POPULATION SYNTHESIS FOR MICROSIMULATING TRAVEL BEHAVIOR

Jessica Y. Guo*
Department of Civil and Environmental Engineering
University of Wisconsin – Madison
U.S.A.
Phone: 1-608-8901064
Fax: 1-608-2625199
E-mail: jyguo@wisc.edu

Chandra R. Bhat
Department of Civil, Architectural and Environmental Engineering
University of Texas - Austin
U.S.A.
Phone: 1-512-4714535
Fax: 1-512-4758744
E-mail: bhat@mail.utexas.edu

** corresponding author*

Abstract

For the purpose of activity-based travel demand forecasting, the representativeness of the base year synthetic population is critical to the accuracy of subsequent simulation outcomes. To date, the conventional approach for synthesizing the base year population is based on the methodology first developed by Beckman *et al.* (1996). In this paper, we discuss two issues associated with this conventional approach. The first issue is often termed as the zero-cell-value problem, and the second issue is related to the inability to control for statistical distributions of both household and individual-level attributes. We then present a new population synthesis procedure that addresses the limitations of the conventional approach. The new procedure is implemented into an operational software system and is used to generate synthetic populations for the Dallas/Fort-Worth area in Texas. Our validation results show that, compared to the conventional approach, the new procedure produces a synthetic population that more closely represents the true population.

Keywords

Synthetic population, Iterative proportional fitting, Microsimulation, Activity-based travel analysis

Main Text: 5803 words + 2 tables + 4 figures (equivalent of 7303 words)
Appendix: 292 words + 5 figures

1 Introduction

Microsimulation is a mechanism for reproducing or forecasting the state of a dynamic, complex, system by simulating the behavior of the individual actors in the system. There has been growing interest in using microsimulation to address policy-relevant issues in several fields. For example, economists have employed microsimulation models of household income structure to analyze tax policies (*e.g.* Creedy *et al*. 2002). Urban and regional scientists have used microsimulation to assess the impacts of employment and welfare policy changes (*e.g.* Martini 1997). Transportation engineers and planners are employing microsimulation, coupled with activity-based travel demand models, to analyze the effects of various demand management policies (*e.g.* Bhat *et al*. 2004, Hensher *et al*. 2004, Los Alamos National Laboratory 2005).

In general, microsimulation involves two major steps: (1) constructing a microdata set representing the characteristics of the decision agents of interest, and (2) simulating the decision agent's behavior of interest to the analyst and updating decision agents' characteristics based on mathematical and/or rule-based models. This paper is concerned with the methodology used to accomplish the first step of microsimulation, often known as population synthesis. For the purpose of activity-based travel demand forecasting, the decision agents to be microsimulated are usually households, and the constituent household members, residing in a study area. Naturally, the representativeness of the synthesized population for the base year of the simulation is critical to the accuracy of the ultimate simulation outcome.

To date, the conventional approach to synthesize base year population is based on a methodology originally developed by Beckman *et al.* (1996). This approach involves integrating aggregate data from one source with disaggregate data from another source. The aggregate data are typically drawn from aggregate census data, such as the Summary Files (SF) of the U.S. and the Small Area Statistics (SAS) file of the U.K.. These data are in the form of one-, two-, or multi-way cross tabulations describing the joint aggregate distribution of salient demographic and socio-economic variables at the household and/or the individual levels. The disaggregate data, on the other hand, usually represent a sample of households with information on the characteristics of each household and each person in it. Examples include the Public-Use Microdata Samples (PUMS) of the U.S. and the Sample of Anonymized Records (SAR) of the U.K.. Beckman *et al.*'s population synthesis approach uses the disaggregate data as "seeds" to create individual population records that are collectively consistent with the cross tabulations provided by the aggregate data. This conventional approach has been incorporated in most deployment initiatives of activity-based travel simulation systems, particularly in the United States.

Most existing population synthesizers based on the conventional approach are application-specific in that they have been developed to create a synthetic population for a fixed combination of variables and for a given geographical area. The lack of re-usability of these population synthesizers implies a need to re-implement a synthesizer whenever the activity-based travel simulation approach is applied to a new study area. This can be rather cumbersome, and can impede the widespread adoption of the activity-based approach. Thus, it is highly desirable to develop a flexible and reusable population synthesizer.

The current study is motivated by the emerging need for a reusable population synthesizer, as well as the very limited advancements in synthesizing methodology since Beckman *et al.*'s original contribution. Specifically, our objectives are twofold. First, we discuss a number of issues underlying the Beckman *et al.* approach and discuss possible solutions to resolve these issues. Second, we describe proposed modifications and enhancements to the Beckman *et al.* approach in the context of designing a flexible and generic population synthesis tool.

The remainder of this paper is organized as follows. Section 2 discusses the conventional approach to solving the population synthesis problem. Section 3 examines a number of issues related to the implementation and application of this conventional approach. Section 4 describes a generic algorithm that we propose for population synthesis. Section 5 presents validation results for our proposed algorithm. Section 6 concludes with summary remarks and a discussion of directions for future research.

2 Conventional Approach

The conventional population synthesis procedure typically starts with identifying the socio-demographic attributes desired of the synthesized households and/or individuals. These are the attributes considered to significantly impact the behavioral outcome of individuals. For the purpose of the subsequent discussion, let the number of attributes desired for the synthesized households be H and denote the attributes by a vector of variables $\boldsymbol{V}=\{V_1,V_2,\ldots,V_H\}$. For example, H can be 2, and the attributes may be V={Household size, Household income}. Similarly, let the number of individual-level attributes be P and denote the attributes by a vector of variables $\boldsymbol{U}=\{U_1,U_2,\ldots,U_P\}$. The variables are typically defined as categorical variables, for example, a 6-way classification of household type or a 7-way classification of race.

As mentioned earlier, the synthesis of socio-demographic attribute values involves integrating an aggregate dataset with a disaggregate dataset. The aggregate dataset comprises a set of cross-tabulations that, at a relatively fine spatial resolution (for example, census blocks), describe the one-, two-, or multi-way distributions of *some*, but not all, of the desired socio-demographic attributes. We refer to these attributes with known distributions as the **control variables** and the spatial units for which the aggregate distribution information is available as the **target areas**. The disaggregate dataset, on the other hand, provides information about all the socio-demographic variables of interest, but only for a sample of households and individuals. The spatial units for which this disaggregate information is available – hereafter referred to as the **seed areas** - are typically larger than the target areas (*e.g.* the PUMS data is available for the Public Use Microdata Areas, or PUMA, which are areas of no less than 100,000 population). For ease in discussion, we assume that each target area t can be uniquely mapped to a single seed area s_t.

The basic population synthesis procedure entails repeating the following steps for each target area t in the study region:

Step 1. Estimate the K-way joint distribution, where K is the number of control variables, such that the resulting distribution (a) satisfies the marginal distributions known about the control variables for t (as informed by the aggregate dataset) and (b) preserves the correlation structure observed in the sample households associated with s_t (from the disaggregate dataset).

Step 2. Select and copy sample households (and their constituent members) from s_t into t so that the resulting joint distribution is consistent with the distribution obtained in Step 1.

Each of these two steps is further discussed below.

2.1 Estimating the Complete Distribution

The problem of estimating a full contingency table (*i.e.* the complete distribution across all control variables), based on known marginal distributions, has been studied since as early as 1940. Deming and Stephan (1940) were the first to apply the now well-known iterative proportional fitting procedure (IPFP) as a way for estimating the cell probabilities p_{ij} in a two-dimensional contingency table, given a sample of n observations in the disaggregate data and known marginal totals $p_{i.}$ and $p_{.j}$ from the aggregate data. The IPFP begins by initializing the cell probabilities with the proportion of observations found in the sample:

$$p_{ij}^{(0)} = \pi_{ij}\text{, where } \pi_{ij} = n_{ij} / n \text{ ....} (1)$$

Each subsequent iteration consists of stepping through the list of marginal distributions and scaling the current cell estimates to make the current table estimate consistent with the marginal distribution (see, for example, Fienberg, 1970, and Beckman *et al.*, 1996, for a detailed discussion of the algorithm). The iterations continue until the relative change in cell values between successive iterations is small. As Mosteller (1968) pointed out, the interaction structure of the initial cell values as defined by the cross product ratios is preserved at each iteration *I*:

$$\frac{n_{ij} n_{hk}}{n_{ik} n_{hj}} = \frac{p_{ij}^{(I)} p_{hk}^{(I)}}{p_{ik}^{(I)} p_{hj}^{(I)}} \text{ ....} (2)$$

Furthermore, according to Ireland and Kullback (1968), the IPFP produces estimates of the p_{ij}'s that minimize the discrimination information:

$$I(p,\pi) = \sum_i \sum_j p_{ij} \ln\left(p_{ij} / \pi_{ij}\right). \text{ ....} (3)$$

In other words, the procedure yields the constrained maximum entropy estimates of the p_{ij}'s, and the resulting contingency table is the one least distinguishable from the contingency table given by the sample (Wong, 1992). The procedure has been shown to converge at the optimal solution and is easily extended to estimating contingency tables of higher number of dimensions (Ireland and Kullback, 1968).

Beckman *et al.* (1996) were the first to apply the IPFP to solve the population synthesis problem. In their paper, they provided a detailed example illustrating how the procedure may be applied to generate the full multi-way distribution for a set of household-level control variables $\boldsymbol{V}'$, where $\boldsymbol{V}' \subseteq \boldsymbol{V}$, leaving all the individual-level socio-demographic variables in $\boldsymbol{U}$ uncontrolled. Values for the uncontrolled variables are directly 'copied' from sample households and individuals. The sample data that provided the observed correlation structure is the PUMS and the marginal totals are extracted from a number of census summary tables. This IPFP-based procedure developed by Beckman *et al.* has since been used in most activity and travel simulation studies to date.

2.2 Selecting Sample Households

The *K*-way joint distribution resulting from the IPFP gives the relative proportion of each homogenous grouping of households in *t*. In Beckman *et al.* (1996), the table of proportions (which are values between 0 and 1) is then converted into a table of integer values representing the expected numbers of households to be created for each demographic group. The conversion (sometimes referred to as *integerization*) of the multi-way distribution table can be achieved by multiplying the proportions by the total number of household expected for the target area. The values are then rounded up (or down) to the next larger (or smaller) integer values. The rounding inevitably introduces deviations from the original correlation structure and marginal totals. Subsequent adjustments to the rounded values are usually required if the resulting marginal totals are to be perfectly consistent with the original marginal totals.

Once the expected number of households in each demographic group is determined, each sample household associated with the corresponding seed area s_t is assigned with a probability of being selected into the target area *t*. The probability is typically a function of the sample weight associated with the household record, the expected number of households to be generated for the given demographic group, and the number of other households in the sample that belong to the same demographic group. Based on the probability values, sample households are then randomly drawn either with or without replacement using a Monte Carlo procedure. The random draw continues until the expected number of households has been obtained for each demographic group.

When a sample household is selected for the target area, its attribute values for the controlled variables as well as the uncontrolled, but desired, variables are used to create a synthetic household for the target area. Values for the person-level variables are also used to create the synthetic individuals that make up the household.

3 Implementation and Application Issues

In this section, we discuss two issues that arise from implementing and applying the basic algorithm described in the preceding section. If left unaddressed, these issues may significantly diminish the representativeness of the synthesized population.

3.1 Incorrect zero cell values

The first issue is inherent to the process of integrating aggregate data with sample data, and the problem occurs when the demographic distribution derived from the sample data is not consistent with the distribution expected of the population. Specifically, consider a demographic group that is present in the population as represented by the aggregate data but not represented in the sample of the disaggregate data. The cell in the contingency table that corresponds to this demographic group will take an initial value of zero and will remain zero throughout the IPFP iterations. However, such 'incorrect' zero cell values will prevent the iterations from ever reaching the given marginal totals of the aggregate data. Consequently, the IPFP will fail to converge.

There are a number of ways to get around this issue. The first, and perhaps the easiest, approach is to terminate the IPFP when a pre-specified maximum number of iterations have been reached. Although this implies that the procedure does not exit at proper convergence, the resulting contingency table estimates usually satisfy the marginal totals reasonably well with a large enough maximum-iteration threshold value. The second approach for overcoming the issue involves replacing the incorrect zero cell values with small, positive, values (*e.g.* 0.01). This 'tweaking' – as referred to in Beckman *et al.* (1996) – allows the IPFP to converge at the expense of an arbitrarily introduced bias in the underlying correlation structure. However, according to Beckman *et al.* (1996), who evaluated and compared the tweaking approach against the maximum threshold approach, the former did not outperform the latter and was therefore not recommended. The third approach is to reduce the occurrences of 'incorrect' zero cell values by appropriately defining the variable class intervals. For example, compared to a 12-way classification of household type, a more aggregate 6-way classification will provide a less sparse contingency table, which is likely to contain fewer incorrect zero cell values. This more aggregate classification, however, results in a coarser representation of household types throughout the microsimulation process. In view of this trade-off between the accuracy of the IPFP results and the level of detail in population representation, one needs to examine the statistical distributions underlying the data and define the control variables accordingly. This process would be aided with a population synthesizer that allows the user to explore and modify his/her choice of control variables without making any code-level changes.

3.2 Individual-Level Variables Uncontrolled

The second issue relating to Beckman *et al.*'s approach arises from the fact that the approach can control for either household-level or person-level variables, but not both. This is because Step 2 of the algorithm is designed to account for only one contingency table; yet the data available for population synthesis typically do not support the estimation of a single contingency table that represents the joint distribution of household-level and individual-level attributes. For example, the U.S. census SF1 provides separate tables on the marginal distribution of household size – a household-level attribute – and the marginal distribution of gender – a person-level attribute. Since the household size table contains household counts and the gender table contains person counts, it is conceptually infeasible to construct a two-dimensional contingency table of household size by gender. This is why past efforts of population synthesis have accounted for only the household-level contingency table during the sample household selection stage, leaving the individual-level variables uncontrolled. This means, for example, the resulting gender distribution in the synthesized population is likely to deviate from the known gender distribution given by SF1. The deviation could severely affect the accuracy of the subsequent microsimulation outcome. Thus, a methodology that controls both the household- and individual-level distributions is needed.

4 Proposed Algorithm and Implementation Considerations

In this section, we describe a population synthesizing system that has been developed in view of the two issues discussed in Section 3. This system features:

1. Generic data structures and accompanying functions to help circumvent the incorrect zero cell value problem by providing users the capability to specify their choice of control variables and class definitions at run-time; and
2. An overall algorithm modified from Beckman *et al.*'s basic algorithm to allow simultaneous control for household- and person-level variables.

These two aspects of the proposed system are discussed in detail below.

4.1 Data structures and operations

The proposed population synthesis system was designed using the object-oriented programming (OOP) paradigm, which promotes highly modular computer code and facilitates direct mapping from real-world objects to programming components. The core data objects in our system design are of three types: Variable, Table, and Tables. A *Variable* object represents a control variable and can either be a household-level or person-level socio-demographic variable. A variable is characterized by a text label, an ID, and its size (*i.e.* the number of values it can possibly take). A *Table* object represents an aggregate cross-tabulation that provides the marginal distributions of the control variables. A table is characterized by an array of variable IDs that define the cross-tabulation and a variable-size multi-dimensional matrix of cell values that describes the actual tabulation. A *Tables* object represents a collection of the *Table* objects that need to be merged to form the complete contingency table.

A number of implementation details are worth noting here. First, we allow the attribute values characterizing the *Variable*, *Table*, and *Tables* objects to be determined at run-time as opposed to being hard-coded in the program. This approach provides flexibility to experiment with different choices of control variables and/or reusability to apply the system to an entirely different empirical context. Second, we develop a recursive algorithm that wraps around the IPFP to merge any given two tables with common variables (see Figure 1). A *recursive algorithm* is an algorithm that solves a problem by calling itself with "smaller" input values and that has a base part to compute the solution for the smallest input without making any calls to itself. In Figure 1, the lines of code between the IF and ELSE statements form the recursive part of the algorithm that ‘strips off’ variables common to two input tables. The lines between the ELSE and END-IF statements form the base part where the IPFP is performed on two tables that have no variables in common. Third, the synthesizer is built with an error reporting mechanism that tracks any non-convergence problems encountered during the IPFP and informs the user of the locations of any incorrect zero cell values.

```
PROCEDURE MergeTables
    IF Table1 and Table2 have a variable $V_k$ in common, THEN
        Initialize NewTable to an empty table
        FOR each value (denoted as $i$) of $V_k$
            Extract Table1' from Table1 that satisfies $V_k=i$
            Extract Table2' from Table2 that satisfies $V_k=i$
            CALL MergeTables with Table1' and Table2' RETURNING NewTable'
            Append NewTable' to NewTable
        END-FOR
    ELSE
        DETERMINE NewTable by performing IPFP on Table1 and Table2
        RETURN NewTable
    END-IF
END-PROCEDURE
```

FIGURE 1 A recursive procedure for merging any two contingency tables with common variables.

4.2 Proposed Algorithm

Figure 2 provides an overview of the algorithm that we have developed for creating the synthetic population for a given target area. The algorithm includes a number of major steps: (1) determine the household-level multi-way distribution, (2) determine the individual-level multi-way distribution, (3) initialize the household- and individual-level counts, (4) compute selection probabilities, (5) select a sample household, (6) check household desirability, (7) add the selected households to the target area, and (8) update the household- and individual-level counts. We discuss each of these steps is in turn below. An example is also provided in the Appendix to demonstrate the application of our proposed algorithm.

4.2.1 Determine Household-Level Multi-Way Distribution

Given the aggregate (*e.g.* U.S. Census Summary Tables) and disaggregate (*e.g.* U.S. PUMS data) input data, this step creates the full multi-way distribution across all the household-level control variables using the IPFP-based recursive procedure outlined in Figure 1. We denote each cell in the resulting household-level multi-way distribution by HH[v_1, v_2, …, v_k, …], where the index v_k is the value of the k^{th} household-level controlled variable, $v_k = 1, \ldots, M_k$. HH[v_1, v_2, …, v_k, …] gives the expected number of households with attribute values of (v_1, v_2, …, v_k, …) in the target area.

4.2.2 Determine Individual-Level Multi-Way Distribution

This step creates the full multi-way distribution across all the individual-level controlled attributes, also using the procedure presented in Figure 1. We denote each cell in the resulting individual-level multi-way distribution by POP[v_1, v_2, …, v_l, …], where the index v_l denotes the value of the l^{th} individual-level variable, $v_l = 1, \ldots, N_l$. POP[v_1, v_2, …, v_l, …] thus gives the expected number of individuals with attribute values of (v_1, v_2, …, v_l, …) in the target area. It should be noted that the cell values in both HH and POP will be used as they are without being rounded to integer values.

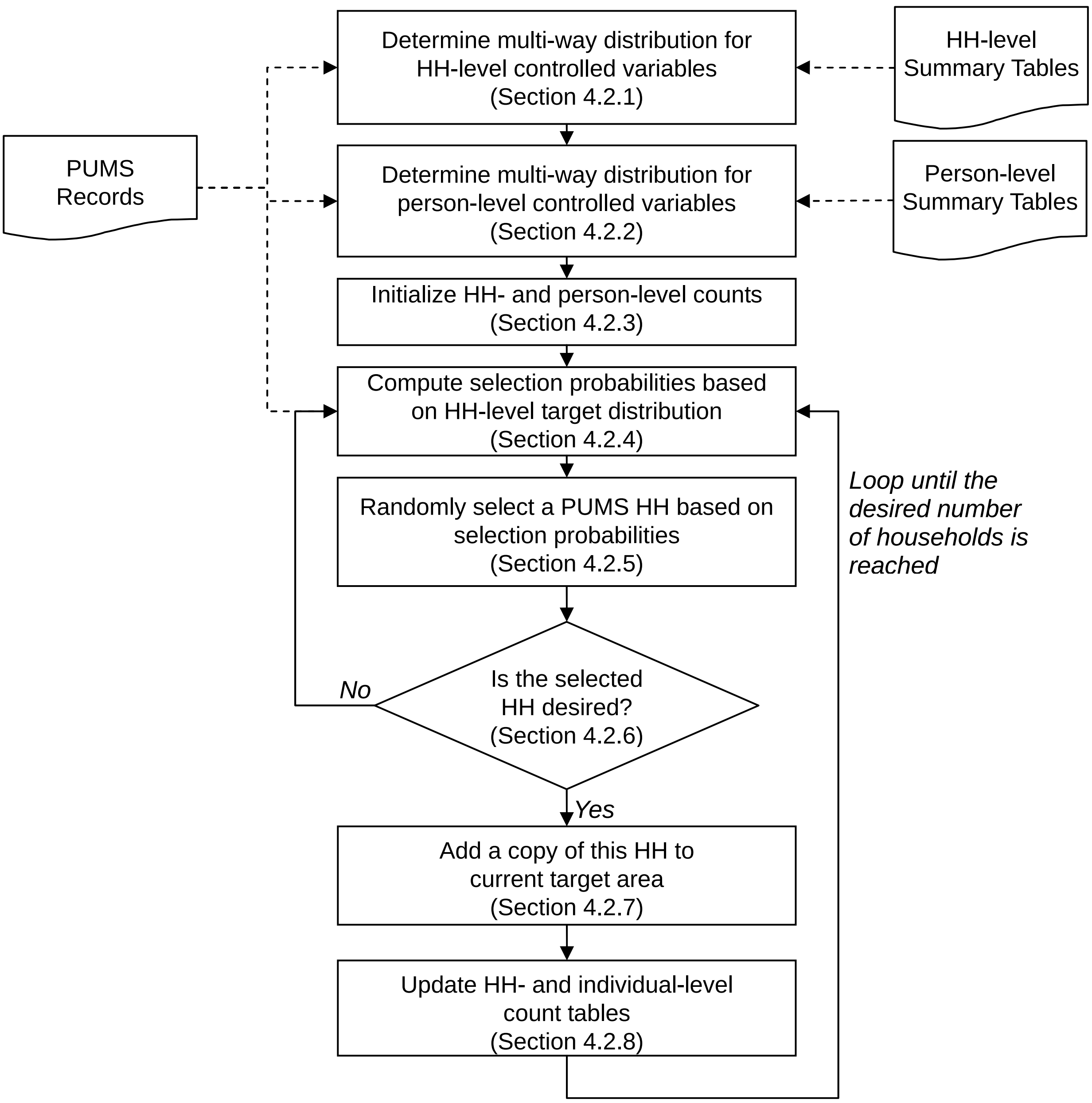

Determine multi-way distribution for HH-level controlled variables (Section 4.2.1)
HH-level Summary Tables
PUMS Records
Determine multi-way distribution for person-level controlled variables (Section 4.2.2)
Person-level Summary Tables
Initialize HH- and person-level counts (Section 4.2.3)
Compute selection probabilities based on HH-level target distribution (Section 4.2.4)
Randomly select a PUMS HH based on selection probabilities (Section 4.2.5)
Loop until the desired number of households is reached
No
Is the selected HH desired? (Section 4.2.6)
Yes
Add a copy of this HH to current target area (Section 4.2.7)
Update HH- and individual-level count tables (Section 4.2.8)

FIGURE 2 Overview of the proposed population synthesis algorithm.

4.2.3 Initialize Household- and Person-Level Counts

Two multi-way tables, *HHI* and, *POPI* are used to keep track of the numbers of households and individuals belonging to each demographic group that have been selected into the target area during the iterative process. At the start of the process, the cell values in the two tables are initialized to zero to reflect the fact that no households and individuals have been created for the target area. During subsequent iterations, these cell values will be updated as households and individuals are selected into the target area (see Section 4.2.8 for further discussion of the updating procedure).

4.2.4 Compute Household Selection Probabilities

Given the target distribution (HH) and the current distribution (*HHI*) of households already selected into the target area, each PUMS sample household in the corresponding seed area is assigned with a probability of being selected into the target area in the current iteration. The probability of household *i* being selected is computed by

$$P_i = \frac{w_i}{\sum_j w_j \cdot Y^j_{v_1,v_2,\cdots,v_k,\cdots}} \cdot \frac{\mathrm{HH}[v_1,v_2,\cdots,v_k,\cdots] - \mathrm{HH\,I}[v_1,v_2,\cdots,v_k,\cdots]}{\sum_{u_1,u_2,\cdots,u_k,\cdots}\left(\mathrm{HH}[u_1,u_2,\cdots,u_k,\cdots] - \mathrm{HH\,I}[u_1,u_2,\cdots,u_k,\cdots]\right)} \quad \ldots\ldots\ldots\ldots(4)$$

In the above equation, w_i is the PUMS weight associated with household *i*. The vector (v_1, v_2, …, v_k, …) reflects the characteristics of household *i*. $Y^j_{v_1,v_2,\cdots,v_k,\cdots}$ takes a value of 1 if the *j*th household is characterized by(v_1, v_2, …, v_k, …) (*i.e.*, the same as the *i*th household), and a value of 0 otherwise. The equation implies that the selection probability of a sample household decreases as more households from the same demographic group are selected into the target area.

4.2.5 Randomly Select a Household

Based on the probabilities computed in the previous step, a household is randomly drawn from the pool of sample households to be considered for "cloning" and added to the population for the target area.

4.2.6 Check Household Desirability

Given a randomly selected household characterized by (v_1,v_2,…, v_k,…), we will add a copy of this household into the population for the target area if the following conditions hold:

1. The number of such households already selected into the target area (as given by $\mathrm{HHI}[v_1,v_2,\cdots,v_k,\cdots]$) is lower than a pre-specified maximum threshold. Ideally, this threshold should be set to the target value given by $\mathrm{HH}[v_1,v_2,\cdots,v_k,\cdots]$ so that the number of households characterized by (v_1,v_2,…, v_k,…) is never higher than desired. However, such a condition may be undesirable for at least two reasons. First, when incorrect zero cell values are found for certain demographic groups, the target total number of households in the area would never be met unless households of other demographic groups are allowed to be over-selected. Second, since the dual goals of satisfying the household-level target distribution and satisfying the individual-level

target distribution may be conflicting in nature, fitting the synthetic population perfectly to the household-level target distribution may prevent the individual-level distribution from being satisfied to any acceptable extent. Therefore, in the proposed algorithm, we allow the threshold values to exceed their respective target values by a user-specified percentage, hereafter referred to as the percentage deviation from target size (PDTS).

2. For each person in the household, the number of such individuals already selected into the target area (as given by $\text{POPI}[v_1, v_2, \cdots, v_l, \cdots]$) is lower than a pre-specified maximum threshold. The threshold values are specified as (1+PDTS) of the corresponding target cell value $\text{POP}[v_1, v_2, \cdots, v_l, \cdots]$.

If any of the above conditions fails, then the household is removed from the consideration set so that it will never be selected again. The selection probabilities of the households remaining in the consideration set are then updated before the next household is randomly selected.

4.2.7 Add Household

If the selected household satisfies the conditions described in Section 4.2.6, then the household is added to the pool of the synthetic population for the target area. As part of this step, the household sample weight is decreased by one to implement the 'random draw without replacement' strategy.

4.2.8 Update Household- and Individual-Level Counts

The cell values in the count tables $\text{HHI}[v_1, v_2, \cdots, v_k, \cdots]$ and $\text{POPI}[v_1, v_2, \cdots, v_l, \cdots]$ that correspond to the selected household and its individuals are incremented accordingly to reflect the reduced desirability of such a household and individuals in subsequent iterations.

5 Validation

The proposed system was used to generate a synthetic population for the Dallas/Fort-Worth Metropolitan Area in Texas. Census block groups and PUMA were used as the target areas and seed areas, respectively. The aggregate data that provide the marginal distributions come from the 2000 U.S. Census SF1 Tables P20, P26, P7, and an aggregate version of P12. Table P20 is defined by four household-level variables: HHR_FAM, HH_TYPE, HH_CHILDREN, and HHR_AGE; P26 is defined by two household-level variables: HH_FAM and HH_SIZE; P7 is defined by the single individual-level variable P_RACE; and P12, which is originally defined by P_GENDER and a twenty-three-way classification of age, is aggregated along the age dimension to form a two-by-ten table of gender by age (P_AGE).

The variables that define these tables are thus considered as controlled variables (see Table 1 for variable definitions). The disaggregate data that inform the correlation structure between the control variables and provide original copies of subsequently synthesized households is the 2000 PUMS data.

5.1 Verification of IPFP

Out of the 3372 target areas, 388 and 151 of them had the zero-cell value problem that led to improper convergence of the household-level and person-level contingency tables, respectively. These problematic target areas are identified by the discrepancy found between the marginal totals in the estimated contingency tables and the control totals given by the summary tables. The discrepancies were found for marginal totals corresponding to the following dimensions: (HH_FAM=1, HH_TYPE=5, HH_CHILDREN=0, HHR_AGE=0), (HH_FAM=1, HH_SIZE=4, 5, 6), and (RACE=4) – that is, the estimated sizes of these population groups as give by the IPFP are zero, yet the actual sizes as given by the aggregate data are greater than zero. Not surprisingly, these are demographic groups relatively smaller than other groups (*e.g.* non-family households that have no children and whose householder is under 65 year of age and does not live alone) and, as a result, have not been represented in the PUMS data for the problematic target areas. The magnitudes of the discrepancy vary for different target areas and for different marginal totals. For example, the discrepancy found in the marginal total for RACE=4 (*i.e.* number of native Hawaiian and other pacific islander alone individuals) ranges from 1 to 10 (see Figure 3 for the distribution of discrepancies).

TABLE 1 Definition of the Control Variables Used in the Validation Study

Variable Label	Size	Value	Value Description
HH_FAM	2	0	family
		1	non-family
HH_TYPE	6	0	not a household (vacant or GQ)
		1	family: married couple
		2	family: male householder, no wife
		3	family: female householder, no husband
		4	non-family: householder alone
		5	non-family: householder not alone
HH_CHILDREN	2	0	no own children under 18
		1	own children under 18 years
HHR_AGE	2	0	15-64
		1	65 and over
HH_SIZE	7	0	1-person
		1	2-person
		2	3-person
		3	4-person
		4	5-person
		5	6-person
		6	7-or-more person
P_RACE	7	0	white alone
		1	black African-American alone
		2	American-Indian and Alaska Native alone
		3	Asian alone
		4	Native Hawaiian and other Pacific Islander alone
		5	Some other race alone
		6	Two or more races
P_GENDER	2	0	male
		1	female
P_AGE	10	0	Under 5 years
		1	5 to 14 years
		2	15 to 24 years
		3	25 to 34 years
		4	35 to 44 years
		5	45 to 54 years
		6	55 to 64 years
		7	65 to 74 years
		8	75 to 84 years
		9	85 and more

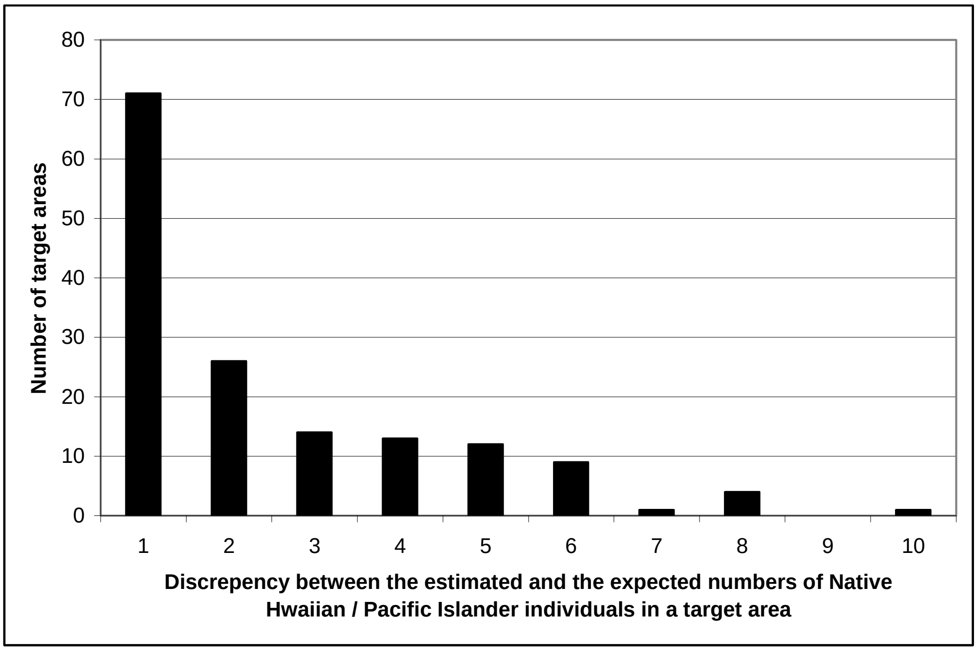


FIGURE 3 The discrepancy found in the number of Native Hawaiian / Pacific Islander per target area.

5.2 Evaluation of Selection Procedures

Four alternative implementations of the household selection procedure were evaluated and compared. In the first implementation, households are selected into the target areas without any assessment of how well they satisfy the individual-level contingency table, POP[v_1, v_2, …, v_l, …]. This represents the conventional approach of not controlling for individual-level variables. The second, third, and forth implementations correspond to setting the PDTS (defined in Section 4.2.6) to 0, 5, and 10 for both the household- and individual-level distributions.

In order to evaluate the performance of the selection procedures independently from that of the IPFP, we focus on the synthetic population generated for 62 census block groups in the Tarrant County that do not suffer from the zero cell value problem. The alternative selection procedures are compared based on the percentage difference between the expected size of each distinct population group and the corresponding size found in the synthetic population. The percentage difference (PD) for cell $(v_1, v_2, \cdots, v_k, \cdots)$ and target area t are formally defined as:

$$PD_{HH,t}(v_1, v_2, \cdots, v_k, \cdots) = \frac{\mathrm{HHI}_t[v_1, v_2, \cdots, v_k, \cdots] - \mathrm{HH}_t[v_1, v_2, \cdots, v_k, \cdots]}{\mathrm{HH}_t[v_1, v_2, \cdots, v_k, \cdots]} \quad \text{, and}$$

$$PD_{POP,t}(v_1, v_2, \cdots, v_l, \cdots) = \frac{\mathrm{POPI}_t[v_1, v_2, \cdots, v_l, \cdots] - \mathrm{POP}_t[v_1, v_2, \cdots, v_l, \cdots]}{\mathrm{POP}_t[v_1, v_2, \cdots, v_l, \cdots]} \text{.....................}(5)$$

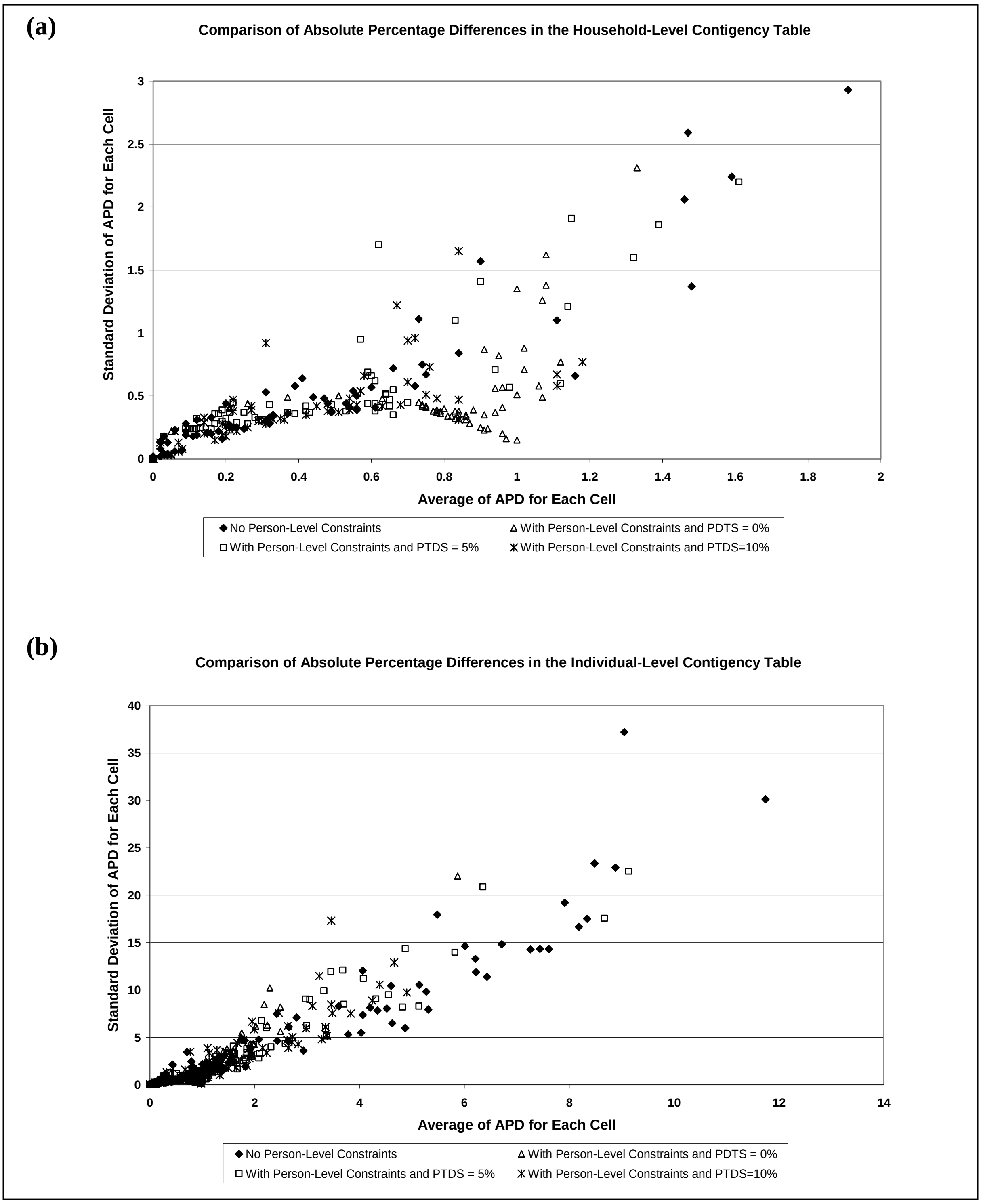
In the first part of the validation exercise, we examine how the *magnitudes* of the percentage differences vary for each distinctive population groups. This is achieved by first computing the absolute percentage differences (APD) for each cell and each target area:

$$APD_{HH,t}(v_1, v_2, \cdots, v_k, \cdots) = \left| PD_{HH,t}(v_1, v_2, \cdots, v_k, \cdots) \right| \text{ , and}$$

$$APD_{POP,t}(v_1, v_2, \cdots, v_l, \cdots) = \left| PD_{POP,t}(v_1, v_2, \cdots, v_l, \cdots) \right| \text{..}(6)$$

We then compute the average and standard deviation using the 62 APD values (one for each target area) that correspond to each cell. The averages and standard deviations for the 336 cells in the household-level contingency table (corresponding to the number of combinations across household control variables in Table 1; $336 = 2 \times 6 \times 2 \times 2 \times 7$) and for the 140 cells in the individual-level contingency table (corresponding to the number of combinations across individual-level control variables in Table 1; $140 = 7 \times 2 \times 10$) are plotted in Figure 4 and Figure 5, respectively. In both figures, each data point corresponds to the average/standard deviation combinations across the 62 target areas for each table cell; and each data series corresponds to one of the four alternative household selection procedures. The data points that are located in the top-right corner of the charts represent the cells (*i.e.* demographic groups) that are difficult to fit. These are typically cells with target values between 0 and 1 (for example, one of such cells represents family households of size 3 with male householder 65 years or older, no wife, and no children under 18). The process of selecting 0 or 1 household/individual belonging to these demographic groups into the synthetic population inevitably results in deviations from the corresponding cell target values. These deviations in turn result in relatively large APD values.

As shown in Figure 4(a), the four household selection procedures are comparable in their household-level APD distributions. This is because all four procedures take the household-level targets into account. The procedure that considers both the household- and individual-level distributions with PTDS=10% results in a slightly less dispersed APD values. In comparison, the differences among the alternative procedures are more pronounced in Figure 4(b). Without taking the individual-level target distributions into consideration, the conventional approach leads to the widest spread of individual-level APD values as expected. On the other hand, when individual-level target distributions are considered during the selection process, the resulting APD values are smaller and less dispersed as the PTDS increases. The charts shown in Figure 4(a) and 4(b) together suggest that the proposed algorithm is capable of producing synthetic populations that better represent the household and individual population subgroups comprising the true population.



(a)

Comparison of Absolute Percentage Differences in the Household-Level Contigency Table

(b)

Comparison of Absolute Percentage Differences in the Individual-Level Contigency Table

FIGURE 4 Comparison of the absolute percentage differences in (a) the household-level contingency table and (b) the individual-level contingency table across the four alternative household selection procedures.

In the second part of our validation exercise, we are interested in how the alternative selection procedures compare in the overall. For each selection procedure, the APD values computed by Equation (6) are averaged across all target areas and all cells to arrive at two overall average APD values: $AAPD_{HH}$ for the household-level and $AAPD_{POP}$ for the individual-level. The four pairs of AAPD values are summarized in Table 2. The selection procedures appear comparable in terms of $AAPD_{HH}$. The procedure with a PTDS value of 0 has the highest $AAPD_{HH}$ value of all, mostly due to the restrictive nature of the selection criteria. Not surprisingly, the conventional procedure of not accounting for individual-level distributions yields the worst $AAPD_{POP}$. The procedure with 10% PTDS outperforms the other three procedures in terms of $AAPD_{HH}$ and $AAPD_{POP}$.

TABLE 2 Average Absolute Percentage Differences (AAPD) Computed for the Alternative Selection Procedures

Selection Procedure	Individual-level distribution considered	PTDS value	$AAPD_{HH}$	$AAPD_{POP}$
1	No	N/A	29.6	267.3
2	Yes	0%	50.2	138.4
3	Yes	5%	30.5	164.0
4	Yes	10%	24.0	125.3

6 Summary and Conclusions

A new algorithm for population synthesis has been presented in this paper. The algorithm represents an extension of the conventional approach (originally developed by Beckman *et al.* in 1996) by controlling for statistical distributions defined by both household- and individual-level variables. Through generic data structures and operators, our implementation allows the user to adjust the choice of control variables and the class definition of these variables at run-time. This flexibility is especially desirable when dealing with the incorrect-zero-cell-value problem and when the population synthesis exercise is to be performed for different study areas. It should be noted that, although our particular application context of interest is activity-based travel simulation, the discussion and the algorithm presented in this paper are relevant to microsimulation in other fields of study.

Our validation results show that the proposed algorithm is capable of producing synthetic populations closer to the true population compared to the conventional approach. The performance of the proposed algorithm, however, depends on the PDTS value used. A higher value of PDTS (10%) appears to strike a better balance at satisfying both the household- and individual-level multi-way distributions than lower values of PDTS (0% and 5%). Further validation analysis is needed to better understand the sensitivity of the algorithm's performance on PDTS values and to identify ways of selecting the most appropriate PDTS value. Investigation is also underway to explore other ways of formulating and solving the population synthesis problem as a constrained optimization problem, where the constraints represent the selection of sample households to meet the desired sizes of population subgroups.

7 Acknowledgements

The authors would like to thank Sanketh Indarapu for his assistance in implementing the proposed population synthesis algorithm, Ipek Sener for helping with the validation analysis, and Lisa Macias for typesetting the manuscript.

8 References

Beckman, R.J., Baggerly, K.A., and McKay, M.D., 1996. Creating synthetic baseline populations. *Transportation Research Part A*, 30(6), 415-429.

Bhat, C.R., Guo, J.Y., Srinivasan, S., Sivakumar, A., 2004. Comprehensive econometric microsimulator for daily activity-travel patterns, *Transportation Research Record*, 1894, 57-66.

Creedy, J., Duncan, A.S., Harris, M., Scutella, R., 2002. *Microsimulation Modelling of Taxation and The Labour Market: The Melbourne Institute Tax and Transfer Simulator*. Cheltenham: Edward Elgar.

Deming, W.E. and Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11, 427-444.

Fienberg, S.E., 1970. An iterative procedure for estimation in contingency tables. *Annals of Mathematical Statistics*, 41, 907-917.

Hensher, D.A., Stopher, P.R., Bullock, P., Ton, T., 2004. TRESIS (transport and environmental strategy impact simulator): Application to a case study in Sydney, presentation at the *83rd Annual Meeting, Transportation Research Board*, Washington, D. C., January 11-15.

Ireland, C.T. and Kullback, S., 1968. Contingency tables with given marginals. *Biometrika*, 55(1), 179-188.

Los Alamos National Laboratory, 2005. TRANSIMS. [http://www.ccs.lanl.gov/transims/index.shtml]

Martini, A., 1997. Microsimulation models and labor supply responses to welfare reforms, *Policy Studies Journal*, 25(1), 39-58.

Mosteller, F., 1968. Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63, 1-28.

Wong, D.W.S., 1992. The reliability of using the iterative proportional fitting procedure. *Professional Geographer,* 44(3), 340-348.

Appendix

For the purpose of illustrating the population synthesis algorithm presented in Section 4.2, we consider a target area of 20 households and 49 people. Household type (HH_FAM) and household size (HH_SIZE) are selected as household-level control variables, while gender (P_GENDER) and race (P_RACE) are selected as individual-level controlled variables. The PUMS sample records for the corresponding seed area are listed in Figure A.1. Based on the sample records and the marginal distributions of the controlled variables, we first determine the complete household- and individual-level multi-way distribution tables, denoted as HH[HH_FAM, HH_SIZE] and POP[P_GENDER, P_RACE] respectively (this corresponds to the steps described in Section 4.2.1 and Section 4.2.2). Both tables are shown in Figure A.2. The next step is to set up and initialize the household- and individual-level count tables, denoted as HHI[HH_FAM, HH_SIZE] and POPI[P_GENDER, P_RACE] respectively (this step corresponds to Section 4.2.3). As shown in Figure A.3, both tables are filled with values of 0 to reflect the fact that no households have yet been selected into the target area.

A selection probability is then calculated for each sample household based on equation (4) (this step corresponds to Section 4.2.4). These probability values and the corresponding cumulative probabilities are shown in Figure A.4. Next, a household is selected based on a random number draw (this step corresponds to Section 4.2.5). With a random value of 0.635, the household with SERIALNO = 13687 is selected. Since the household satisfies both the household level selection condition (HHI[1,2]<HH[1,2]) and the individual-level selection condition (POPI[0, 0]<POP[0,0] and POPI[1, 0]<POP[1,0]), the household is now added to the target area (this step corresponds to Section 4.2.6 and Section 4.2.7). The current iteration completes with updating the count tables (see Figure A.5; this step corresponds to Section 4.2.8).

(a) PUMS Housing Unit Record

SERIALNO	HWEIGHT	PERSONS	HHT	Other attributes
2599	6	2	Family: married couple	…
2797	9	3	Family: married couple	…
13687	18	4	Family: married couple	…
21197	18	1	Nonfamily: female living alone	…
15458	6	1	Nonfamily: male living alone	…
24526	6	2	Family: married couple	…
39951	15	2	Family: female householder	…

(b) PUMS Person Record

SERIALNO	PNUM	SEX	RACE	Other attributes
2599	1	male	white alone	…
2599	2	female	white alone	…
2797	1	male	white alone	…
2797	2	female	Some other race alone	…
2797	3	male	Some other race alone	…
13687	1	male	white alone	…
13687	2	female	white alone	…
13687	3	male	white alone	…
13687	4	male	white alone	…
21197	1	female	Black or African American alone	…
15458	1	male	white alone	…
24526	1	male	Asian alone	…
24526	2	female	white alone	…
39951	1	male	Black or African American alone	…
39951	2	male	Black or African American alone	…

Figure A.1 Sample household and person records for the seed area.

(a) HH[H_FAM, H_SIZE]

		H_SIZE (household size)			
		0 (1 person)	1 (2 person)	2 (3 persons or more)	Total
H_FAM (whether household is a family)	0 (No)	3	0	0	3
	1 (Yes)	0	8	9	17
	Total	3	8	9	20

(b) POP[P_GENDER, P_RACE]

		P_RACE			
		0 (white alone)	1 (black alone)	2 (other)	Total
P_GENDER	0 (Male)	16.4	7.6	3	27
	1 (Female)	14.6	7.4	0	22
	Total	31	15	3	49

Figure A.2 Steps 1 and 2: determine household-level and individual-level multi-way distribution tables for the target area.

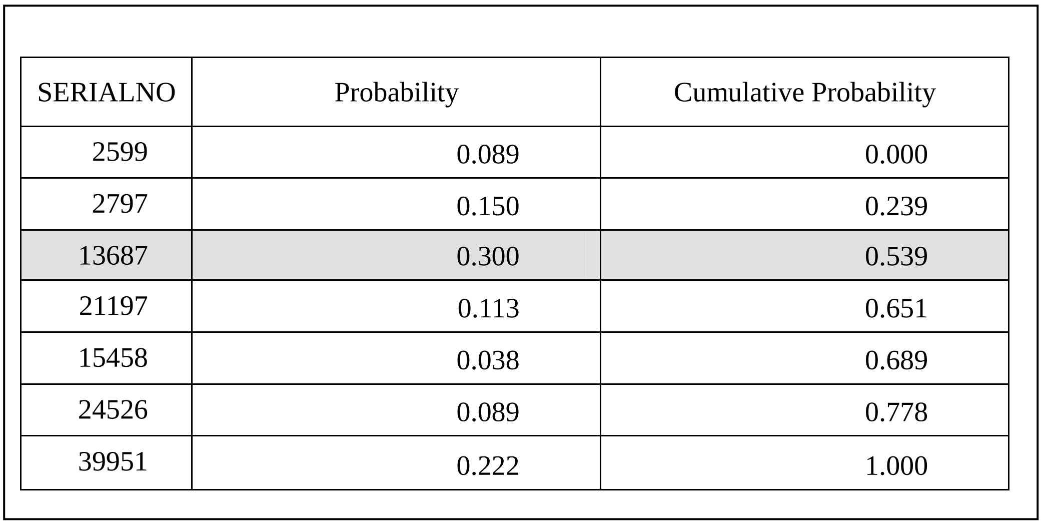
(a) HHI[H_FAM, H_SIZE]

		H_SIZE (household size)			
		0 (1 person)	1 (2 person)	2 (3 persons or more)	Total
H_FAM (whether household is a family)	0 (No)	0	0	0	0
	1 (Yes)	0	0	0	0
	Total	0	0	0	0

(b) POPI[P_GENDER, P_RACE]

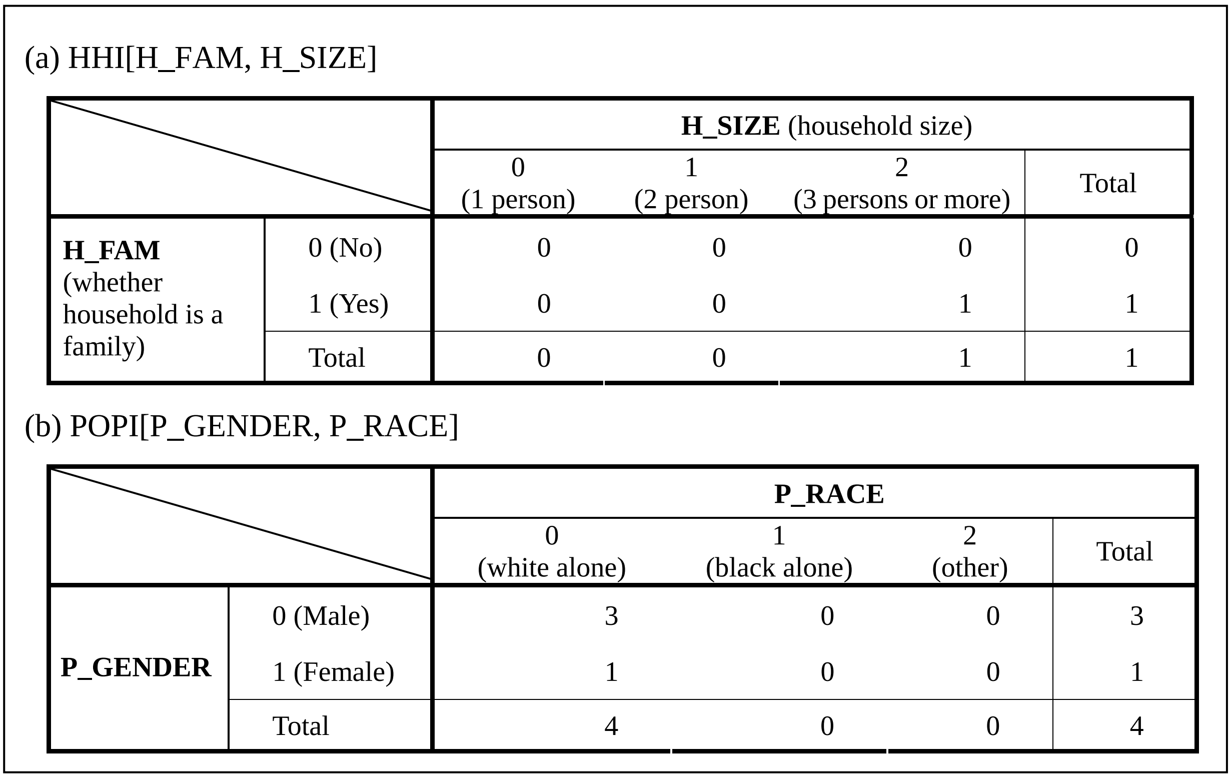
		P_RACE			
		0 (white alone)	1 (black alone)	2 (other)	Total
P_GENDER	0 (Male)	0	0	0	0
	1 (Female)	0	0	0	0
	Total	0	0	0	0

Figure A.3 Step 3: initialize household-level and individual-level count tables.



SERIALNO	Probability	Cumulative Probability
2599	0.089	0.000
2797	0.150	0.239
13687	0.300	0.539
21197	0.113	0.651
15458	0.038	0.689
24526	0.089	0.778
39951	0.222	1.000

Figure A.4 Step 4: compute the household selection probabilities.



(a) HHI[H_FAM, H_SIZE]

		H_SIZE (household size)			
		0 (1 person)	1 (2 person)	2 (3 persons or more)	Total
H_FAM (whether household is a family)	0 (No)	0	0	0	0
	1 (Yes)	0	0	1	1
	Total	0	0	1	1

(b) POPI[P_GENDER, P_RACE]

		P_RACE			
		0 (white alone)	1 (black alone)	2 (other)	Total
P_GENDER	0 (Male)	3	0	0	3
	1 (Female)	1	0	0	1
	Total	4	0	0	4

Figure A.5 Step 8: update the household-level and individual-level count tables.