Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels

Joshua Auld and Abolfazl Mohammadian

This paper details a new methodology for controlling attributes on multiple analysis levels in a population synthesis program. The methodology determines how household- and person-level characteristics can jointly be used as controls when populations are synthesized as well as how other multiple-level synthetic populations, such as firm and employee or household and vehicle, can be estimated. The use of multilevel controls is implemented through a new technique involving the estimation of household selection probabilities on the basis of the probability of observing each household, given the required person-level characteristics in each analysis zone. The new procedure is a quick and efficient method for generating synthetic populations that can accurately replicate desired person-level characteristics.

Population synthesis is recognized as an integral component in activity-based modeling. Beginning with the development of the TRANSIMS population synthesizer (*1*), increased focus has been directed at developing synthetic populations for use in travel demand microsimulation (*2–5*) and many other agent-based microsimulation applications (*6, 7*). Population synthesis generally uses a sample of households at an aggregate geography combined with marginal data on household characteristics at a disaggregate geography to generate a set of households that satisfy known marginals at the small-area level. Population synthesizers often use a well-known statistical technique, iterative proportional fitting (IPF) (*8*), and probabilistic selection to generate synthetic populations, although other procedures have recently been developed (*9*). Either way, a population synthesizer creates copies of sample households and locates them geographically to replicate the full population of the study area. For a more in-depth discussion of the IPF procedure and basic population synthesis techniques see Beckman et al. (*1*), Arentze et al. (*10*), and Hobeika (*11*), among others. The original population synthesis program in which the current work is implemented is discussed at length in Auld et al. (*12*). This program implemented the basic IPF procedure and probabilistic selection and was developed for use in an activity-based model system (*13*).

The increasing focus on population synthesis has resulted in recognition of some limitations of the basic synthesis method. This paper aims to improve the methodology behind the basic population synthesis routine to account for multiple levels of analysis units–control variables, which was a limitation to earlier population synthesizers. The paper includes a discussion of the literature on the issue, a description of a newly developed method to address it, validation of the new method and evaluations of its computational performance, and finally, a discussion of the value of the new method and directions for future work.

PREVIOUS WORK IN POPULATION SYNTHESIS

The methodology behind most population synthesizers used in travel demand modeling is generally derived from the synthesizer developed by Beckman et al. (*1*) for the TRANSIMS project, although some recent work has also addressed the combinatorial optimization approach (*7, 9*) or combinations or permutations of both (*14, 15*). During the development of different population synthesizers, many limitations of the basic methodology have been observed. Subsequent research has focused on attempts to correct for these deficiencies and extend the usefulness of synthesis methods (*14, 16*). Several problematic issues relating to population synthesis that have been observed at various times include zero-cell issues arising from using sample data, biases introduced as a result of rounding the joint distributions, biases introduced as a result of simulation, and lack of multiple levels of control (*1, 9, 15, 16*). Different strategies have been proposed to address these issues, for example, the zero-cell problem has been addressed by tweaking the joint distribution from the IPF procedure (*1, 16*) and by limiting the number of control variable categories (*12, 16*).

The limitation of population synthesis methods to only one analysis level has recently begun to receive more attention. Traditionally, population synthesizers consider control variables for only one level because joint distributions between household- and person-level control variables cannot be constructed. Therefore the IPF procedure and selection procedure as found in Beckman et al. (*1*) cannot be implemented directly for household- and person-level variables simultaneously (*16*). Researchers have attempted to overcome this in several ways, including household reconstruction methods (*15*) or using population characteristics to impute household-level distributions (*10*). Recent work has focused on methods to address the issue directly in the synthesis procedure, rather than as a reconstruction step. Guo and Bhat account for person-level controls by developing joint distributions for individuals and households separately, and then synthesizing households while considering whether the person- or household-level constraints would be violated beyond a given

Department of Civil and Materials Engineering, University of Illinois at Chicago, 842 West Taylor Street, Chicago, IL 60607-7023. Corresponding author: J. Auld, auld@uic.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2175, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 138–147.
DOI: 10.3141/2175-16

threshold, although only the household distribution is considered when drawing households (*16*). Ye et al. developed the only previous attempt of which the authors are aware to directly and simultaneously control on multiple levels (*14*). They used an iterative reweighting procedure to heuristically solve for household weights considering both household and person constraints together before the household selection procedure. The methodology presented here is a new, efficient procedure for considering joint multilevel controls implemented directly in the selection stage, which builds on the basic IPF and household draw procedure and which, to the best of the authors' knowledge, has not been implemented previously. For details of the basic procedure see Auld et al. (*12*). The new procedure is discussed in the following sections.

MULTILEVEL CONTROL METHODOLOGY

This section discusses the methodology used for multilevel control, implemented in the basic population synthesis program described in Auld et al. (*12*). Multilevel control allows population characteristics to be replicated when the synthetic population is created for more than one analysis level, with one level such as households serving as the base level of analysis that contains the sublevel analysis unit, such as persons. There is, however, no requirement that the analysis be used only for synthesizing households or individuals. Any situation in which marginal and sample data are available for a base- and sublevel of analysis (i.e., firms or employees, households or vehicles, buildings or tenants, etc.) can be synthesized by using the program. The only limitations are that the membership size of the sublevel within the base level must be used as a control (i.e., household size if using household or individual) and the sample data for the base- and sublevels must be linked by unique identifiers. The second requirement results from the fact that the program uses a procedure in which the base units are generated and their component subunits are copied with them rather than each subunit being synthesized separately. Because the subunits are copied with the base unit, there must be a link between the base- and subunit sample data. For clarity the base- and sublevels of analysis are referred to hereafter as simply household level and person level.

Household Selection Probability Considering Person-Level Constraints

One feature most population synthesizers share is the creation of synthetic households through probabilistic selection. This procedure involves setting a probability for selecting a sample household into the synthetic population on the basis of the sample weight of the household, the number of total households required, the number of households of the current type already generated, and so forth. This is the basic procedure followed in the synthesizer by Beckman et al. (*1*) and others. Selection probabilities are assigned for households that are then replicated through simulation. The probabilities increase with the weight of the household and decrease as the required frequency of the current household type is reduced through the simulation process. The required frequency of each household type is taken from the estimated household joint distribution created through the IPF process. Population synthesizers may depart from this basic methodology, as in the procedure developed by Ye et al., in which the frequencies determined in the IPF procedure are used in a heuristic iterative solution to set household weights such that person-level constraints are satisfied (*14*). Even in that case, however, simulation is still used to create the synthetic households by using the reweighted IPF results. The general selection probability as described in Beckman et al. is shown in Equation 1 (*1*).

$$P_{i,C} = \frac{W_i}{\sum_{k=1}^{N_C} W_k} \tag{1}$$

where

$P_{i,C}$ = probability of selecting household *i* of household type *C*,
W_i = household weight for household *i*, and
N_C = remaining households in subregion sample of type *C*.

This equation states that the probability of selecting the current household *i* of a given demographic type *C* is equal to the weight of the current household divided by the sum of the weights of all other households in the sample of the same type. This selection procedure ensures that households with a higher sample weight are selected more frequently when the households are synthesized. This selection probability does not account for differences between households on the person level. Therefore, a new selection probability, shown in Equation 2, was developed that explicitly accounts for the person-level distribution when the households are synthesized.

$$P_{i,C} = \frac{W_i \prod_{j=1}^{N_{\text{per},i}} \frac{\text{MWAY}^*_{\text{per}}\left(v_{1,j}, v_{2,j}, \ldots, v_{n,j}\right)}{N_{\text{remain}}}}{\sum_{k=1}^{N_C} \left(W_k \prod_{l=1}^{N_{\text{per},k}} \frac{\text{MWAY}^*_{\text{per}}\left(v_{1_l}, v_{2_l}, \ldots, v_{n_l}\right)}{N_{\text{remain}}} \right.} \tag{2}$$

where

$P_{i,C}$ = probability of selecting household *i*, of household type *C*,
W_i = household weight for household *i*,
$N_{\text{per},i}$ = number of people in household *i*,
$\text{MWAY}^*_{\text{per}}(v_{1,j}, \ldots, v_{n,j})$ = remaining cell frequency in zonal person-level joint distribution,
$v_{i,j}$ = index of control variable *i* for person *j*,
N_{remain} = number of individuals not yet created in zone, and
N_C = remaining households in subregion sample of type *C*.

The selection probability defined in Equation 2 has the same form as in Equation 1, with the addition of the product terms in the numerator and denominator. These product terms are essentially the probability of observing a household composed of each individual household member given the remaining persons to be synthesized according to the person-level joint distribution, $\text{MWAY}^*_{\text{per}}$. This selection probability is derived from a straightforward application of Bayes theorem, that is, the probability of selecting the current household *H* is the probability of observing household *H* given the current household type *C*. This is equivalent to the probability of observing each member in the household together divided by the sum of the probability of observing each household member together for all households of the same type, assuming no correlation between the probabilities for individual household members. This assumption is generally incorrect in actuality and would cause problems if households were being reconstructed on the basis of individual probabilities. However, because only the individual probabilities are being used to weight household selection, that does not matter—even if unlikely households are weighted the same as likely households on the basis of their individual members, the likely household type is naturally more likely to be observed in the sample data and will therefore be more likely to be generated, as expected. So assumption of independent individual probabilities is corrected by the household weighting term to produce the proper results. This can be reduced even further with the proper choice of household-level control variables. This new selection probability allows the household selection procedure to generate households with individuals that most closely match the required person-level joint distribution. This is best demonstrated with an example, shown in Figure 1.

In this example, 25 households of the same type are synthesized from a sample of four households with the person-level joint distribution shown in Part *a.* The basic procedure is shown in Part *b*, in which all four households have the same selection because they have the same weight and the person-distribution is ignored, so that the same number of each household is generated with the resultant synthesized person-level distribution clearly not matching the expected. Part *c* then shows the results when the new selection probability is used. Now the households with more frequent person types in the person-level distribution (HH1, HH4) are generated more than the

a) Starting Data

Microdata sample:

HH1: 1 employed male, 1 employed female, weight = 1

HH2: 1 unemployed male, 1 employed female, weight = 1

HH3: 1 unemployed male, 1 unemployed female, weight = 1

HH4: 1 employed male, 1 unemployed female, weight = 1

Person-Level Joint Distribution:

	Employed	**Unemployed**	**Total**
Male	20	5	25
Female	10	15	25
Total	30	20	50

HH-Level Joint Distribution

		Total
HHSize = 2	25	25

b) Selection with Household-Level Control Only

Selection Probabilities (Equation 1):

$P(HH1)$ = P(HH2) = P(HH3) = P(HH4) = 1/(1+1+1+1) = 0.25

Synthesized Person-Level Distribution:

	Employed	**Unemployed**	**Total**
Male	12.5	12.5	25
Female	12.5	12.5	25
Total	25	25	50

HH1_count = 0.25 x 25 = 6.25

HH2_count = 0.25 x 25 = 6.25

HH3_count = 0.25 x 25 = 6.25

HH4_count = 0.25 x 25 = 6.25

c) Selection with Household- and Person-Level Controls

Selection Probabilities (Equation 2):

$$P(HH1) = \frac{(1)[(20/50)(10/50)]}{(1)[(20/50)(10/50)]+(1)[(5/50)(10/50)]+(1)[(5/50)(15/50)]+(1)[(20/50)(15/50)]} = 0.32$$

$$P(HH1) = \frac{(1)[(5/50)(10/50)]}{(1)[(20/50)(10/50)]+(1)[(5/50)(10/50)]+(1)[(5/50)(15/50)]+(1)[(20/50)(15/50)]} = 0.08$$

$$P(HH1) = \frac{(1)[(5/50)(15/50)]}{(1)[(20/50)(10/50)]+(1)[(5/50)(10/50)]+(1)[(5/50)(15/50)]+(1)[(20/50)(15/50)]} = 0.12$$

$$P(HH1) = \frac{(1)[(20/50)(15/50)]}{(1)[(20/50)(10/50)]+(1)[(5/50)(10/50)]+(1)[(5/50)(15/50)]+(1)[(20/50)(15/50)]} = 0.48$$

Synthesized Person-Level Distribution:

	Employed	**Unemployed**	**Total**
Male	20	5	25
Female	10	15	25
Total	30	20	50

HH1_count = 0.32 x 25 = 8

HH2_count = 0.08 x 25 = 2

HH3_count = 0.12 x 25 = 3

HH4_count = 0.48 x 25 = 12

FIGURE 1 Example of selection probability calculation.

others. In this simple situation the person distribution is matched exactly. The example shows that with the new selection probability the person-level marginals and joint distribution are matched when the household and person data are consistent.

Updated Household Selection Procedure

The new household selection probability requires much more calculation than the basic household selection probability definition. Under the base methodology, the sum of the weights in the household sample can be calculated once before the synthesis procedure begins and the number can be reused, but the new methodology requires the sum to be recalculated every time the probability is calculated because the product changes whenever a household/person is synthesized. Therefore a new selection procedure was needed to ensure that populations could be synthesized more efficiently. The new procedure is described in this section.

The procedure behind the new synthesis methodology is as follows for each subregion (i.e., geographic area at which sample data are available):

1. Generate subregion-level household (HH) and person joint distributions.
   – Create subregion-level household joint distribution from household sample data,
   – Create subregion-level person joint distribution from person sample data,
   – Use IPF to fit household joint distribution to HH marginals from marginal data, and
   – Use IPF to fit person joint distribution to person marginals from marginal data.
2. Get next geographic zone in the subregion.
3. Generate zone-level HH and person joint distributions.
   – Seed zone HH joint distribution with subregion joint distribution,
   – Seed zone person joint distribution with subregion joint distribution,
   – Use IPF to fit HH joint distribution to zone marginal data, and
   – Use IPF to fit HH joint distribution to zone marginal data.
4. Run household selection procedure.
   – Get next household, *H*, randomly from the subregion sample;
   – Calculate household selection probability *P* by using Equation 1;
   – Make *N* attempts to add copy of H with probability *P*, with *N* as remaining houses of current type needed in HH joint distribution;
   – Reduce cell in zone HH joint distribution by number of *H* added;
   – Remove *H* from subregion sample; and
   – If households remain in subregion sample, return to first substep in Step 4.
5. Add all removed households back to subregion sample.
6. If iterations are less than max and households still needed, return to Step 4.
7. If zones remaining in subregion, go to Step 2.
8. If subregions remaining, get the next subregion and go to Step 1, or else finish.

The procedure allows for simultaneous household and person control (as described in the previous section) while enhancing the efficiency of the algorithm. In the traditional household selection procedure, the list of households in the subregion sample is searched through many times to generate the required number of households (*1, 16*). The search procedure generally occurs as follows:

1. Get current household from household list.
2. Set selection probability based on Equation 1 multiplied by remaining frequency of household type divided by total remaining households.
3. Determine whether household is added on the basis of selection probability.
4. Return to Step 1 if households are still needed in zone.

This procedure generally requires much iteration through the subregion household sample, a process that takes a fairly long time to complete when the new selection probability calculation described previously is used. The new selection procedure simply searches once through the sample household list for each zone. For each household in the sample, the procedure calculates the selection probability and then makes a number of attempts to copy the household equal to the remaining frequency in the household joint distribution for the household type. Each time a copy of the household is successfully added, the probability is updated. After all attempts have been made, the household is removed from consideration, so that it does not figure into the selection probability calculation for later households. This continues until all households in the list have been searched, at which point the full population is synthesized. The list is searched in random to ensure that any biases in the ordering of the sample data are not transferred to the synthetic population.

This process guarantees that the full population is synthesized in one pass through the sample list, greatly reducing the computation run time. However, as a result of random rounding during the synthesis procedure (i.e., if 3.4 households are required, this will be realized as either three or four households), marginal totals are sometimes violated (*12*). Therefore a marginal constraint is added to the selection procedure at the household and the person level. This constraint takes the form of an additional rule: if a household is to be added, neither the household nor any individual in the household can cause any of the household- or person-level marginals to be exceeded by more than a user-defined tolerance. If the marginal constraints will be violated, the household is not added. This generally leads to the result that less than the full number of households is generated, usually as a result of inconsistencies and incompatibilities in the data. Therefore the selection procedure is run for up to a user-defined maximum number of iterations, at which point the marginal constraints are relaxed and the full number of households is generated.

Another problem sometimes arises as a result of the nature of the selection probability. When the probability for a household is calculated, if one of the household members is not needed (i.e., has a remaining frequency of zero in the joint distribution), the selection probability for that household goes to zero. This is an intentional feature of the procedure and is almost always desirable, but can occasionally cause problems when there are incompatibilities between the household- and person-level data. For example, zones such as Block Group 170312704002 in Cook County, which has seven households of household size four but only 20 total people, will cause the selection procedure to fail. In this example, after the fifth household is generated, there are no people left in the person-level joint distribution, so the household selection probability goes to zero and no households are selected no matter how many iterations are run. Therefore, on the final iteration of the procedure, if there are still households remaining to be generated, the program disregards all person-level controls and generates the remaining households on the basis of only the household weights by using the selection procedure seen in Equation 1.

PERSON-LEVEL CONTROL VALIDATION RESULTS

To assess the validity of the new person-level control methodology, a synthetic population created with the new routine was validated against the same population created without person-level control. The validation for the person-level control procedure was conducted on 846 block groups in the Chicago-land six-county region, where household- and person-level marginal control incompatibilities were minimal. Many block groups had populations less than the population estimated from the household size control variable, an error that causes less than the full number of households to be generated (because all person-level probabilities are set to zero before all households are generated). The selected block groups have a total of 553,387 households containing 1,498,482 individuals, approximately 20% of the total six-county population. These block groups were selected such that there were no group quarters population and the differences between estimated population totals based on the household size control variables and the population totals in the person-level marginals were less than 2%, to separate out error due to the procedure from error caused by data issues. Block groups with group quarters are excluded from this analysis only because including a marginal variable relating to group quarter status does not add anything to the person-level validation. When synthetic populations are generated for actual modeling purposes, it is a straightforward, although cumbersome, procedure to add a group quarters control marginal at the household level that enables block groups with substantial group quarters populations to be generated. In this manner, the validations run below are comparing the differences in procedure rather than differences due to data issues.

Two separate populations were synthesized, one by using only household controls referred to as POP-HH and one with an additional set of person-level controls referred to as POP-PER. The household controls used for both populations were

- Household size—seven categories,
- Household income—16 categories,
- Household number of workers—five categories, and
- Total household joint distribution size—560 cells.

The person-level controls used in generating POP-PER were

- Gender—two categories,
- Age—eight categories,
- Race—seven categories, and
- Total person joint distribution size—112 cells.

These variables were selected for demonstration purposes only; the purpose of this exercise is to confirm that using person-level controls improves the person-level fit results, not to validate the use of this particular set of control variables. Any set of household- and person-level variables for which adequate sample and marginal data exist can be used because the synthesis program is designed to be as general as possible (*12*).

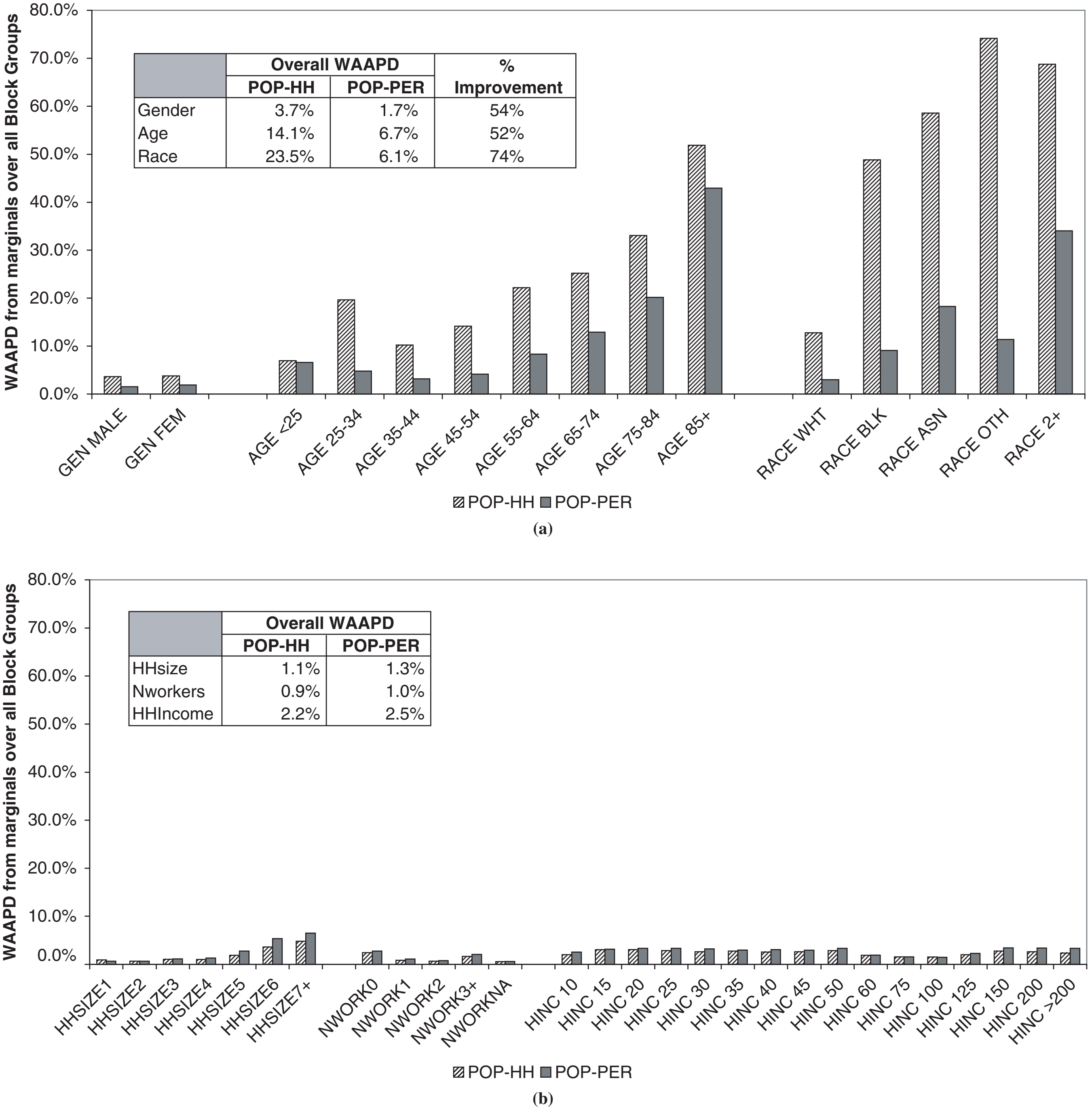
Both synthetic populations were able to exactly match the total number of households required, with each generating the actual total of 553,387 households. In addition the total number of individuals generated was almost exact for each synthetic population, as expected even for the non-person control population as a result of the inclusion of a household size variable as a control. The POP-HH population contains 1,500,308 people, 0.1% more than required, whereas the POP-PER population contains 1,487,815 people, 0.7% less than required. The marginal fit comparison, in regard to weighted average absolute percent difference (WAAPD) between the known and synthesized marginal totals over all block groups, for both populations is shown in Figure 2. The Native American, Alaskan, and Hawaiian categories in the race control are not shown because these categories represent less than 0.25% of the population in the region although both exhibited improvement similar to the other categories.

The person-level comparison, shown in Figure 2*a,* demonstrates a substantial improvement in fit between the POP-HH and POP-PER marginal totals on the person level, as expected. Overall there is an improvement in fit of between 52% and 74% over each person-level category, showing that the new routine allows a marked improvement in fitting to person-level marginal control totals. As seen in the figure, even under person-level control, the average error associated with certain marginal categories can still be large, although always less than with no person control. This is the result mainly of rounding errors and difficulty satisfying the marginal constraints for infrequent categories. The largest errors in the marginal fit are seen for the over-85-years-of-age category and the two-or-more-races category for the age and race marginals, respectively, which each represent less than 2% of the total population. In fact all marginal categories that have a WAAPD of more than 15% contain less than 5% of the population, meaning that the large errors are the result mainly of small category sizes.

The household-level comparison in Figure 2*b* shows that the improvement in marginal fit using person-level controls comes at a minimal cost to the accuracy of the household-level marginals. All marginal control totals are matched fairly precisely in the POP-HH and POP-PER synthetic populations, with larger errors again seen in the less frequent categories. All household marginal categories had under a 7.0% WAAPD value.

One point about the procedure should be noted concerning the relaxation of the person-level constraints used to ensure convergence when selecting households. It is clear that allowing the person-level constraints to be violated introduces errors into matching the expected person-level marginals, causing most of the differences seen in Figure 2*a.* However, analysis shows that in general it is a very small number of generated households and individuals that contribute to these violations, so the impacts are most likely not particularly large. For the POP-PER synthetic population, on average more than 97% of households (2% standard deviation) and 95% of individuals (3% standard deviation) were generated before the person-level constraints were relaxed.

The previous analysis shows only how the population matches the marginal characteristics. Therefore each synthetic population was also evaluated on how well the required household- and person-level joint distributions were matched. This is evaluated by estimating the absolute percent difference between the synthesized and expected (from IPF) frequencies for each cell in each block group. This value is then averaged over all block groups to obtain an average absolute percent difference (AAPD) value for each cell in each joint distribution. The AAPD values for each synthetic population are then plotted against the average cell frequency, along with a



	Overall WAAPD		%
	POP-HH	POP-PER	Improvement
Gender	3.7%	1.7%	54%
Age	14.1%	6.7%	52%
Race	23.5%	6.1%	74%

(a)

	Overall WAAPD	
	POP-HH	POP-PER
HHsize	1.1%	1.3%
Nworkers	0.9%	1.0%
HHIncome	2.2%	2.5%

(b)

FIGURE 2 WAAPD comparison for (*a*) person-level marginals and (*b*) household-level marginals.

theoretical estimated AAPD from rounding error calculated as shown in Equation 3 below:

$$\mathrm{AAPD}_i = \frac{\sum_{j=1}^{N_{\mathrm{BG}}} \mathrm{APD}_{i,j}}{N_{\mathrm{BG}}}$$

$$\mathrm{APD}_{i,j} = \left(1 - p_{i,j}\right)\frac{x_{i,j} - \left(x_{i,j} - p_{i,j}\right)}{x_{i,j}} + p_{i,j}\frac{\left(x_{i,j} + 1 - p_{i,j}\right) - x_{i,j}}{x_{i,j}}$$

$$= \frac{2\left(p_{i,j}\right)\left(1 - p_{i,j}\right)}{x_{i,j}} \tag{3}$$

where

$\mathrm{APD}_{i,j}$ = expected absolute percentage difference from value in cell i for block group j from rounding,
AAPD_i = average APD for cell i over all block groups from rounding,
$p_{i,j}$ = $x_{i,j}$ (mod 1),
$x_{i,j}$ = value in cell i of person-level joint distribution for blockgroup j, and
N_{BG} = number of block groups (zones).

Equation 3 states that the expected absolute percent difference for each cell in the joint distribution for each block group is the probability of rounding the cell down multiplied by the error caused by this plus the probability of rounding the cell up multiplied by the error caused from rounding up, where the probability is determined by the decimal portion of the actual cell value (i.e., a cell value of 1.2 will be rounded down 80% of the time and rounded up 20% of the time, so that 80% of the time the error is 0.2/1.2, or 16.7%, and 20% of the time the error is 0.8/12, or 67%, for an average of 26.7%). The values for each block group are then averaged to obtain the AAPD value for each cell. These values are plotted, along with the AAPD values from the POP-HH and POP-PER populations in Figure 2 for the household- and person-level joint distributions. Note that these values are plotted against average cell frequency, so that a cell with an integer average frequency will still have expected average rounding error.

Figure 3*a* shows the results of the comparisons of the AAPD values for each cell in the household distribution matrix for the POP-HH and POP-PER synthetic populations. The figure shows that the populations produced through both procedures replicate the household-level joint distribution reasonably well, with the AAPD values approaching the theoretically expected value as a result of random rounding. In fact, the population generated with person controls actually slightly outperforms the base procedure in satisfying the household distribution with an average AAPD over all cells of 89% compared with 125% for the POP-HH population. That is possibly the result of a more targeted search being performed through the use of the person-level controls and constraints.

The results presented in Figure 3*b* show that, as expected, the fit of the POP-PER synthetic population to the person-level joint distribution is much better than the fit of the POP-HH population, as a result of the use of the person-level controls. The overall AAPD improves from 407% for the POP-HH to 118% for the POP-PER population, which is a significant improvement. The cell AAPD values for the POP-PER population are generally much closer to the expected rounding error, whereas large differences can be seen in the POP-HH AAPD. Although the POP-PER AAPD values also generally follow the expected pattern of decreasing error with increasing average cell size, that is not the case with the uncontrolled population, with large errors seen even for several cells with large average sizes, which reinforces the problem with not controlling for person-level characteristics. This result is not due merely to the error caused by large variances in the household size between zones because that is accounted for in the calculation of the expected AAPD value.

Overall, the validation analyses presented in Figures 2 and 3 show that the additional use of person-level controls when a population is generated improves the fit of the resulting population to known person-level characteristics when compared with the same synthetic population generated without person-level controls. The increase in fit to the person-level known marginal totals and estimated joint distribution is very substantial, with little to no sacrifice in the ability to match household-level characteristics. In fact, the ability to match the household joint distribution is somewhat improved through the use of the person-level controls.

A final validation exercise was performed to determine whether the new, more-efficient selection procedure outlined in the section on the updated household selection procedure had any negative impact on the fit of the synthetic populations when compared with the traditional selection procedure. For this validation analysis the selection procedure refers only to the manner in which the sample households are searched; both procedures tested here still use the new household selection probability calculation, which accounts for person-level characteristics. Also, because the test is conducted to determine the validity of the selection procedure rather than the overall synthesis procedure, the marginal constraints were turned off when the test synthetic populations were generated. Three different synthetic populations were generated for 46 block groups in Public Use Microdata Areas (PUMAs) 3408, 3409, 3518, and 3519 in the Chicago region, which had no group quarters population and minimal discrepancies between household size counts and population levels. The three populations were person-level control under the new selection procedure (PER-NEW), person-level control under the traditional selection procedure (PER-OLD), and no person control (PER-NONE).

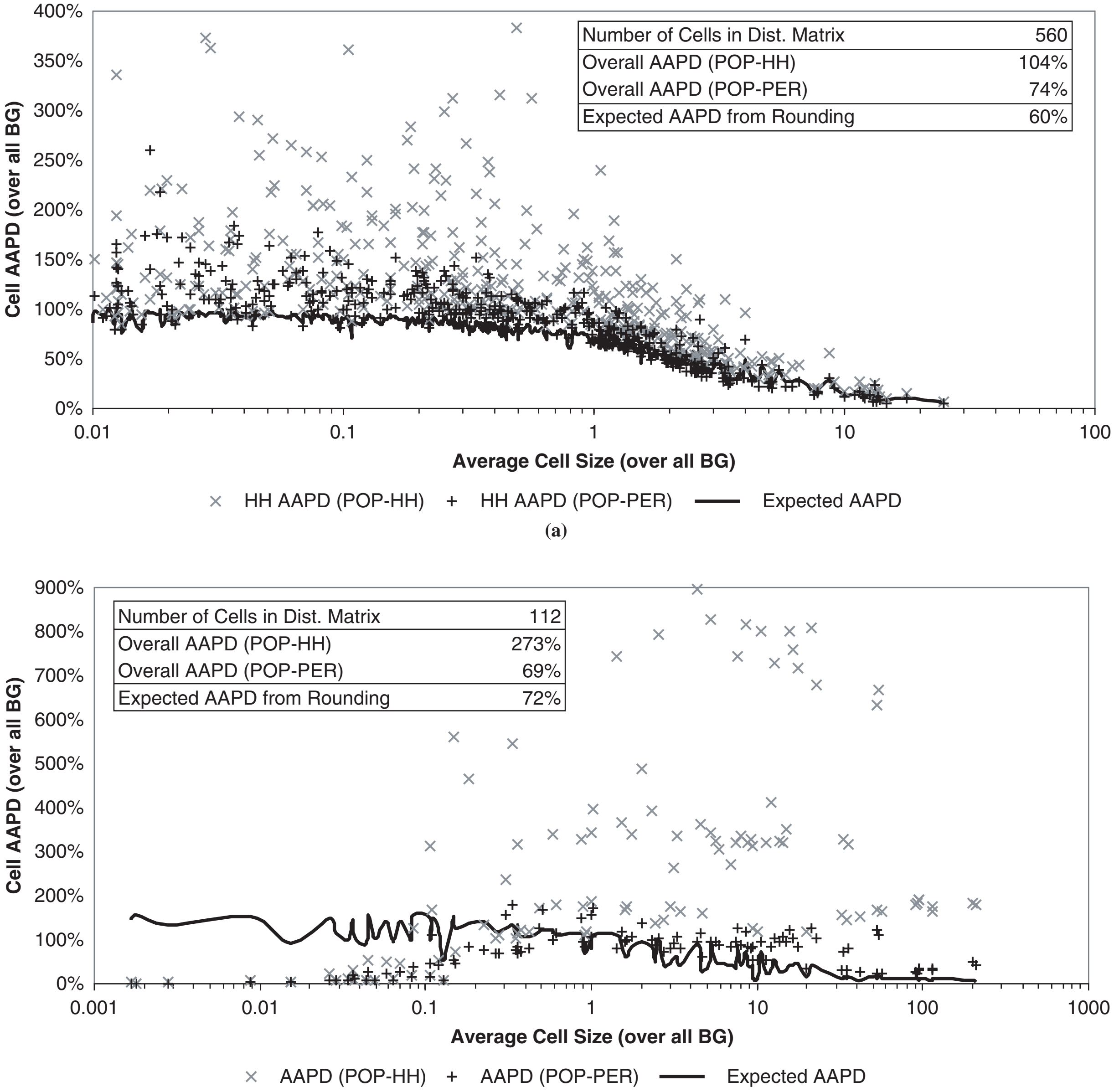
To test for potential biases in the new selection procedure, the Freeman–Tukey test statistic was used to compare the fit of the generated household and person joint distributions with the expected distributions from the IPF process for each procedure. The advantages of this statistic for use in analyzing goodness of fit for synthetic population have been described in Voas and Williamson (*17*) and Ryan et al. (*7*). The test statistic is calculated as

$$\mathrm{FT}^2 = 4\sum_{i}^{N_{\mathrm{cells}}}\sum_{j}^{N_{\mathrm{zones}}}\left(\sqrt{\hat{u}_{ij}} - \sqrt{u_{ij}}\right)^2$$

$$\mathrm{FT}^2 \sim \chi^2\left(N_{\mathrm{cells}} \times N_{\mathrm{zones}} - 1\right) \tag{4}$$

where the statistic is four times the sum of the square of the differences between the square root of actual (u_{ij}) and estimated ($\hat{u}_{ij}$) frequencies over all cells i and zones j and has a chi-square distribution. The test statistic is calculated and compared with a critical value for a given significance level from the χ^2 distribution to evaluate the fit of the synthesized population to the person-level joint distribution. The results for all three synthetic populations are shown in Table 1 for the household- and person-level distributions at a significance level of .05.

According to Table 1, the null hypothesis for the Freeman–Tukey test—that is, the synthesized joint distribution and joint distribution resulting from IPF at the person level have the same distribution—



Number of Cells in Dist. Matrix	560
Overall AAPD (POP-HH)	104%
Overall AAPD (POP-PER)	74%
Expected AAPD from Rounding	60%

(a)

Number of Cells in Dist. Matrix	112
Overall AAPD (POP-HH)	273%
Overall AAPD (POP-PER)	69%
Expected AAPD from Rounding	72%

(b)

FIGURE 3 AAPD comparison for (*a*) household-level joint distribution and (*b*) person-level joint distribution.

TABLE 1 Comparison of Synthetic Population Fit for Different Selection Procedures

	Household-Level Distribution[a,c]			Person-Level Distribution[b,c]		
Population	Crit Value	FT^2 (σ)	H_o^d	Crit Value	FT^2 (σ)	H_o^d
PER-NONE	26,134	4,799 (54)	Accept	5,319	24,786 (434)	Reject
PER-OLD	26,134	5,734 (68)	Accept	5,319	4,044 (106)	Accept
PER-NEW	26,134	6,651 (82)	Accept	5,319	4,840 (102)	Accept

[a] 25,759 degrees of freedom for household-level distribution.

[b] 5,151 degrees of freedom for person-level distribution.

[c] FT^2 values averaged over 20 runs; standard deviation of FT^2 value shown in parentheses.

[d] Null hypothesis accepted if FT^2 is less than critical value at significance level of .05, i.e., probability of observing FT^2 statistic due to random chance is greater than 5%.

is accepted for both populations with person-level controls and rejected for the population without controls, and the household-level distribution is matched for all populations. The results in Table 1 clearly show that using person-level controls improves the fit of the synthesized person-level joint distribution to the estimated distribution, whereas not controlling for person-level characteristics results in poor fit to the estimated distribution, as expected. More important, the good fit to the joint distribution is obtained for both selection procedures. Although the fit obtained by using the new procedure is slightly worse than that obtained by using the traditional procedure, it is still good and results in a run time of 0.7 min to synthesize the 85,590 individuals in the example above as compared with 18.6 min using the other procedure. The run time for synthesizing the entire population in the Chicago region with the traditional procedure assuming the same rates obtained above would be approximately 30 h for a single run compared with the 1.4 h achieved with the new procedure. The long run times using the traditional selection procedure combined with the potential need for running multiple different permutations of a synthetic population and for averaging over multiple runs for the same population motivates the use of the more efficient selection procedure, although the traditional selection procedure can still be used to generate a final synthetic population in combination with initial testing and development done by using the faster procedure. For this reason, both selection procedures are implemented in the actual synthesis program with the choice left to the user.

COMPUTATION PERFORMANCE

Beyond validating the accuracy of the new methodology, it is necessary to evaluate its computation performance. To determine the performance characteristics of the new algorithm, the run times for generating the synthetic populations described in the previous section, POP-HH and POP-PER were compared with run times for generating the full Chicago population, with and without person-level controls. The same program settings, other than the use of person control, were used in each run. Each synthetic population was generated by running the population synthesis program on an Intel Centrino Duo 2.0-GHz processor.

The non-person-controlled population, POP-HH, which contained 1,500,308 synthetic individuals, took 13 min to generate. In contrast, the population with person-level controls, POP-PER, with 1,487,815 people, took more than 28 min. For the full populations, the non-person-controlled full population took about 33 min to generate 7,972,057 individuals, and the person-controlled full population took 84 min to generate 7,889,221, out of a total actual population of 8,091,720. All of the synthetic populations had a household-level joint distribution size of 560 cells and a person-level joint distribution size of 112 cells.

Although it is difficult to compare results across different synthesizers, these run times appear to compare favorably as far as the authors can tell. During the validation of the Atlanta Regional Council population synthesizer, a synthetic population of 1.35 million households controlled only at the household level was run in 17.4 min with a household-distribution size of 316 cells, about half the time it took to synthesize the 2.9 million households in the Chicago region by using only household controls in the new synthesizer (*18*).

The only comparable results available for synthesizers that control for person-level characteristics were presented in Ye et al. for a synthetic population of 2.9 million individuals in Maricopa County, Arizona (*14*). This synthetic population was generated by using a household-distribution size of 280 cells (more than three control variables) and a person-joint-distribution size of 140 cells (over the same three control variables used in this study but with two additional age categories). The overall run time was 16 h, which is substantially longer than the 1.4 h to generate the Chicago population of 7.9 million individuals with approximately the same number of control variables and distribution matrix sizes.

CONCLUSIONS

This paper has detailed the development of a new methodology for using control variables at multiple analysis levels when synthesizing populations with an existing population synthesizer (*12*). The new procedure improves the fit of the synthesized person-level characteristics when compared with synthesis procedures that do not account for person-level controls. Validation of the new methodology shows that the improved fit to the person controls comes at no cost to the fit against the household-level controls. In addition, the introduction of a new household selection procedure has greatly increased efficiency while maintaining good fit to the required person-level controls without some of the run time issues that are found with the use of other methods. Although the discussion in this paper is limited mainly to the household/person synthesis, this methodology can be applied to any analysis with multiple levels of control. Future work is expected on generating shipping firms/vehicles and business firms/employees, for example, by using the same synthesis program. In fact, the applicability of the program is limited only by the availability of data. Overall, the new methodology seems to be an improvement on existing population synthesis techniques for controlling characteristics on multiple levels of analysis.

ACKNOWLEDGMENTS

The authors acknowledge the financial support provided to this project from the Chicago Metropolitan Agency for Planning and the National Science Foundation Integrative Graduate Education and Research Traineeship Program in Computational Transportation Science at the University of Illinois at Chicago.

REFERENCES

1. Beckman, R. J., K. A. Baggerly, and M. D. McKay. Creating Synthetic Baseline Populations. *Transportation Research Part A,* Vol. 30, No. 6, 1996, pp. 415–429.
2. Roorda, M. J., E. J. Miller, and K. Habib. Validation of TASHA: A 24-Hour Activity Scheduling Microsimulation Model. Presented at 86th Annual Meeting of the Transportation Research Board, Washington, D.C., 2007.
3. Bhat, C. R., J. Y. Guo, S. Srinivasan, and A. Sivakumar. Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1894,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 57–74.
4. Yagi, S., and A. Mohammadian. Modeling Daily Activity–Travel Tour Patterns Incorporating Activity Scheduling Decision Rules. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2076,* Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 123–131.
5. Frick, M., and K. W. Axhausen. Generating Synthetic Populations Using IPF and Monte Carlo Techniques: Some New Results. Presented at 4th Swiss Transport Research Conference, Monte Veritá, Switzerland, March 25–26, 2004.

6. Wheaton, W. D., J. C. Cajka, B. M. Chasteen, D. K. Wagener, P. C. Cooley, L. Ganapathi, et al. *Synthesized Population Databases: A U.S. Geospatial Database for Agent-Based Models.* RTI Press Publication No. MR-0010-0905. RTI Press, Research Triangle Park, N.C., 2009.
7. Ryan, J., H. Moah, and P. Kanaroglou. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. *Geographical Analysis,* Vol. 41, No. 2, 2009, pp. 181–203.
8. Deming, W. E., and F. F. Stephan. On a Least Squares Adjustment of a Sampled Frequency When the Expected Marginal Totals Are Known. *Annals of Mathematical Statistics,* Vol. 11, 1940, pp. 427–444.
9. Voas, D., and P. Williamson. An Evaluation of the Combinatorial Optimization Approach to the Creation of Synthetic Microdata. *International Journal of Population Geography,* Vol. 6, No. 5, 2000, pp. 349–366.
10. Arentze, T., H. J. P. Timmermans, and F. Hofman. Creating Synthetic Household Populations: Problems and Approach. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2014,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 85–91.
11. Hobeika, A. *TRANSIMS Fundamentals: Chapter 3 Population Synthesizer, Technical Report.* U.S. Department of Transportation. Washington, D.C., July 2005. http://tmip.fhwa.dot.gov/transims/transims_fundamentals/ch3.pdf. Accessed Aug. 1, 2007.
12. Auld, J. A., A. Mohammadian, and K. Wies. Population Synthesis with Subregion-Level Control Variable Aggregation. *ASCE Journal of Transportation Engineering,* Vol. 135, No. 9, 2009.
13. Auld, J. A., and A. Mohammadian. Framework for the Development of the Agent-Based Dynamic Activity Planning and Travel Scheduling Model. *Transportation Letters: The International Journal of Transportation Research,* Vol. 1, No. 3, 2009, pp. 243–253.
14. Ye, X., K. C. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
15. Pritchard, D. R., and E. J. Miller. Advances in Agent Population Synthesis and Application in an Integrated Land Use and Transportation Model. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
16. Guo, J. Y., and C. R. Bhat. Population Synthesis for Microsimulating Travel Behavior. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2014,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 92–101.
17. Voas, D., and P. Williamson. Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical and Environmental Modeling,* Vol. 5, No. 2, 2001, pp. 177–200.
18. Bowman, J. L., and G. Rousseau. Validation of the Atlanta Population Synthesizer. Prepared for TRB Conference on Innovations in Travel Modeling, Austin, Tex., May 21–23, 2006.

The Transportation Demand Forecasting Committee peer-reviewed this paper.