based trips are those that do not begin or end at home.

Table 8. Mean Trip Rates for a Three-way Classification

Workers	Household Size	Trip Purpose	Number of Households	Mean Trip Rate	Variance	St. Error
0	1	hb-work	8	1.50	0.29	0.19
	1	hb-shop	17	1.65	0.99	0.24
	1	hb-school	28	2.29	1.47	0.23
	1	hb-other	39	2.64	3.08	0.28
	1	nh-based	39	3.10	5.20	0.37
0	2	hb-work	11	1.91	0.89	0.28
	2	hb-shop	25	2.44	2.34	0.31
	2	hb-school	8	2.25	1.36	0.41
	2	hb-other	34	4.24	4.97	0.38
	2	nh-based	32	3.16	3.36	0.32
0	3	hb-work	1	2.00	ND**	ND
	3	hb-shop	2*	5.50	0.50	0.50
	3	hb-school	1	1.00	ND	ND
	3	hb-other	3	6.00	7.00	1.53
	3	nh-based	3	4.00	1.00	0.58
0	4	hb-work	0	ND	ND	ND
	4	hb-shop	1	1.00	ND	ND
	4	hb-school	2	2.00	2.00	1.00
	4	hb-other	3	4.33	2.33	0.88
	4	nh-based	3*	2.00	0.00	0.00
1	1	hb-work	63	1.95	0.88	0.12
	1	hb-shop	25	1.64	1.24	0.22
	1	hb-school	30	2.10	1.33	0.21
	1	hb-other	54	2.30	2.85	0.23
	1	nh-based	60	2.65	2.77	0.22
1	2	hb-work	26	2.04	0.68	0.16
	2	hb-shop	23	2.04	1.04	0.21
	2	hb-school	10	2.20	0.84	0.29
	2	hb-other	36	4.17	7.00	0.44
	2	nh-based	36	4.56	11.11	0.56
1	3	hb-work	14	2.07	0.69	0.22
	3	hb-shop	12	2.00	1.45	0.35
	3	hb-school	7	1.14	0.14	0.14
	3	hb-other	15	3.67	8.10	0.73
	3	nh-based	16	5.00	17.60	1.05
1	4	hb-work	9	2.56	1.28	0.38
	4	hb-shop	5	2.60	3.80	0.87
	4	hb-school	3*	2.00	0.00	0.00

	4	hb-other	13	5.54	6.94	0.73
	4	nh-based	8	3.25	3.93	0.7
2	2	hb-work	33	2.91	3.40	0.32
	2	hb-shop	28	2.29	1.25	0.21
	2	hb-school	5	2.40	1.30	0.51
	2	hb-other	35	3.60	5.42	0.39
	2	nh-based	39	3.13	2.69	0.26
2	3	hb-work	8	3.13	4.70	0.77
	3	hb-shop	6	2.50	1.10	0.43
	3	hb-school	1	2.00	ND	ND
	3	hb-other	9	3.78	5.94	0.81
	3	nh-based	8	4.38	7.13	0.94
2	4	hb-work	18	2.78	3.12	0.42
	4	hb-shop	20	3.20	3.85	0.44
	4	hb-school	6	3.33	6.27	1.02
	4	hb-other	24	6.71	25.35	1.03
	4	nh-based	24	5.71	15.87	0.81
3	3	hb-work	2*	4.50	24.50	3.50
	3	hb-shop	0	ND	ND	ND
	3	hb-school	0	ND	ND	ND
	3	hb-other	2	4.00	18.00	3.00
	3	nh-based	2	1.50	0.50	0.50
3	4	hb-work	6	2.50	1.50	0.50
	4	hb-shop	4	2.25	1.58	0.63
	4	hb-school	0	ND	ND	ND
	4	hb-other	5	7.20	27.70	2.35
	4	nh-based	4	2.75	2.25	0.75
4	4	hb-work	3	2.33	2.33	0.88
	4	hb-shop	2	1.50	0.50	0.50
	4	hb-school	0	ND	ND	ND
	4	hb-other	2*	3.50	0.50	0.50
	4	nh-based	2	4.50	12.5	2.50

*standout categories; ND: no data;

A regression tree for the previous three-way classification is shown in Table 9. The response variable is the number of trips. The independent variables are: number of workers per household (five categories), household size (four categories) and trip purpose (five categories). Of the 100 possible category combinations only 70 are possible, but 5 categories have missing observations on the trip purpose variable. Thus only 65 categories are included in the analysis.

Table 9. Regression Tree for the Three-way Classification

Regression Tree Terminal Node	Number of Workers per Household Categories	Household Size Categories	Trip Purpose* Categories	Number of Groups in Terminal Node**	Number of Households	Mean Trip Rate per Household
1	0, 1, 4	1, 2, 3, 4+	1, 2, 3	336	235	2.91
2	2, 3	1, 2, 3, 4+	1, 2, 3	137	89	4.28
3	0, 1	1	4, 5	192	119	4.26
4	0, 2, 3	2, 3	5	84	84	3.25
5	0, 2, 3	2, 3	4	83	83	3.98
6	1	2, 3	4, 5	103	58	7.74
7	0, 1, 3, 4	4+	5	17	17	3.06
8	0, 1, 4	4+	4	18	18	5.11
9	3	4+	4	5	5	7.20
10	2	4+	4, 5	48	26	11.46

Residual mean deviance = 4.38. *Trip purpose: 1=home-based work; 2=home-based shopping; 3=home-based school; 4=home-based other; 5=non-home based. **The term group is meant to denote category combinations.

A re-expression of the results in Table 9 is shown in Table 10. The number shown in each category combination is the trip rate corresponding to the respective regression tree terminal node (at which point further group splitting stops) in Table 9.

Table 10. Trip Rates by Household Size, Number of Workers and Trip Purpose (after CART Analysis)

Workers	Trip Purpose	Household Size			
		1	2	3	4+
0	hb-work	2.91			
	hb-shop				
	hb-school				
	hb-other	4.26	3.98		5.11
	nh-based		3.25		3.06
1	hb-work	2.91			
	hb-shop				
	hb-school				
	hb-other	4.26	7.74		5.11
	nh-based				3.06
2	hb-work	-	4.28		
	hb-shop	-			
	hb-school	-			
	hb-other	-	3.98		11.46
	nh-based	-	3.25		
3	hb-work	-	-	4.28	
	hb-shop	-	-		
	hb-school	-	-		
	hb-other	-	-	3.98	7.20
	nh-based	-	-	3.25	3.06
4	hb-work	-	-	-	2.91
	hb-shop	-	-	-	
	hb-school	-	-	-	
	hb-other	-	-	-	5.11
	nh-based	-	-	-	3.06

-: indicates category not possible in this classification.

A common strategy for building trees is to fit one with a large numbers of nodes and then prune the tree according to some optimality or complexity criteria. Pruned trees typically do not predict the values in the observed data as well as larger ones, but they may be more robust to overfitting than larger ones. The CART algorithm treats pruning as a tradeoff between two issues: getting the right size of a tree and getting accurate estimates of the true probabilities of misclassification. This process is known as minimal cost complexity pruning.

For moderate size samples (of order 1000), the above method can be used in combination with cross validation (Lachenbruch and Mickey, 1968). The idea is that, instead of using one sample (training data) to build a tree and another sample (pruning data) to test the tree, one can form several pseudo-independent samples from the original sample and use these to form a more accurate estimate of the classification error. A ten fold cross-validation is recommended for the CART algorithm. This is done by holding out 10% of the data, fit a tree to the other 90% of the data, and dropping through the tree the held-out data. While doing so, we note at what level the tree gives the best results. Then we hold out a different 10% and repeat.

The fully-grown tree of the previous three-way classification would have 40 terminal nodes. After pruning, the tree kept would have 10-12 terminal nodes accounting for about 50% of the total variation in the number of trips (Figure 1). The tree with 10 terminal nodes is shown in Tables 9 and 10. Little gain in residual deviance would have resulted had we chosen to retain more terminal nodes by continuing group splitting.

It is obvious from Tables 9 and 10 that home-based work, shopping and school trips cluster differently from home-based other and non-home based trips independently of family size. Perhaps this is not unexpected since the second group is comprised of trips seemingly more discretionary in nature than the first group. In addition, single-member households are making such discretionary trips at different rates than two- and three-member families, as well as larger families.

It would be interesting to investigate, in future work, whether these observations reflect local travel behavior or can be corroborated in larger, perhaps national-scale surveys. Clearly, if the former is true, it could limit the potential for transferability of household trip generation rates. On the contrary, if the latter proves to be true, then what a relief for the local transportation planners!

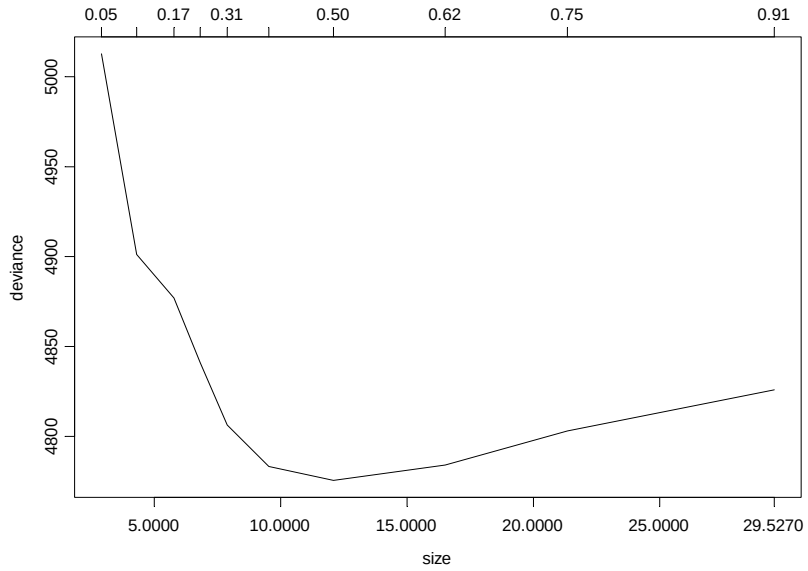

Figure 1. 10-fold cross validation of regression tree

In numerous studies, vehicle availability has been reported to be an important predictor in trip generation. Intrigued by the small samples in Table 2, we decided to conduct a CART analysis to obtain more reliable trip generation rates for this particular three-way classification. The 2002 travel survey (Morocoima-Black and Kang, 2003) collected data on vehicles available for use by members of the household. Four levels of vehicle availability are reported: none, one, two, three-or-more vehicles in the household.

Results from the CART analysis of Table 2 are shown in Tables 11 and 12. The number in each category combination of Table 12 is the trip rate for each category combination corresponding to the respective regression tree terminal node in Table 11. Tree growing and pruning followed the same rules as previously discussed. The 10 terminal nodes in Tables 11 and 12 captured more than 70% of the total deviance. Again, very little additional deviance reduction was achieved with additional tree growing.

Table 11. Regression Tree Analysis of Table 2

Regression Tree Terminal Node	Number of Workers per Household Categories	Household Size Categories	Vehicle Availability Categories	Number of Households	Mean Trip Rate per Household
1	0, 1	1	0, 1, 2, 3+	140	6.16
2	0	2	0, 1	14	10.14
3	0	2	2, 3+	26	8.73
4	1, 2	2	0, 1, 2, 3+	91	9.89
5	0, 1	3, 4+	0, 1	23	9.09
6	0, 1	3	2, 3+	9	14.89
7	0, 1	4+	2, 3+	7	8.86
8	2, 3	3	0, 1, 2, 3+	13	10.85
9	2	4+	0, 1, 2, 3+	28	15.46
10	3, 4	4+	0, 1, 2, 3+	9	11.44

Table 12. Re-expression of Table 2 (after CART analysis)

Workers	Vehicle Availability	Household Size			
		1	2	3	4+
0	0	6.16	10.14	9.09	
	1				
	2		8.73	14.89	8.86
	3+				
1	0		9.89	9.09	
	1				
	2			14.89	8.86
	3+				
2	0	-		10.85	15.46
	1	-			
	2	-			
	3+	-			
3	0	-	-		11.44
	1	-	-		
	2	-	-		
	3+	-	-		
4	0	-	-	-	
	1	-	-	-	
	2	-	-	-	
	3+	-	-	-	

-: indicates category not possible in this classification.

Tables 11 and 12 show that vehicle availability influences the trip making behavior of two-member, and larger-size households. In those cases, the trip rates are different for one-car or less households as compared to multiple-car households. This is an important finding since it is uncommon in traditional trip generation models to uncover such distinctive behavior.

From a practical standpoint, it is interesting to contrast the situations in Table 2 and Tables 11 and 12. Table 2 shows 56 category combinations of which 18 (flagged with 'ND' in Table 2) had no samples. After the CART analysis, each category combination has at least 7 households to draw trip rates from. This is one way to boost small-sample household travel surveys reliability.

6. ROW-COLUMN DECOMPOSITION ANALYSIS AS AN IMPUTATION METHOD

The third problem addressed in this paper is the issue with missing (no samples) information or very little (very few samples) in small-scale household travel surveys. We propose to ameliorate such issues using row-column decomposition analysis as an imputation method.

Row-column decomposition analysis is a simple procedure for the analysis of travel data and its implementation has become trivial with today's spreadsheet functionality. The method has been utilized for trip generation (as elementary analysis in Sen and Johnson, 1977), trip distribution (as elementary analysis in Sööt and Sen, 1978), and mode split (as elementary analysis in Sen and Johnson, 1977). Tukey (1977) has discussed such procedures under the name of row-PLUS-column fit.

Row-column decomposition analysis of Table 1 is carried out in two steps shown in Tables 13 and 14, respectively. Table 13 shows the first step of a row-column decomposition analysis. The means for each column have been computed and subtracted from the original cell values of Table 1. The means themselves, called column fits, are written in the bottom of Table 13.

Table 13. Step 1: Column Means Subtracted

Workers	Household Size			
	1	2	3	4+
0	0.10	-0.51	3.81	-3.00
1	-0.10	0.69	-1.25	-1.94
2	U*	-0.17	-1.38	4.46
3	U	U	-1.19	0.83
4	U	U	U	-0.34
Column Fit	6.18	9.73	12.19	11.00

*U - unavailable

Examples

Column fit = means of column; e.g., 6.19=(6.28+6.09)/2 (from Table 1)
Cell value = observed value – column fit; e.g., 0.10=6.28-6.18

Table 14 completes the row-column decomposition analysis started in Table 13. Now the means of the rows in Table 13 are subtracted – the means themselves being noted in the right-hand margin as row effects. The grand mean is computed as the mean of the column fits; the values of the column fit less than the grand mean are shown in the bottom row of Table 14 as column effects. The numbers left in the cells of the cross tabulations in Table 14 are the residuals.

Table 14. Step 2: Row Means Subtracted

Workers	Household Size				Row Effects
	1	2	3	4+	
0	-0.00	-0.61	3.71	-3.10	0.10
1	0.55	1.34	-0.60	-1.29	-0.65
2	U	-1.14	-2.35	3.49	0.97
3	U	U	-1.01	1.01	-0.18
4	U	U	U	0	-0.34
Column Effects	-3.60	-0.04	2.41	1.23	Grand Mean 9.78

Examples

Row effect = mean of step one row; e.g., 0.10=(0.10-0.51+3.81-3.00)/4
Residuals = cell value of step one – row effect; e.g., -0.00=0.10-0.10
Grand mean = mean of column fits; e.g., 9.78=(6.18+9.73+12.19+11.00)/4
Column effect = column fit – grand mean; e.g., -3.60=6.18-9.78
Original cell value = grand mean + row effect + column effect + residual; e.g., 6.28=9.78+0.10-3.60-0.00

Table 14 can be used to reconstruct Table 1 by the use of the formula

$$y_{ij} = \mu + a_i + b_j + r_{ij} \quad (4)$$

where, y_{ij} is the original observation in the *i*th row and *j*th column of Table 1; μ is the grand mean; a_i is the effect of the *i*th row; b_j is the effect of the *j*th column; and r_{ij} is the residual in the *i*th row and *j*th column.

Any decomposition of the y_{ij}'s in the form of Equation (4) is a row-column decomposition analysis (Sen and Johnson, 1977). Tukey (1977) discusses further details on various forms of row-column decomposition analysis. He also discusses the types of transformations that may be required to achieve the type of additivity implied by Equation (4). Sen and Johnson (1977) argue that row-column decomposition analysis is a regression technique and that row-column decomposition analysis by means is a least- squares technique. The method does not lend itself for formal hypothesis testing, but as a method for simple model building compares well with traditional methods used in the analysis of travel data, such as regression, category analysis and analysis of variance (Sen and Johnson, 1977).

As we see in Table 14, (number of workers, household size) categories (0, 3) and (0, 4+) with 3 observations each (Table 1) have high residuals. In category (0, 3), two of the three households were retired families with high incomes, and the third one was a family with several children headed by a student with a part-time job. In the (0, 4+) category, of the three families surveyed one was a retired high-income family and two families with several kids headed by students with no jobs (international students).

Category (2, 4+) also has a high residual value. This is also corroborated in Table 3 where several cases in this category were identified as unusual. In this category we found families with two kids that are either overly or minimally active regarding trip making. Such cases would need further scrutiny as discussed earlier.

We repeated the analysis several times after removing unusual cases reported in Table 3 and observed some improvement regarding the size of the residuals, but not at the desirable level. Perhaps some level of re-expression for the dependent variable (trip rate) is needed to achieve the type of additivity implied by Equation (4) as Tukey (1977) suggested. Note that all the independent variables in this case are categorical (0/1 variables) and any transformation would be meaningless for them. After a logarithmic transformation of the trip rates in Table 1 (we did not remove any unusual observations) we performed a row-column decomposition analysis in Table 15.

Table 15. Row-column decomposition analysis
(Logarithmic Transformation of Trip Rates in Table 1)

Workers	Household Size				
	1	2	3	4+	
0	1.84	2.22	2.77	2.08	
1	1.81	2.34	2.39	2.20	
2	U*	2.26	2.38	2.74	
3	U	U	2.40	2.47	
4	U	U	U	2.37	
Step 1: Column Means Subtracted					
Workers	Household Size				
	1	2	3	4+	
0	0.02	-0.05	0.29	-0.29	
1	-0.02	0.07	-0.09	-0.17	
2	U	-0.02	-0.11	0.37	
3	U	U	-0.09	0.10	
4	U	U	U	-0.01	
Column Fit	1.82	2.27	2.49	2.37	
Step 2: Row Means Subtracted					
Workers	Household Size				Row
	1	2	3	4+	Effects
0	0.03	-0.04	0.30	-0.28	-0.01
1	0.04	0.12	-0.04	-0.12	-0.05
2	U	-0.10	-0.19	0.29	0.08
3	U	U	0.09	0.09	0.01
4	U	U	U	0.00	-0.01
Column Effects	-0.42	0.04	0.25	0.13	Grand Mean 2.24

*U – unavailable

The residuals of the logarithmic transformation in Table 15 show a substantially better fit than the raw values would yield. Notice that once the rather strong effect of household size has been accounted for, the effect of the number of workers is small. Since we transformed the dependent variable (trip rate) it would be only prudent to recheck about heteroscedasticity and perhaps continue the search for a better transformation and repeat the previous steps.

Row-column decomposition analysis can also be applied to *n*-way tables. For example, a row-column decomposition analysis of Table 2 would treat the three-way classification (number of workers, family size, and auto availability) as a two-way table (e.g., [number of workers, auto availability], family size) and proceed with the rest of the steps as described above. Similarly, a row-column decomposition analysis of a four-way table could be translated into a 'two-way' analysis using as rows the category combinations of three of the four variables. Additional dimensions could be accommodated in an analogous fashion.

To our knowledge, imputation in trip generation is either uncommon or seldom discussed. As we have seen, opportunities for imputation arise naturally in trip generation modeling and

techniques developed for missing data and contingency table imputation could find fertile ground here. For example, Brownstone (1998) showed how Rubin's (1987) multiple imputation methodology could help alleviate problems caused by survey non-response and missing data. Wang and Shao (2003) used hot-deck imputation for missing data in a two-way contingency table, while Cox (2002) used linear programming and Markov Chain Monte Carlo methods for the same problem.

On the other hand, research into the use of CART-type methods for imputation has not been conclusive and perhaps other algorithmic approaches appear to be more promising (Michie *et al., 1994*). Reiter (2003) reports that sequential CART models have promise as a method for generating partially synthetic data sets. He also reports that the primary drawback of the approach is the sequential nature of the imputations, which can introduce conditional independence structures into the released data. This issue also affects the use of CART models, or any sequential imputation scheme, for imputation of missing data.

In addition, Wilmot and Shivananjappa (2003) tested the accuracy of CART-type and neural network imputation procedures on a sample of households from the 1995 Nationwide Personal Transportation Survey. The analysis produced mixed results for the variables tested (household income, number of vehicles, educational status, and age).

Meanwhile, the simplicity in implementation of the model in Equation (4) shows that row-column decomposition analysis could be used for imputation in cases where one or more independent variables have missing information or in cases where the number of observations is small. Indeed once the residuals have been minimized and a satisfactory model has been obtained, Equation (4) can be used to impute the missing cell value by the value of the grand mean augmented by the values of the respective row and column effects. We believe that the method compares favorably (in terms of cost-effectiveness) against other more complicated imputation approaches and could become a valuable tool especially among transportation modelers in smaller MPOs.

7. CONCLUSIONS

The analyst of small-scale household travel surveys is facing a number of issues when estimating or validating trip-generation rates as an input to the trip-generation modeling process. Such issues include data quality screening, reliability of trip rates in the presence of small samples, and imputation issues. This paper has proposed methods to tackle each of these tasks. The methods are fairly easy to implement in a cost-effective manner.

REFERENCES

Breiman, L., J.H. Friedman, R.A. Olshen, and C.J. Stone (1984). *Classification and Regression Trees (CART).* Pacific Grove, CA: Wadsworth.

Brownstone, D. (1998). Multiple Imputation Methodology for Missing Data, Non-Random Response and Panel Attrition. *Theoretical Foundations of Travel Choice Modeling,* pp. 421-449. University of California Transportation Center. http://www.uctc.net/papers/594.pdf

Clarke, L.A. and D. Pregibon (1992). Tree-based Models. In *Statistical Models in S,* eds. J.M. Chambers and T.J. Hastie, pp. 377-419. Pacific Grove, CA: Wadsworth & Brooks/Cole.

Champaign Urbana Urbanized Area Transportation Study (2004). *Long Range Transportation Plan 2025*. December, 2004.

Cox, L.H. (2002). Imputing Missing Values in Two-Way Contingency Tables Using Linear Programming and Markov Chain Monte Carlo. UNECE Work Session on Statistical Data Editing, Working paper no. 39.

Friedman, J.H. (1991). Multivariate Adaptive Regression Splines (with discussion). *The Annals of Statistics 19,* pp. 1-141.

Lachenbruch, P. and R. Mickey (1968). Estimation of Error Rates in Discriminant Analysis. *Technometrics,* 10:1–11.

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (eds.) (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood Series in Artificial Intelligence. Ellis Horwood.

Morocoima-Black, R. and E. Kang (2003). *2002 Champaign-Urbana-Savoy Travel Survey.* Final Report. CUUATS, July 1, 2003.

Reiter, J.P. (2003). Using CART to Generate Partially Synthetic, Public Use Microdata. *Proceedings of the Federal Committee on Statistical Methodology Research Conference, 2003.*

Ripley, B.D. (1996). *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Sen, A. and C. Johnson (1977). A Simple Method for Analysis of Travel Data. *Transportation Research,* Vol. 11, pp. 189-196.

Sen, A. and M. Srivastava (1990). *Regression Analysis: Theory, Methods and Applications.* Springer-Verlag, New York.

Sööt, S. and A. Sen, (1978). Elementary Analysis in Trip Distribution Modeling. *Transportation Engineering Journal,* ASCE, Vol. 104, No. TE6, Proc. Paper 14114, November, 1978, pp. 789-797.

Thakuriah, P., A. Sen, S. Sööt and E. Christopher (1993). Non-Response Bias and Trip Generation Models. *Transportation Research Record 1412*, pp. 64-70.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Massachusetts.

Venables, W.N., and B.D. Ripley (1997). *Modern Applied Statistics with S-Plus.* New York, NY: Springer-Verlag.

Wang, H and J. Shao (2003). Two-Way Contingency Tables Under Conditional Hot Deck Imputation. *Statistica Sinica 13*, 613-623.

Wilmot, C.G. and Shivananjappa, S. (2003). Comparison of Hot-Deck and Neural-Network Imputation. In: *Transport Survey Quality And Innovation*, pp. 543-554. Elsevier.