SIMPLE SYNTHETIC POPULATIONS WITHOUT THE USE OF RANDOM DRAWS

Vincent L. Bernardin, Jr., Ph.D.
Bernardin, Lochmueller & Associates, Inc.
6200 Vogel Road
Evansville, IN 47715
VBernardin2@BLAinc.com
Ph: 812-479-6200
Fax: 812-479-6262

Submitted: October 25, 2009
2,135 Words, 1 Table, 1 Figure

Abstract

This short paper presents the methodology employed in the hybrid travel model developed for Knoxville, Tennessee, for producing synthetic populations of households without introducing simulation variation through the use of random draws. The avoidance of simulation variation is an important advantage for practical forecasting tools since it can greatly reduce the amount of computing time necessary for many practical applications, such as alternatives analyses. The approach adopted here has its own distinct limitations, mostly related to its sensitivity to the dimensionality or number of characteristics of the population represented. It may not be appropriate for extremely detailed modeling of populations, but it may, nonetheless, be very valuable as a practical tool for reducing aggregation bias in travel forecasting models.

Introduction

In recent years there has been a shift away from the application of demand models directly to traffic analysis zones in favor of representing individual households (and sometimes persons) and modeling travel behavior at their level. The shift is driven by the basic fact that people travel, not zones. More precisely, the shift is to avoid the aggregation bias that occurs when non-linear demand models (such as logit models) are applied to aggregate or average characteristics rather than to populations with a range of attributes around their group averages. For example, a mode choice model may predict no significant transit mode share when applied to a zone with 100 households with an average of 2.2 cars per household. However, the same mode choice model, applied to the same 100 households individually, may predict a significant number of transit trips if 5 of the households have no vehicles. Examples like this illustrate that the effects of aggregation bias can be quite significant, not only in the statistical sense, but practically, and have helped motivate the shift to modeling disaggregate synthetic populations.

Aggregation inevitably results in the loss of information, and aggregation bias in travel models has been a concern and received attention in the form of academic research since the 1970s (Koppelman, 1974). Several sorts of aggregation are typically present in traditional trip-based travel models including spatial, demographic and temporal aggregation. Demographic aggregation, which motivates the use of synthetic populations, refers to the application of behavioral models to the entire set of travelers within a zone rather than individual travelers, implicitly treating all travelers within a zone as though they were the same, representing distributions of demographic variables only by their means. Although demographic aggregation and the resulting information loss is typically avoided in model estimation, this causes an inconsistency in the estimation and application of component models and results in biases due to the nonlinear nature of these models.

Many four-step models attempt to avoid aggregation bias simply by omitting many demographic variables and their effects, but this simply creates a different problem by making the models insensitive to important characteristics of the traveling population and, possibly, introducing a variety of unknown biases from specification errors. Some traditional models, however, do partially reduce, if not eliminate, the problem of demographic aggregation through the use of market segmentation.

All but the original activity-based model, instead, avoid demographic aggregation by modeling travel at the level of individual travelers. This avoids aggregation bias but has typically involved Monte Carlo simulation which introduces simulation error and ultimately dramatically increases the computational cost of producing average results. (The same approach is generally used throughout activity-based models, not only in population synthesis.) The use of random draws reduces the computational burden of a single model application, but makes the results of the model application variable so that the model must be run multiple times to produce an average result which can be used, for instance, for purposes of making comparisons.

The practical consequences of this are not insignificant. A single application of an activity-based model requires generally on the order of a day to complete, even on machines with a dozen or more

processors (1). The limited research on simulation variation in activity-based models indicates that the number of runs necessary to produce a reasonable confidence interval for facility/corridor specific forecasts may be relatively small, perhaps as few as ten runs (2,3). However, this still means that to compare just two alternative corridor alignments would take approximately 20 days of computing time, even using multi-processor machines.

It has occasionally been suggested that fixing the random seed may be used as a sort of short-cut or quick solution to avoid the need to make multiple runs. However, while fixing the random seed may allow users to reproduce results, it does nothing to solve the inherent problem in simulation modeling or produce an unbiased, average result. A fixed random seed produces not an unbiased comparison between alternatives, but rather, simply a stable and repeatable biased comparison. This is illustrated in that a different fixed random seed would produce a different resulting comparison.

The hybrid travel model recently developed for Knoxville, Tennessee, and a similar one now being developed for Evansville, Indiana, have attempted to reduce the aggregation bias of traditional models while avoiding the simulation variation of activity-based models (4,5). The result is a compromise in several regards. The hybrid models do avoid simulation variation, but are only able to avoid demographic aggregation bias in some of the model components such as tour and stop generation and tour mode choice while other components such as stop sequence choice and departure time choice remain aggregate.

The remainder of this paper focuses on the population synthesizer which allows the successful disaggregation and avoidance of random draws in the disaggregate portion of Knoxville's hybrid model. Although the approach could be extended to deal with individual travelers, the Knoxville model generates only a synthetic population of household.

Knoxville's Population Synthesizer

For each traffic analysis zone, individual households are created based on the demographic information associated with that zone. Each household has a total number of persons, a number of workers and of students, a binary variable indicating whether or not any of the household members is over the age of 65 and an income variable that indicates whether the household belongs to the lower (under $25,000/year), middle ($25,000 - $50,000/year) or upper (over $50,000/year) income category, each of which comprises approximately a third of the households in the Knoxville region. The number of vehicles available to each household is modeled separately, after the population synthesis, based on these and other variables.

Primary Inputs

- Zonal Average Household Size
- Zonal Average Workers per Household
- Zonal Average Students per Household
- Zonal Percent of Households w/ Senior
- Zonal Average Household Income

Secondary Inputs

- Population Density
- Percent of Zone within .5 mi of Bus
- Urban Design Factor

Output

Synthetic households for each TAZ with

- Number of persons (1-5+)
- Number of workers (0-3+)
- Number of students (0-2+)
- Presence of seniors (0, 1)
- Income Group (low, mid, high)

The synthetic population is developed in two steps. First, a set of ordered response logit models with shadow prices predict for each variable (such as household size, number of workers, etc.) the number of households which have each level of that variable (one person, two persons, etc., zero workers, one worker, two workers, etc.). Second, iterative proportional fitting is used to develop the synthetic population based on a seed population of households and the marginal distributions for each variable provided by the logit models.

The process is able to avoid simulation by allowing the production of more or less individual households than exist in the real population, creating consistency instead by weighting the households so that their weighted sum is the total actual number of households in each zone. Rather than sample from the distribution produced by iterative proportional fitting, using random draws, the distribution itself is used as the synthetic population simply converting the probabilities for each household type to the number of households of that type by multiplying by the number of households in the zone.

It follows that the limitation of this approach is that the size of the synthetic population, measured by the number of household types (regardless of whether there are ten or 0.2 households of that type), grows quickly with the number of characteristics which describe the households. However, the problem is not as restrictive as might initially be thought. If all combinations of number of persons, workers, students, presence of seniors and income group were possible, there would be 360 household types for each zone. However, many of these types, such has single person households with two or more workers, are not possible. The observed seed distribution of households actually contained only 157 different household types. Hence, the issue of size for synthetic populations of this type is not insignificant, but neither is it generally prohibitive.

Ordered Response Logit Models of Marginal Distributions

Aggregate ordered response logit (ORL) models were developed to model the discrete distributions of each household characteristic variable noted above. These models essentially replace the stratification curves used in many traditional travel models to cross-classify households for trip generation. The models are fairly simple, largely driven by the aggregate zonal average variable describing the distribution which they represent (e.g., the model which determines the number of households with

zero, one, two or three or more workers is driven largely by the zonal average number of workers per household).

The models also include some other, secondary demographic variables which are related to the distributions of the primary variable as well. For instance, even for a given average number of students per household for a zone, the number of zero student households is generally greater in zones with more households with seniors (age 65 and older); whereas, in contrast, the zero student households tend to decrease with the zone's average income, all other things being equal.

Ordered response logit models are a special form of nested logit models designed to accommodate the correlation pattern typical of ordinal data, such as the number of persons, workers, etc., in a household. They were tested against simpler multinomial logit models which assume independence across alternative categories, and in each case, the ordered response model provided better goodness-of-fit to the observed data. ELM software (www.elm-works.com) was used for all logit model estimation.

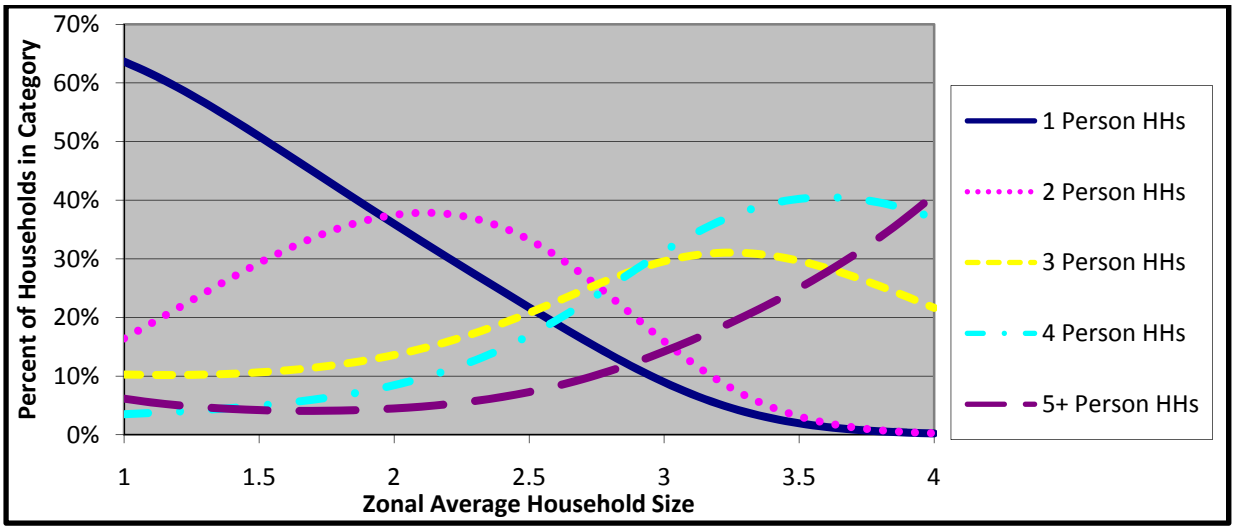

Figure 1. Percent of Households by Number of Persons vs. Zonal Average Household Size (before shadow prices)

The model parameters, t-statistics and goodness-of-fit measures for the number of persons variable are shown in Table 1. It is not unusual or unexpected that the goodness-of-fit is low for models of this type, which is attempt to explain disaggregate phenomena based on aggregate variables. However, a reasonable level of confidence can still be had in the synthetic populations which they produce since they are both constrained to agree with zonal average characteristics (through the use of shadow prices) and only applied to factor the observed seed distribution in the subsequent round of iterative proportional fitting. The implied distribution of households (assuming regional average secondary zonal demographic characteristics) before the application of shadow prices are shown in Figure 1. While the need for the shadow prices is evident for extreme zonal averages, the distributions are clearly reasonable in general.

Table 1. Aggregate Ordered Response Logit Model for Household Size

Household Size	Alternative	Parameter	t-statistic
-- Logsum Parameters			
Nest_1	alt_2, Nest_2	0.9	Constrained
Nest_2	alt_3, Nest_3	0.8	Constrained
Nest_3	alt_4, alt_5	0.7	Constrained
-- Alternative Specific Parameters			
CONSTANT	alt_1	1.4991	1.15
CONSTANT	alt_2	-4.2750	-2.18
CONSTANT	alt_3	-0.4124	-0.29
CONSTANT	alt_4	-1.9605	-1.35
Zonal Average Household Size	alt_1	2.5378	2.05
Zonal Average Household Size	alt_2	4.9789	2.96
Zonal Average Household Size	alt_3	1.5143	1.26
Zonal Average Household Size	alt_4	1.9344	1.58
Zonal Average Household Size, Squared	alt_1	-0.9999	-3.55
Zonal Average Household Size, Squared	alt_2	-1.3571	-3.70
Zonal Average Household Size, Squared	alt_3	-0.3655	-1.39
Zonal Average Household Size, Squared	alt_4	-0.3655	Constrained
Population Density	alt_1	0.0581	2.07
Log of Zonal Average HH Income	alt_1	-0.3076	-2.41
Log of Zonal Average HH Income	alt_2	0.3827	3.43
Percent of Households with Senior	alt_3	-1.5443	-2.62
-- Model Statistics	statistic		
Log Likelihood at Zero	-4730.5		
Log Likelihood at Constants	-4363.7		
Log Likelihood at Convergence	-4229.1		
Rho Squared w.r.t. Zero	0.106		
Rho Squared w.r.t Constants	0.031		

Shadow Prices

To insure consistency with zonal averages, the models also include “shadow prices” which guarantee the average characteristics of the synthetic population will agree with averages for each zone. The concept of shadow prices is taken from economics and optimization science. Technically, they are simply lagrangian multipliers associated with constraints in an optimization problem, in this case, constraints that the observed zonal averages be reproduced.

Conceptually, consider the situation in which the basic relationship between the demand for some good and its price is known (from various observations), and yet, for some (other) observation or observations, the observed demand is lower than what is predicted based on the known relationship with its price. One way this situation can be addressed, if there is confidence in the basic demand function and the contrary observations, is that an additional, unobserved “shadow price” in addition to the observed price can be postulated to account for the observed demand. This shadow price becomes an additive correction term in the demand function.

In these models, the shadow prices are developed iteratively, where the formula for the shadow prices added to the utility function of alternatives less than the true zonal average is given:

$$s_i = s_{i-1} + (TrueAvg - AltAvg)\ln\left(EstAvg_{i-1}/TrueAvg\right)$$

or for alternatives greater than the true zonal average:

$$s_i = s_{i-1} + (TrueAvg - AltAvg)\ln\left(TrueAvg/EstAvg_{i-1}\right)$$

where *TrueAvg* is the zonal average from the TAZ geographic layer, $EstAvg_{i-1}$ is the resulting zonal average in iteration *i-1*, and *AltAvg* is the average for that alternative (generally equal to the alternative number, except for the last category, e.g., 5+ persons, 3+ workers, etc.).

Iterative Proportional Fitting

The synthesis of the population is completed using traditional iterative proportional fitting in multiple dimensions. The inputs to the iterative proportional fitting procedure are the marginal distributions produced by the ordered response logit models and a seed or sample population of households and persons. The combined sample from the 2000 and 2008 household surveys, properly weighted, is used for this purpose. The use of the household survey sample as a seed distribution for iterative proportional fitting offers consistency with the models of the marginal distributions which were estimated from the same data and helps ensure convergence.

The use of shadow prices in the generation of the marginal distributions guarantees that the synthetic population created by iterative proportional fitting will agree with the TAZ layer not only on the number of households, but also the number of persons, workers, students and households with seniors in each zone.

References

1. Rossi, T., B. Winkler, T. Ryan, K. Faussett, Y. Li, D. Wittl, M.A. Zeid. Deciding on Moving to Activity-Based Models (or Not). Presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
2. Castiglione, J., J. Feedman and M. Bradley. 2003. Systematic Investigation of Variability due to Random Error in an Activity-Based Microsimulation Forecasting Model. In *Transportation Research Record* 1831, TRB, National Research Council, Washington, D.C., pp. 76–88.
3. Veldhuisen, J., H. Timmermans, and L. Kapoen. 2000. Microsimulation Model of Activity-Travel Patterns and Traffic Flows: Specification, Validation Tests, and Monte Carlo Error. *Transportation Research Record*, No. 1706, TRB, Washington, D.C., pp.126–135.
4. Bernardin, V. and M. Conger. 2010. From Academia to Application: Results from the Calibration and Validation of the First Hybrid Accessibility-based Model. Presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C., 2010.
5. Bernardin, V. An Accessibility-Based Approach to Travel Demand Forecasting: A New Alternative to Four-Step and Activity-Based Methods. Presented at the 87th Annual Meeting of the Transportation Research Board, Washington, D.C., 2008.