A COMPARISON OF POPULATION SYNTHESIZERS USED IN MICROSIMULATION MODELS OF ACTIVITY AND TRAVEL DEMAND

JOHN L. BOWMAN, PH. D.
5 Beals Street, Apartment 3
Brookline, MA 02446, USA
T: (617) 232-3478 * E-mail:john_l_bowman@alum.mit.edu

Abstract

Microsimulation models that forecast the activities and travel of urban populations create synthetic populations and then use them to simulate the behavior of the households and persons in that synthetic population. The features of eight population synthesizers are compared, and suggestions are made for incorporating the best features into future population synthesizers.

Author’s note

The author applauds the fine work of the developers of all the population synthesizers reviewed in this paper, and thanks them for their assistance by providing documentation and, in some cases, corrections. This is a working paper; not all corrections have yet been incorporated.

INTRODUCTION

Disaggregate land use and activity-based travel models, including microsimulations, represent the decisions and actions of individual persons and households. They first create a synthetic population (synpop), then predict outcomes for each member of the population, and finally aggregate the results for policy analysis. If the disaggregate model predicts only short-term activity and travel, then it creates a base year synthetic population, typically from census data, and uses aggregate land use forecasts to create a synthetic population for each forecast year. If it also predicts land use, then it creates only a base year synthetic population and predicts changes in the population for forecast years.

This paper compares eight population synthesizers (popsyns), which are introduced in Table 1. The Table 1 names are used herein to refer to either a population synthesizer, its model system, or an agency, depending on the context. TRANSIMS, under development since 1994, generates a synthetic population to send through a regional traffic microsimulator, and was recently augmented to create forecast year synthetic populations. SFCTA is currently used in an activity-based travel demand microsimulator by SFCTA for forecasting and policy analysis. Metro was used in an activity-based model system with sample enumeration, and also with demand microsimulation. TRESIS uses sample enumeration methods to predict work and residential locations in addition to travel choices, and has been applied in several Australian locations. MORPC is used in a newly implemented tour-based travel demand simulator. Oregon2 creates a base year synthetic population which is subsequently evolved using dynamic models, in a state-wide land use and travel microsimulator currently under development. CEMDAP and ARC are parts of activity-based travel demand microsimulators being developed for regional travel forecasting.

Table 1. The eight compared population synthesizers

Name	Year	Sponsor, client and location	Developer and citation
TRANSIMS	1995	U.S. Government Portland Metro Portland, Oregon	Los Alamos National Laboratories Beckman et al (1996), LANL (2003)
SFCTA	1998	San Francisco County Transit Authority (SFCTA) San Francisco, California	Mark Bradley Bradley et al (1999), Bradley (2003)
Metro	1999	Portland Metro Portland, Oregon	Mark Bradley Bradley (1999, 2003)
TRESIS	2000	Australian Government Sydney and other regions	Institute of Transport Studies (ITS), University of Sydney Ton and Hensher (2003), ITS (2004), Hensher et al (2004)
MORPC	2002	Mid-Ohio Regional Planning Commission (MORPC) Columbus, Ohio	PB Consult PBConsult et al (2003)
Oregon2	2003	Oregon Department of Transportation (ODOT) Oregon state-wide	PBConsult and HBA Specto PBConsult et al (2004)
CEMDAP	2003	Texas Department of Transportation (TxDOT) Dallas/ Fort Worth	Center for Transportation Research (CTR), University of Texas at Austin Bhat et al (2003a, 2003b)
ARC	2003	Atlanta Regional Commission (ARC) Atlanta, Georgia	John L Bowman and PBConsult PBConsult et al (2003)

The comparison focuses on the population synthesizers themselves, rather than the model systems in which they appear, although some of the popsyn differences can be attributed to the different requirements placed on them by their model systems. Several aspects of the popsyns are compared, including their basic approach, how they generate base year and forecast distributions—demographic and geographic—of population characteristics, the procedure for generating households from the distribution, their validation procedures and results, and their software implementation.

BASIC APPROACH

The typical synthesis procedure involves two main steps; first a demographic distribution of households is estimated for each transportation analysis zone (TAZ) or census block group, and then a matching sample of households is drawn from a set of household records for which nearly complete census information is available. This produces a synthetic population in which each synthetic household and its members have many clearly defined characteristics for use in the model system, and together they match the estimated demographic distribution within each TAZ.

The demographic distribution is defined discretely by the cartesian product of several categorical control variables (dimensions), with each multidimensional category (or cell) defined as a unique value combination of the one-dimensional control variables. The number of households in each cell is estimated through an iterative proportional fitting procedure (IPF). The IPF procedure starts with an intitial joint distribution available for aggregate census geographical units—Public Use Microdata Areas (PUMA). It then cycles iteratively through a set of control totals, one total for each category of each control variable. For each control total it adjusts the joint distribution to satisfy the total by proportionally adjusting all cells with attributes matching the control total's category. If the control totals are mutually consistent, then IPF eventually converges so that all control totals are satisfied and the correlation structure of the initial joint distribution is preserved. Control totals are taken from census tables for the base year, and for the forecast years they come from demographic and land use forecasts, which may be less detailed. Once the joint distribution is determined, with each cell indicating the number of households of a certain type residing in the TAZ or block group, then for each cell its number of households is drawn randomly from the corresponding set of household records available in the 5% Census Public Use Microdata Sample (PUMS) of its PUMA.

Most of the popsyns deviate from this typical procedure in one or more ways. TRANSIMS and Oregon2 complete the base year synthesis three times, once each for families, non-family households, and group quarters residents. This makes it easy to use distinct type-appropriate base year controls available in the census for each household type. Then, however, either the forecast controls must also be generated for each type, or else the forecast year IPF must act on a differently defined joint distribution, potentially introducing differences between base and forecast year synthetic populations caused only by the different IPF structures. Oregon2 does not face this problem because it synthesizes only a base year population. It then uses a variety of models to gradually evolve the synthetic population, instead of synthesizing a forecast population.

TRESIS (which uses Australian census data) estimates a 3-dimensional distribution for the entire region (rather than for each TAZ) without employing IPF. It starts with a 2-dimensional 100% census summary table, and subdivides each cell acording to the distribution in the 1% sample of households. This strictly limits the possible number of controlled dimensions, and the accuracy of the third dimension is limited by the 1% sample. Also, since the joint distribution is estimated for the entire region, the popsyn itself removes geographic variation from the demography of the sample; it relies instead on the subsequent residential and work-location choice models to model such variation.

MORPC does not use census data directly for its base year controls. Instead, it replicates the procedure used for generating forecast controls, starting with estimates of zonal population, households, average income and workforce, and then converting these into a categorical distribution in each dimension via functions and tables estimated from base year data for the region. Unfortunately, this reduces geographic variation from the base year, but it makes it easier to compare base and forecast year model results in cases where the forecasts are far less detailed than the base year census data.

CEMDAP estimates the demographic distribution for each census tract (which contains many block groups). It substitutes the entire PUMS of the tract's PUMA for the multidimensional distribution, using the IPF procedure to reweight the PUMS. This allows any variable available in the sample households to be used as a control variable in the IPF procedure, as long as the control values are available. This makes it easy to include control the numbers of persons of various types in addition to controling the numbers of households of various types. Different control variables can be used in the base and forecast years without changing the underlying structure of the joint distribution.

The ARC user specifies a list of categorical variables (dimensions) for which control is needed, and a list of cells comprising the joint distribution. Each cell is defined by the categories it includes in each dimension. Categorization can eliminate infeasible cells and can aggregate to avoid many sparsely populated cells. Categorization is also chosen to match as closely as possible the available base year census controls and forecast input data. The configuration of dimensions and cells can be changed without requiring programming changes, allowing optimization as part of the validation process, and accommodating changes in available forecast input data. ARC also allows simultaneous use of regional control totals (that may come from regional demographic forecasts) and TAZ control totals (that may come from land use forecasts) in the IPF procedure.

ESTIMATING THE BASE YEAR JOINT DISTRIBUTION

In addition to the differences in approach described in the preceding section, the popsyns differ in how they estimate the base year distribution. As shown in Table 2, they differ substantially with respect to the control variables and categories used for the base year IPF. Nearly all the popsyns control for household income, household size and number of workers. Half or more control for age of householder and presence of children. Six other household variables are used by only two or three popsyns, and CEMDAP alone employs four control variables on the number of persons of various types. Most of the controls are one-dimensional, but ARC, with its flexible joint distribution definitions, is able to make heavy use of available 2- and 3-dimensional controls.

Some of the popsyns also differ from the previous section's description of the base year IPF procedure. TRANSIMS, Portland and Oregon2 use a preliminary first stage IPF procedure to adjust the joint distribution of the 5% PUMS so that it matches the census summary table control values at the PUMA level of aggregation. In the main second stage IPF procedure, the PUMA-level joint distribution is enforced as a set of controls along with all the disaggregate controls for each of the tracts (TRANSIMS) or block groups (Oregon2) or TAZs (SFCTA, Portland) in the PUMA. This assures that the resulting tract-level distributions match the PUMA level controls when aggregated.

MORPC also uses a 2-stage IPF procedure, but for a different reason. The first stage estimates a two-dimensional distribution by household size and number of workers. The result is then used as input for estimating the control values of the third dimension—income—so that the income controls used in the 3-dimensional second stage IPF depend on household size and number of workers.

ARC uses yet another 2-stage IPF procedure, again for a different reason. A preliminary first stage IPF reconciles inconsistent controls; the original controls serve as the seed distribution and a set of consistent overriding meta-controls serve as the controls. This assures that the second stage IPF, which uses the adjusted (and now consistent) controls, converges even if the original controls were not mutually consistent. This can happen quite easily, especially if the controls come from different sources. ARC uses base year controls from three different sets of census tables (SF1 100%, SF3 16%, and CTPP).

One other minor difference exists between the base year popsyns. CEMDAP and perhaps other popsyns use unweighted 5% PUMS as the seed distribution for IPF. The resulting seed is somewhat biased because the records in PUMS were not sampled with equal probability. This can be overcome easily by using the census supplied PUMS weights to estimate the seed distribution.

Table 2. Control variables and categories of the base year population synthesizers

		Number of categories for control variables of various types																
			Household variables (# of households by...)											Person variables			
	Control number	Multidimensional controls	Family vs nonfamily	Age of householder	Income	HH size	# workers	Children	Gender of hoseholder	Couple vs not	Race of householder	Institutionalized vs not	Group quarters type	Employed vs not	Age	Race	Hispanic vs not
TRANSIMS and Oregon2																	
family HH	1			7													
	2				7												
	3						4										
	4	60						4	2	2	5						
nonfamily HH	1			7													
	2				7												
	3										5						
	4	4				2			2								
	5	6		3		2											
Group qtrs residents	1	10										2	5				
	2	6		3								2					
SFCTA	1			3													
	2				4												
	3	9				3	4										
Metro	1			4													
	2				4												
	3					4											
TRESIS (no IPF)	1	150			5	2	6	2		2							
MORPC	1				3												
	2					9											
	3						5										
CEMDAP	1				16												
	2						7										
	3		2														
	4													2			
	5														3		
	6															4	
	7																2
ARC	1	7	2	2		2		2									
	2	9	2			5											
	3	8		2	4												
	4	8	2		4												
	5	13				4	4										
	6	16			4	4											
	7	16			4		4										
	8	39			3	4	4										

ESTIMATING THE FORECAST YEAR JOINT DISTRIBUTION

Since the primary purpose of synthetic populations is to support the forecasting, whether and how popsyns synthesize forecast populations is of great importance. Interestingly, here is where major differences among the popsyns—and the systems to which they belong—emerge. Two of the eight popsyns do not synthesize a forecast population. As mentioned earlier, Oregon2 uses models to simulate evolution of the base year population. TRESIS uses models to estimate changes in residential and job location, without adjusting population demographics. The procedures of the other six are varied enough that there is no 'typical' approach. They have all taken different approaches for dealing with the fundamental problem, which is to take maximum advantage of the scarce forecast information available for incorporation into the synthetic population.

Portland, designed from the outset with the awareness of this problem, employs a three-dimensional joint distribution for which the Portland land use model already predicts a joint distribution. Therefore, for the forecast year, IPF is unnecessary; rather, the joint distribution from the land use model can be used directly to draw households (as described in the next section).

As described in previous sections, the base year synthesis procedure of MORPC was designed so that it uses the base year information in the same format as is available (in Columbus) for forecast years. Therefore, the procedure for estimating forecast year joint distribution is identical to the base year procedure.

For each of its base year control variables, SFCTA has available forecasts that can not be used directly. Instead of households by income category, it has population by income category. Instead of househlders in three age categories, it has persons over and under age 62. Instead of households by household size and households by number of workers, it has population and employment. In each case, SFCTA uses a procedure that uses the available information to adjust the related base year control values. It then performs forecast year IPF exactly like the base year IPF, with the adjusted controls instead of the base year controls.

Unlike Portland, MORPC and SFCTA, the base year TRANSIMS popsyn was developed without a specific regional application or a documented design for synthesizing a forecast population. The TRANSIMS procedure for synthesizing a forecast population was developed later, specifically for forecasting in Portland. The TRANSIMS base year joint distribution involves many control variables that differ by household type (family, nonfamily and group quarters). It is quite complex and is substantially different than the available Portland forecast joint distribution described above. TRANSIMS adopts the structure of the Metro joint distribution and control variables, and transforms the base year joint distribution (after base year IPF) to serve as the seed distribution for the forecast year IPF. The transformation relies on the relationship between the two categorization schemes that is embedded in the 5% PUMS, benefitting from the fact that each PUMS household can be precisely located in both of the joint distribution structures. The resulting seed distribution for the forecast IPF is specific to each PUMA; any tract-specific demographic information from the base year is lost. TRANSIMS then uses the marginal distributions of the three metro forecast variables as controls to estimate a forecast distribution that satisfies the Metro controls and maintains the correlation structure of the transformed PUMA-level base year joint distribution.

Although CEMDAP documentation does not refer to a forecast year synthesis procedure, its structure is well suited for synthesizing a forecast population. As mentioned above, its use of the entire PUMS to represent the joint distribution allows any variable available in the sample to be used as a control variable, including estimates of the number of persons of various types (eg, by employment status or age cohort) that are typically available from land use models.

As also mentioned above, the flexible structure of ARC allows a joint distribution categorization that closely matches available base year census controls and forecast input data, and allows simultaneous use of regional control totals and TAZ control totals, both of which are used in the ARC application. However, with the chosen structure for the ARC application, forecasts of population by age and workforce must still be transformed. The seed distribution for forecast year IPF is the output of the base year IPF, and ARC allows the user to control the level of aggregation of the seed.

GENERATING THE HOUSEHOLDS

Once the popsyns have estimated a base or forecast year distribution, they generate synthetic households to match the distribution. This process typically includes four steps. Since the estimated distribution specifies non-integer values for the number of households of each demographic category within a geographic unit, the first step is usually to adjust all those values to integers. Second, a Monte Carlo procedure is employed to draw the correct number of households of each type from the census sample. Third, the useful household and person variables are extracted from the drawn census household and retained for use by the model system. Finally, in some cases, an additional procedure is used to assign each household to a more precise location within its geographic unit.

Integerizing the distribution

The simplest approach, employed by Oregon2, and perhaps by TRANSIMS, skips the task of integerizing the distribution. Instead, the distribution is retained as a discrete probability mass function, and a Monte Carlo draw is made from this distribution for every household in the geographic unit. The drawback of this approach is that in small geographic units the actual number of draws from each demographic category can differ substantially from the expected number of draws. All of the other popsyns avoid this problem by rescaling the distribution so that it sums to the total needed number of housholds, and adjusting every non-integer value to the next integer value up or down. However, they all adjust in different ways.

TRESIS uses a simple rounding procedure, except all non-zero values less than 1 are rounded up to 1. This method is biased for values less than 1, and can yield aggregations that differ substantially from the control values.

CEMDAP uses an unspecified rounding procedure that somehow controls for household size distribution, but not other control values. For each geographic unit, households are then randomly replicated or deleted to exactly match the total number of households implied by the controls.

MORPC rounds up, starting with demographic categories with the largest fractional component, but avoiding rounding up if it would cause a control value to be exceeded. This procedure closely matches control totals, but is biased, and the bias could be material if certain unusual demographic categories have values much less than one in most geographic units.

SFCTA and Metro avoid the bias by using a Monte Carlo draw for every demographic category; denoting the fractional value of the category as p, then it is rounded up with probability p, and rounded down with probability 1-p. As Bradley (2003) notes, “When done across all cells, this procedure may not result in the correct number of total households. So, the process is started over and iterated until the correct total number is achieved. There is no control that the marginals are still perfectly matched across all cells, only the total number of households.”

The design for ARC is to use the adjustment protocol of MORPC, so that control values are satisfied, but to implement it with Monte Carlo draws, so that bias is avoided.

Drawing households

CEMDAP is different than the other popsyns because by using the census sample to represent the joint distribution, no Monte Carlo draw is required. CEMDAP simply replicates each sample record so that there is one occurrence per unit value of its integerized weight. All other popsyns employ a Monte Carlo procedure, but several differences exist.

TRANSIMS and Oregon2 sample randomly with replacement from all households of the same demographic category within the same PUMA. This is simple and unbiased, but results in a synthetic population in which some sample households within a category occur far more frequently than others in the same category. Also, the forecast synthetic population can be extremely different than the base year synthetic population, simply because of stochastic variation. The other popsyns deal with one or both of these problems in various ways.

SFCTA and Metro also use random sampling with replacement, but they deal with the second problem by generating and saving a random seed for each demographic category within each geographic unit. The seed serves as a starting point in a fixed sequence of pseudo-random numbers. The same seeds are used for the base year and all forecast year popsyns. For a given demographic category, the number of households may differ between base and forecast, but the year with more households simply adds more households.

MORPC ignores this problem, but deal with lumpy sampling of households by sampling without replacement within PUMA. If all households in a category get sampled, then they are all released for a second round of sampling. MORPC also deals with a problem that only it encounters: non-zero draw probabilities in cells for which there are no sample households. They occur because during IPF, MORPC adds a tiny fractional value to all technically feasible categories that would otherwise be zero (because of no corresponding PUMS record). If the Monte Carlo procedure requests a draw from an empty category, then a second draw is attempted from a category with the next smaller household size and number of workers. This slightly biases the synthetic population.

The design for ARC deals with both problems. It adopts the SFCTA and Metro approach of retaining draw seeds for base and forecast years. It also draws without replacement, like MORPC, but with a couple differences. First, sampling without replacement starts over for each TAZ, so that the previously described feature works properly. Second, the user can restrict the number of times that a household can be drawn into the same TAZ; if more draws are needed, households of the same category are drawn from the most similar PUMA, as defined by the user. Finally, an unbiased smoothing procedure is employed that assures approximately uniform sampling frequencies among households within a given PUMA demographic category.

The author has been unable to identify the drawing protocol used in TRESIS.

Extracting variables from census and locating hoseholds within the geographic unit

All of the popsyns extract a subset of the available variables for each sample household. Most of them extract a pre-programmed set, that varies among popsyns. For TRANSIMS and ARC, on the other hand, the list of variables is a user-supplied input table.

By construction, each household in the synthesized population is associated with a geographic unit. Several popsyns employ a supplemental procedure to assign each household to a more specific location within the geographic unit. These include TRANSIMS, Metro and Oregon2. CEMDAP uses a similar (Monte Carlo) procedure to translate block group assignments into TAZ assignments.

VALIDATION

The author has found no written validation results for Metro, MORPC or Oregon2.

CEMDAP is the only implemented popsyn that produces validation statistics as a standard output of the synthesis procedure. However, it only partially validates—no variance statistics, and only univariate statistics—base year control variables, which it should be able to replicate almost perfectly. At the tract level, synthesized control variables differ from actual control variables by 1-5% most of the time, with a maximum discrepency of 12%. This says nothing about the ability to synthesize uncontrolled variables in the less well informed forecast year scenarios.

SFCTA partially validates only control variables, but does it for base year and forecast year, at three levels of aggregation. Marginal distributions match within 1% at the PUMA level. For small TAZs, some marginal totals differ by up to 22%. Here again, the synthesis of uncontrolled forecast year variables is untested.

TRESIS partially validates all control variables (not jointly) and several uncontrolled variables at the regional level, its level of synthesis. In the two primary dimensions, the controlled variables match perfectly. In the third dimension, an 18% discrepency occurred in the number of households with 5 or more workers; the small 400 household regional sample is unable to replicate the frequency of this rare event. In uncontrolled estimates, discrepencies ranged from two to 17% in nine household categories, and from one to 47% in 21 person categories.

Of the implemented popsyns, TRANSIMS underwent the most rigorous validation. Beckman et al (1995) correctly focused attention on uncontrolled variables, studying the univariate distribution of household size and the bivariate distribution of household size by number of vehicles. To study household size, they synthesized one census tract (actual population 5592) 100 times. The synthesizer controlled households of size 1, so it matched perfectly. The average number (across 100 synthetic populations) of households of sizes 2, 3, 4 and 5+ differed from actual by .6%, 2.6%, 2.8% and 8.9% respectively, and the maximum discrepencies were approximately 3%, 11%, 9% and 21%. Validating multivariate distributions is difficult because the census tables, that can supply ‘actuals’ for validation, have few multidimensional tables. To validate the bivariate distribution by household size and number of vehicles, Beckman, et al used the entire 5% PUMS from 22 PUMAs in the San Francisco Bay area, treating it as a 100% census from 22 pseudotracts in one pseudoPUMA. A 5% pseudoPUMS was drawn from the pseudoPUMA. TRANSIMS was used to synthesize the population of the 22 pseudotracts, and the joint distribution was compared to the actual joint distribution in the 22 pseudotracts. In all except 3 of the pseudotracts, the discrepency between actual and estimated number of households in a given household size by number of vehicles category was always less than about 5%, and usually less than 2%. The three pseudotracts came from the most urban PUMAs—the City of San Francisco—where TRANSIMS overestimated vehicle ownership in small households. For example, in two pseudotracts the number of one-person households with no car was underestimated by nearly 20%. The variance across pseudotracts, or across multiple synthetic populations for the same pseudo tract, were not reported. The results of this test are probably worse than they would be if the data represented true tracts from a true PUMA, which would probably be more homogeneous than the pseudotracts. Although the reported validation results are good, they ignore the univariate and bivariate distributions of any other uncontrolled variables, which may not be synthesized as accurately as household size and number of vehicles, and they also ignore forecast year validation, which would probably be less accurate because the forecast controls, even if correct, would be less detailed than the base year census controls.

Because of legitimate concerns at ARC about the reliability of forecasts based on microsimulation, the design of ARC, which is still under development, provides more rigorous validation than the other popsyns. A validation component (validator) is included as an integral part of the application software, producing validation statistics whenever a population is synthesized.

ARC forecast year synthesis will be validated through a backcast, as follows. The forecast population will be synthesized from the base year (2000) synthetic population and 1990 Census statistics summarized to emulate available ARC demographic and land use forecasts. The validation procedure will compare details of the resulting synthetic population to those available in the 1990 census tables. This procedure removes from the synthesized population all error introduced by bad backcast controls, focusing attention on the ability to accurately synthesize the forecast population, given accurate forecast inputs at the available level of detail.

The ARC validator calculates the values of approximately 100 summary household and personal characteristics in the synthetic population at four levels of geographic aggregation (tract, PUMA, county and super-county). It then calculates the percentage difference between each of these values and corresponding validation values taken from census tables, and summarizes the differences (as mean, variance, max and min) for each geographic level. For example, one summary characteristic is the percentage of households that have only one person, with the person being age 65 or older. Some of the characteristics provide alternative ways of aggregating demographic categories. For example, one row combines households of size 5+, whereas others split the group into size 5, size 6 and size 7+. With a variety of geographic aggregations and a variety of demographic aggregations available, it can be decided, for each characteristic, which geographic-demographic aggregation combinations are reliable and which are not. For example, household size categories 5, 6 and 7+ may be reliable only at the supercounty level, but size 5+ may be reliable at the PUMA level.

TOWARD AN IMPROVED APPROACH

The comparisons in this paper were drawn only from documentation, rather than from hands-on use and validation of the synthesizers in head-to-head competition. Therefore any conclusions can only be tentative, and are subject to correction under a more rigorous comparison. Nevertheless, this section makes some preliminary recommendations for improvements that could be made as these synthesizers are enhanced, or as new synthesizers are developed.

The first and strongest recommedation relates to validation. For a population synthesizer to be effective, the bias and covariance of its estimated population attributes should be known (and acceptable) for various levels of demographic and geographic aggregation, including its estimates of any controlled and uncontrolled base year and forecast year variables that are used by models or for segmenting model outputs. Proper validation involves evaluating these statistics during development, so that any necessary adjustments can be made and to demonstrate acceptability. The validation statistics should also be produced when the synthesizer is used, so that its reliability for particular uses can be evaluated if necessary. Unfortunately, but not surprisingly, none of the compared popsyns is properly validated according to the above criterion, although the ARC validator design is promising.

A second, and related, recommendation is for the implementation of flexible software that offers the user significant power to adjust basic attributes of the popsyn without additional programming. The primary reason for this is to make it possible to easily adjust and tune the popsyn during validation. Without such flexible software, attempts at validation will probably continue to be as weak as in the current batch of synthesizers. Aspects that may benefit the most from flexibility include: (a) the definition of the cells in the joint distribution; (b) the definition of the control variables and their categories (including the provision to use different variables and categories, and a mix of geographic aggregations) in the forecast year than in the base year, since less detail is available in forecast years; and (c) the definition of variables for which validation statistics are automatically produced.

In addition to these two recommendations, attractive features of some of the popsyns can be noted. It may be possible to develop a superior popsyn that incorporates the most attractive features from several of the compared popsyns:

- The flexible joint distribution and control variable definitions of ARC are attractive. However, CEMDAP's use of the PUMS itself for the joint distribution may be the most elegant and flexible approach of all, easily enabling controls on the numbers of persons and numbers of households of various types, and allowing different controls for base year and forecast year without concern for the structure of the joint distribution.

- The preliminary IPF used by TRANSIMS is attractive because it assures that the disaggregate distributions estimated in the main IPF match disaggregate AND PUMA level values of the control variables.

- The preliminary IPF used by ARC is also attractive because it helps assure that the main IPF converges even if the controls are not mutually consistent.

- Drawing from a rescaled integerized joint distribution should yield less simulation error than simply drawing all households randomly from the distribution. The MORPC approach of maintaining the IPF controls through this procedure is desirable, but it should be enhanced to use Monte Carlo rounding to avoid bias, as is done by SFCTA and Metro.

- Controling the draws through sampling without replacement (MORPC), retaining draw seeds (SFCTA and Metro) and providing user control (ARC) are also attractive; however, here again, CEMDAP's method of using the PUMS for the joint distribution may be a more elegant solution, avoiding the problems that the other procedures are designed to overcome.

Another recommendation, if it can be called that, would be to make census sf1, sf3 and CTPP tables all available for the same small geographic unit, either block group or TAZ, or both. This would enable controls from all three sets of tables to be used together without suffering from the loss of information caused by using GIS to convert all files to the same geography.

APPENDIX

This appendix presents, for all compared popsyns, comparable tabular summaries of the features described in the paper.

(to be added)

REFERENCES

Beckman R.J., K.A. Baggerly and M.D. McKay (1996). Creating Synthetic Baseline Populations. Transportation Research A 30, 415-429, 1996.

Bhat, C.R., J. Guo, S. Srinivasan, and A. Sivakumar (2003a). Synthetic Population Generation for Micro-Simulation Activity-Based Travel Demand Modeling Systems. Technical report.

Bhat, C.R., J. Guo, S. Srinivasan, and A. Sivakumar, (2003b) "Activity-Based Travel Demand Modeling for Metropolitan Areas in Texas: Software-related Processes and Mechanisms for the Activity-Travel Pattern Generation Micro-simulator," Report 4080-5, prepared for the Texas Department of Transportation, October

2003.

Bradley, M. (1999). Methodology and Results of Generating a Prototypical Population, working paper, July 8, 1999.

Bradley, M. (2003). A Discussion of the Population Synthesis Approach for Atlanta, working paper, July 28, 2003.

Bradley, M.A., J.L. Bowman, and T.K. Lawton (1999). A Comparison of Sample Enumeration and Stochastic Microsimulation for Application of Tour-Based and Activity-Based Travel Demand Models. Presented at the European Transport Conference, Cambridge UK, September 1999.

Hensher, D.A., P.R. Stopher, P. Bullock and T. Ton (2004). TRESIS (Transport and Environmental Strategy Impact Simulator): Application to a Case Study in NE Sydney, presentation at the 83rd Annual Meeting, Transportation Research Board, Washington, D. C., January 11-15, 2004.

Institute of Transport Studies, Faculty of Economics and Business, The University of Sydney (2004). TRESIS software documentation, available at http://www.its.usyd.edu.au/softwares/softwares.asp.

Los Alamos National Laboratories (LANL). TRANSIMS-Version 3.0, Vol 3--Modules, chapter 2—Population Synthesizer, (LA-UR 00-1725). Jan 8, 2003.

Parsons Brinckerhoff/ PB Consult, J. Bowman and M. Bradley (2004). Progress Report for the Year 2003, Regional Transportation Plan Major Update Project for the Atlanta Regional Commission, General Modeling: Task 2 – Activity / Tour-Based Models, January 5, 2004.

PB Consult (2003). Task 2: Household and Population Synthesis Procedure, PB Consult / Parsons Brinckerhoff, report prepared for the Mid-Ohio Regional Planning Commission as part of The MORPC Model Improvement Project. March 12, 2003.

PBConsult, HBA Specto Inc. and EcoNorthwest (2003). HA Module Description at Finalization, First Draft Report Submitted to the Oregon Department of Transportation, July, 2003.

Ton, T. and D.A. Hensher (2002). A Spatial and Statistical Approach for Imputing Origin-Destination Matrices from Household Travel Survey Data: A Sydney Case Study, in Proceedings of the 25th Australasian Transport Research Forum, Canberra, October (CD-Rom).

Ton, T. and D.A. Hensher (2003) Synthesising population data: The Specification and Generation of Synthetic Households in TRESIS, in Proceedings of the 9th World Conference of Transport Research., Elsevier, Oxford.