# Sensitivity of Synthetic Population Generation Procedures in Transportation Models – Implications of Alternative Constraints

by

Rumpa Rani Dey

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2012 by the
Graduate Supervisory Committee:

Ram M. Pendyala, Chair
Soyoung Ahn
Michael S. Mamlouk

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

The growing use of synthetic population, which is a disaggregate representation of the population of an area similar to the real population currently or in the future, has motivated the analysis of its sensitivity in the population generation procedure. New methods in PopGen have enhanced the generation of synthetic populations whereby both household-level and person-level characteristics of interest can be matched in a computationally efficient manner. In the process of set up, population synthesis procedures need sample records for households and persons to match the marginal totals with a specific set of control variables for both the household and person levels, or only the household level, for a specific geographic resolution. In this study, an approach has been taken to analyze the sensitivity by changing and varying this number of controls, with and without taking person controls. The implementation of alternative constraints has been applied on a sample of three hundred block groups in Maricopa County, Arizona. The two datasets that have been used in this study are Census 2000 and a combination of Census 2000 and ACS 2005-2009 dataset. The variation in results for two different rounding methods: arithmetic and bucket rounding have been examined. Finally, the combined sample prepared from the available Census 2000 and ACS 2005-2009 dataset was used to investigate how the results differ when flexibility for drawing households is greater. Study shows that fewer constraints both in household and person levels match the aggregate total population more accurately but could not match distributions of individual attributes. A greater

number of attributes both in household and person levels need to be controlled. Where number of controls is higher, using bucket rounding improves the accuracy of the results in both aggregate and disaggregates level. Using combined sample gives the software more flexibility as well as a rich seed matrix to draw households which generates more accurate synthetic population. Therefore, combined sample is another potential option to improve the accuracy in matching both aggregate and disaggregate level household and person distributions.

# DEDICATION

This thesis is dedicated to my husband, Sanjay Paul. I give my deepest expression of love and appreciation for the encouragement that you gave and the sacrifices you made during this graduate program. Thank you for the support and company during late nights of typing.

# ACKNOWLEDGMENTS

The author takes great pleasure to acknowledge several individuals who contributed and extended their valuable assistance in preparation, continuation and completion of the study.

First and foremost the author wishes to express her utmost gratitude to her supervisor, Prof. Ram M. Pendyala who was abundantly helpful and offered invaluable assistance, support and guidance from the initial to the final level which enabled the author to successfully complete her graduate studies at Arizona State University. Deepest gratitude is also due to the members of the supervisory committee, Dr. Soyoung Ahn and Prof. Michael S. Mamlouk for their continuous encouragement and assistance.

The author considers it an honor to work with Dr. Karthik C. Konduri who has shared valuable insights and has always been a helping hand throughout all the challenging research works and projects.

Finally the author is beholden to Sanjay Paul, without whose help the research would have been incomplete and impossible and who has been her inspiration as she hurdle all the obstacles in accomplishment this thesis work.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

# 1. INTRODUCTION

## 1.1 Background

In the field of activity based analysis and transportation research, the behavioral unit considered is the individual traveler, which leads to microsimulation model systems that are capable of simulating activity travel patterns of individual persons. As these model systems operate at the level of the individual traveler, one needs household and person attribute information for the entire population in a region to calibrate, validate, and apply such model systems. As, such information is not available at the disaggregate level, there has been growing interest in the generation and use of synthetic populations. In these microsimulation models, synthetic populations are initially created and the prediction of outcomes for each unit of the population is done. The results are then aggregated to guide policy related analysis and decision making. The fundamental goal in the development of a population synthesizer is to synthesize the required population as accurately and precisely as possible, for as many variables as possible that are known to determine travel behavior.

Population synthesis, within the context of transportation modeling, land use modeling and similar domains, is the process of creating a representation of a complete, disaggregate population by combining a sample of disaggregate members of a population in a way as to match key distributions for the entire population (Beckman et al. 1996). Population synthesis involves generating a synthetic population by expanding the disaggregate sample data to mirror known

aggregate distributions of household and person variables of interest (Konduri et al. 2010).

Synthetic population can be formed from the random samples by choosing or selecting households and persons from the random samples such that the joint distribution of the critical attributes of interest in the synthetic population match known aggregate distributions of household and person attributes available through a Census (Ye et al. 2009). For example, in the United States, marginal distributions of population characteristics are readily available from Census Summary Files (SF) for any region. For some combinations of critical variables, the Census SF may also directly provide joint distributions against which synthetic population joint distributions can be matched. However, more often than not, such joint distributions of critical attributes of interest are not directly available and the analyst must generate these joint distributions from the known marginal distributions of interest (Ye et al. 2009).

### 1.2 Motivation

There are two primary factors motivating this paper. First, it is desirable to have an analysis to test the accuracy of the synthetic population both in aggregate and disaggregate level by implementing alternative constraints in both household and person levels. Second, how closely and precisely the synthetic population at disaggregate level could be matched with the known marginal distribution with changing or varying possible parameters in the process.

## 1.3 Research Objectives

In the context of explicitly generating a synthetic population for transportation planning and modeling, the objectives of this research are following

- Investigate available meaningful household and person attributes and develop a list of alternative constraints and test the sensitivity of synthetic population generation procedure for different combinations.
- Find out significant improvements of synthetic population at disaggregate level to match the marginal distributions by changing the parameters in the process of generating synthetic population.

This would help the planners get a clear idea about how the accuracy of synthetic population depends on alternative constraints and different input parameters.

## 1.4 Organization of the Thesis

The remainder of the document is organized as follows. The second chapter reviews previous literature on different procedures of synthetic population generation. Chapter three describes the methodology and algorithm used in developing PopGen. Chapter four describes the experimental design part for testing the sensitivity by alternative constraints, alternative rounding procedures and alternative sample data inputs. Chapter five presents different analysis on the synthetic population for the study area generated by different experiments mentioned in chapter four. This chapter also contains the single findings, the linkage of different parameters and detailed results of the analysis. Finally conclusions from the sensitivity test are discussed in chapter five.

## 2. LITERATURE REVIEW

This chapter provides an overview of existing research efforts that deal with different methods for the representation of synthetic population as well as the comparison of various synthesizing procedures developed by different researchers and professionals in the field of transportation planning.

### 2.1 Population Synthesizing Approaches

The Iterative Proportional Fitting (IPF) procedure was first presented by Deming and Stephan in 1940. Deming and Stephan (1940) proposed an iterative least square adjustment method to fill the each cell in the contingency table with constraining row and column totals to match known marginal distributions. The resultant table of data is a joint probability distribution obtained when the probabilities are convergent within an acceptable limit. Beckman et al. (1996) outlined a methodology for the creation of a of a synthetic baseline population of individuals and households which is employed in an activity-based travel demand model (in TRANSIMS) by expanding the original IPF algorithm based on the aggregate Census Summary File (SF3) and disaggregate data (PUMS). That is, while the traditional IPF procedure from Deming and Stephan fits only one block group at a time, Beckman's IPF can simultaneously consider all block groups making up the PUMA (Kao et al. 2012). In its basic formulation, IPF can estimate only one level of aggregation, i.e., it can control either for agent-level or for group-level attributes but not for both simultaneously. Sometimes it suffices to

convert all agent-level attributes into group-level attributes; in this case, IPF can be used on the group-level distribution.

The population synthesizer for the Albatross model, presented by Arentze et al. (2007), is an example of a synthesizer for a European region. The household-level distribution is computed from the person-level distribution in a preprocessing step. They applied the conventional IPF as a population synthesizer for their own rule-based and activity-based travel demand model Albatross. They modified the IPF to fit the model input requirements, that is, known marginal distributions of individuals are converted to marginal distributions of households on relevant attributes and derived marginal household distributions are used as constraints of a multiway table of household counts.

Evers and Santapaola (2007) modified the conventional IPF algorithm for combining contingency tables with missing dimensions from a variety of data sources (e.g., traffic data, census data, etc.) with the example of traffic count data on German motorways.

Guo and Bhat (2007) discussed two issues associated with this conventional IPF approach: the zero-cell-value problem, and the inability to control for statistical distributions of both household and individual-level attributes. They presented a new population synthesis procedure that addresses the limitations of the conventional IPF approach developed by Beckman et al. in 1996 by controlling for statistical distributions defined by both household- and individual-level variables. The algorithm represented an extension of the

conventional IPF 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. Their validation results showed 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.

Wheaton et al. (2009) developed US synthesized human agent database providing a realistic agent population for use in agent-based models. That is, they implemented the conventional IPF algorithm to synthesize the household and individuals in the 50 states and the District of Columbia on a county-by-county basis and assigned them into each agent. Each household agent has been randomly located on a GIS map.

Srinivasan and Ma (2009) developed a heuristic data-fitting algorithm which is first described that can be used to synthesize populations by simultaneously controlling for household-level and person-level characteristics.

They applied the algorithm for both base-year and target-year population synthesis. State-of-the-practice methods for population synthesis fundamentally involve the development of a joint multi-way distribution using IPF. The development of this joint multi-way distribution requires that all controls are at the same "universe" (such as households). This condition is violated when both household- and person-level controls are present. Thus, a heuristic data-fitting algorithm was developed to systematically draw households from a "seed" dataset (such as the PUMS) such that several control tables (at household and person levels) are satisfied. According to Srinivasan and Ma (2009) in each iteration, a "fitness" value is calculated for each household in the seed dataset. This value is a measure of the extent to which the household contributes to satisfying the target values in all the different control tables simultaneously. The household with the highest fitness is drawn into the synthetic population of the census tract – thus a "greedy" heuristic is employed. When adding a household would violate several control tables, its fitness would have a negative value. The synthesis procedure stops when all households in the seed data have negative fitness and hence none can be selected into the synthesized population. Empirical testing indicates that, with this stopping criterion, the number of synthesized households is approximately equal to the actual number of households in the census tract.

For validation, their developed heuristic data-fitting methodology was applied synthesize both base-year and target-year populations for thirteen census tracts in Florida. They found that the greedy-heuristic procedure results in

synthetic populations that match rather closely with the true distributions. Further, the results also highlight the improvements that can be achieved by controlling for both household and person level attributes.

Ye et al. (2009) developed a heuristic approach, called the Iterative Proportional Updating (IPU) algorithm for generating a synthetic population while simultaneously matching both household-level and person-level joint distributions of control variables of interest. The next chapter describes their algorithm elaborately.

Mohammadian et al. (2010) reinforced the conventional IPF algorithm by accounting for multiple-levels of analysis units and control variables for both household-level and person-level, concluding that their methodology improved the fit to the person-controls at no cost to the fit against the household-level controls. Their methodology details how both household- and person-level characteristics can jointly be used as controls when synthesizing populations, as well as how other multiple level synthetic populations, such as firm/employee, household/vehicle, etc. can be estimated. The use of person-level, or any other sub-level, controls is implemented through a new technique involving the estimation of household selection probabilities based on the probability of observing each household given the required person-level characteristics in each analysis zone. They described their procedure to be a quick and efficient method for generating synthetic populations which can accurately replicate desired person-level characteristics. They also detailed the development of a new

methodology for using control variables at multiple analysis levels when synthesizing populations with an existing population synthesizer. The new procedure improves the fit of the synthesized person-level characteristics when compared to synthesis procedures that do not account for person-level controls. The introduction of a new household selection procedure has significantly increased the efficiency of the procedure while maintaining a good fit to the required person-level control variable joint-distribution without some of the run-time issues that are found in the constrained optimization type synthesizers. Their new methodology was an improvement on existing population synthesis techniques for controlling characteristics on multiple levels of analysis.

Another approach to population synthesis is to employ combinatorial optimization techniques, as shown by Voas and Williamson (2000). This approach is compared to synthetic reconstruction in Ryan et al. (2009) and Huang and Williamson (2001).

Abraham et al. (2012) described an approach using the combinatorial optimization algorithm; a versatile technique capable of simultaneously matching targets at multiple agent levels, such as properties of households as well as for individuals within the households. The software also supports simultaneous targets defined for multiple geographical levels. They demonstrated the use of the software in two applications; the synthesis of the 2000 population of California and the synthesis of the California 2008 employment in Oregon and surrounding areas. They found their algorithm acceptably fast and efficient in matching the

targets with a high degree of accuracy. The software they developed works to identify a list of units whose aggregate attribute values match a pre specified set of corresponding target values. This list forms a synthetic population of such units consistent with the target values. Each unit included in this list is drawn from a sample of such units, with the potential that any particular unit in the sample is included in the list 0, 1 or more times as appropriate. The software proceeds by iteratively considering one of three operations: adding a unit from the sample to the list, subtracting a unit from the list, or a 'swap' where a unit in the list is swapped out and a unit from the sample is swapped in. The match of the list to the target values is scored using a goodness-of-fit function. The population synthesis procedure based on combinatorial optimization has proven to be fast, flexible and practical for real-world use in very large model areas with unique challenges.

Another popular weighting procedure for expanding survey data to match marginal totals is entropy based weighting procedure. Many transportation researchers have investigated Entropy related models and, over the years, Entropy maximizing techniques have been used to develop models of trip distribution, mode split, and route choice. The Entropy maximization methodology proposed by Bar-Gera et al. (2009) presents a way to estimate weights that match the exogenously given distributions of the population including both household and person level marginal distributions. Entropy maximization principles trace their roots to statistical thermodynamics. The development and the application of Entropy maximization techniques have been conducted in the field of

transportation in numerous studies. One of the earliest efforts to use the principles of Entropy maximization in the field of transportation planning was carried out by Wilson (1969, 1970), in the estimation of origin-destination distributions by gravity models. Oppenheim (1995) presented a comprehensive discussion of the Entropy formulation and its equivalence to gravity trip distribution, logit mode choice and the logit stochastic traffic assignment in his 'Urban Travel Demand Modeling'.

Jornsten and Lundgren (1989) presented the similarity between the Entropy maximization methodology and the traditional logit-type framework to model mode splits. Further they presented that the logit model can be obtained as a special case of the general Entropy model. Fang and Tsao (1995) considered the linearly constrained Entropy maximization problem with quadratic cost and present a globally convergent algorithm which was both robust and efficient (Bar-Gera et al. 2009). The usual path flow based Entropy function was decomposed into a link flow based function and the likeness between the decomposed form and the LOGIT assignment were presented using Markov properties that form the basis of Dial's algorithm (Dial 1971). Rossi et al. (1989) proposed Entropy maximization as a condition for the most likely route flow solution among all user-equilibrium solutions. A time dependent combined model for trip distribution and traffic assignment was proposed by Li et al. (2002). The origin-destination matrix was estimated using the observed Entropy value and minimizing the total system travel time. Bar-Gera et al. (2009) proposed an algorithm for estimating

survey weights with multiple constraints using Entropy optimization approach. In their study, they found that the strict formulation can be used to estimate the weights when constraints imposed by distributions of population characteristics are feasible and relaxed formulation can be used to estimate weights when the constraints are infeasible such that distributions of the population characteristics are satisfied to within reasonable limits. This entropy maximization procedure is also one of the main methodologies of generating synthetic population using PopGen.

### 2.2 Summary of Literatures

As microsimulations become more and more used, the development of synthetic population generation methods become a growing field of interest as it is an important step of these models. Several algorithms are available in the literature mentioned above, and the choice of one of them depends on the final application, the available data and the size of the population to synthetize. However, no literature was found which focused on the sensitivity of these synthetic population generation procedures. This is why this study was motivated to test the sensitivity of these synthetic population generation procedures.

## 3. POPGEN

This chapter provides an overview of the software PopGen: a population synthesizing software developed by ASU transportation systems group. The chapter also contains the methodology and algorithm used in PopGen and the process how it works.

The synthetic population generators described in the literatures typically use census-based marginal distributions on household attributes to generate joint distributions on variables of interest using standard iterative proportional fitting (IPF) procedures. Households are then randomly drawn from an available sample in accordance with the joint distribution such that household-level attributes are matched perfectly. However, these traditional procedures do not control for person-level attributes and joint distributions of personal characteristics. The team for developing PopGen adopted a heuristic approach, called the Iterative Proportional Updating (IPU) algorithm to generate synthetic populations whereby both household-level and person-level characteristics of interest can be matched in a computationally efficient manner. The algorithm involves iteratively adjusting and reallocating weights among households of a certain type (cell in the joint distribution) until both household and person-level attributes are matched.

### 3.1 Methodology for PopGen

This section presents the methodology and algorithm implemented in PopGen. First, the Iterative Proportional Fitting (IPF) procedure is described with an

example and then the Iterative Proportional Updating (IPU) procedure, which is the basic idea behind PopGen is illustrated in a step-by-step procedure followed by an example.

### 3.1.1 Illustration of Iterative Proportional Fitting (IPF) Procedure

The IPF procedure which lies at the heart of most synthetic population generators, involves the estimation of household and person level joint distributions that match the given household and person level marginal frequency distributions. The following example describes how the IPF procedure expand the seed matrix to match the given marginal control totals while maintaining the joint distribution implied by the seed matrix.

Figure 1 shows a simple example of sample seed data and the summary of marginal distributions of total 100 households. There are two household variables, one is in the left most column, the household size and second one is in the top left, the household income. The household size marginal distributions are 30, 40 and 30 for household size categories 1, 2 and 3 or more respectively. The household size marginal distributions for low income and high income categories are 60 and 40 respectively. The seed matrix is shown in the figure which we need to expand my balancing row and column respectively. Figure 2 shows the column adjustments for income variable. The column factor is calculated by dividing the column marginal by the sum of that column for the seed matrix. Like here the total number of households for low income marginal is 60. Sum of low income

households for the seed matrix is 7. So the factor will be 60 divided by 7 which equals to 8.57. Now, all the cells for the low income column of the seed matrix will be multiplied by this value. The column factor for high income column will be calculated in the same way. Here the total number of households for high income marginal is 40. Sum of high income households for the seed matrix is 6. So the factor will be 40 divided by 6 which is equal to 6.67. So, all the cells for the high income column of the seed matrix will be multiplied by this value.

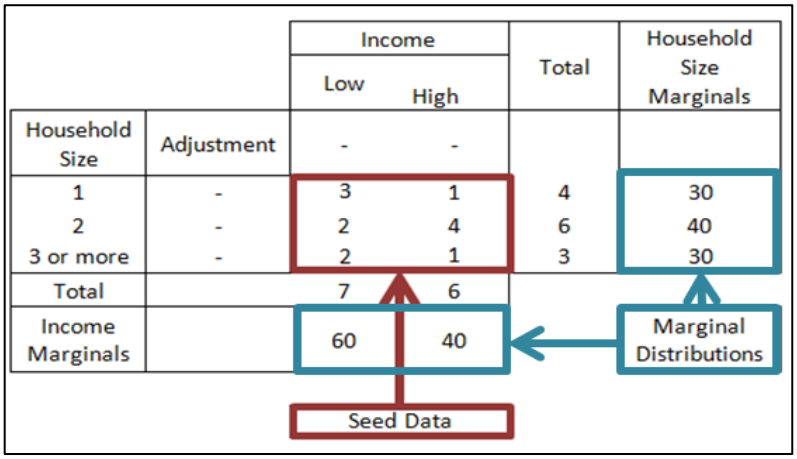


Figure 1: Sample Seed Data and Summary Marginal Distributions

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | 60/7 = 8.57 | 6.67 | | |
| 1 | - | 3 X 8.57 = 25.7 | 6.7 | 32.4 | 30 |
| 2 | - | 17.1 | 26.7 | 43.8 | 40 |
| 3 or more | - | 17.1 | 6.7 | 23.8 | 30 |
| Total | | 60 | 40 | | |
| Income Marginals | | 60 | 40 | | |

Figure 2: Adjustment for Income

Figure 3 shows the row adjustment for household size. The procedure for row adjustment is similar to the column adjustments. The row factor is calculated by dividing the row marginal by the sum of that row for the seed matrix. Like here the total number of households for household size 1 marginal is 30. Sum of household size 1for the seed matrix after the column adjustment is 32.4. So the factor will be 30 divided by 32.4 which equals to 0.93. Now, all the cells for the household size 1 row of the seed matrix will be multiplied by this value. The row factor for household size 2 and 3 or more will be calculated in the same way. Here the total numbers of households for household size 2 and 3 or more are 40 and 30 respectively. Sum of household sizes 2 and 3 or more for the seed matrix are 43.8 and 23.8 respectively. So the factors will be 40 divided by 43.8 which is equal to 0.91 and 30 divided by 23.8 which is equal to 1.26. So, the cells for the seed matrix (after column adjustment) for household size 2 will be multiplied by 0.91 while those for household size three or more will be multiplied by 1.26.

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | - | - | | |
| 1 | 30/32.4 = 0.93 | 23.82 | 6.18 | 30 | 30 |
| 2 | 0.91 | 15.65 | 24.35 | 40 | 40 |
| 3 or more | 1.26 | 21.60 | 8.40 | 30 | 30 |
| Total | | 61.08 | 38.92 | | |
| Income Marginals | | 60 | 40 | | |

Figure 3: Adjustment for Household Size

| | | Income | | Total | Household Size Marginals |
|---|---|---|---|---|---|
| | | Low | High | | |
| Household Size | Adjustment | - | - | | |
| 1 | 1.00 | 23.60 | 6.40 | 30 | 30 |
| 2 | 1.00 | 15.20 | 24.80 | 40 | 40 |
| 3 or more | 1.00 | 21.30 | 8.70 | 30 | 30 |
| Total | | 60.00 | 40.00 | | |
| Income Marginals | | 60 | 40 | | |

Figure 4: Multiway Frequency Table Matching Known Marginal Distributions

Figure 4 shows the final multiway frequency table matching known marginal distributions for the households. The convergence is achieved after three iterations. This way, the household and the person level joint distributions that match the given household and person level marginal frequency distributions can be obtained. However the IPF procedure will naturally result in two different sets of weights, one set for matching household distributions and one set for matching person-level distributions.

### 3.1.2 Illustration of Iterative Proportional Updating (IPU) procedure

The household weights computed in IPF procedure will never match the person weights computed by the same procedure. As a result, a synthetic population that is generated based on the application of household weights will yield joint distributions of person attributes that do not match the given person-level marginal distributions. This is because the traditional procedure involves simply

selecting all persons in the chosen households according to the household weights. In other words, the person weights are forced to be equal to the corresponding household weights, when in fact they are different. The desire to generate a synthetic population whereby both household and person-level attribute distributions are matched against known marginal distributions is one of the primary motivating factors for the creation of IPU algorithm. According to Ye et al. (2009) the general formulation of the algorithm are as follows:

"The steps for IPU procedure:

1. Generate a frequency matrix $D$ showing the household type and the frequency of different person types within each household for the sample. The dimension of the matrix generated will be $N \times m$, where $N$ is the number of households in the sample and $m$ is the number of population characteristic (household type and person type) constraints. An element in the matrix $d_{i,j}$ represents the contribution of household $i$ to the frequency of population characteristic (household type/person type) $j$.
2. Obtain joint distributions of household type and person type constraints using the standard IPF procedure and store the resulting estimates into a column vector $C$ where $c_j$ represents the value of the population characteristic $j$ and $j = 1, 2, \ldots, m$.
3. Initialize the weights vector represented by the column vector, $W$, such that $w_i = 1$ where $i = 1, 2, \ldots, N$.

Also, initialize a scalar, $\delta = \frac{\sum_j \left[ \left| \left( \sum_i d_{i,j} w_i - c_j \right) \right| / c_j \right]}{m}$ and set the value of the scalar, $\delta_{min} = \delta$.

4. Initialize a scalar, $r = 1$, representing the iteration number.
5. For each column $j$ ($j = 1, 2, \ldots, m$), record the indices (i.e., the row number or, in the context of the simple example, the household ID) into a column vector $S_j$, including only those that actually belong to household or person type $j$. Let an entry in such a column vector be denoted by $s_{qj}$ where $q$ is an index corresponding to non-zero elements in the $j_{th}$ column. For instance, in the simple example considered in Figure 5, $S_1$ would include elements (households) 1, 2, and 3; $S_2$ would include elements 4, 5, 6, 7, and 8; and so on.
6. Initialize a scalar $k = 1$ to serve as a constraint counter.
7. Retrieve the indices $s_{qk}$ of all the non-zero elements in the $k_{th}$ column stored in $S_k$ of Step 5 where $q$ is the index corresponding to non-zero elements in the $k_{th}$ column.
8. Calculate the adjustment $\rho$ for the $k_{th}$ constraint, $\rho = \frac{c_k}{\sum_q d_{s_{qk},k} \times w_{s_{qk}}}$
9. Update the weights with respect to the $k_{th}$ constraint as $w_{sqk} = \rho\ w_{sqk}$ . Recall that all initial weight values are set to one.
10. Update $k = k + 1$.

11. If $k \leq m$, i.e., the weight have not been adjusted with respect to all population characteristic constraints, then go to Step 7; otherwise, proceed to Step 12.
12. Set the value of a scalar, $\delta_{prev} = \delta$.
13. Calculate the new value of δ corresponding to the current iteration,

$$\delta = \frac{\sum_j \left[ \left| \left( \sum_i d_{i,j} w_i - c_j \right) / c_j \right| \right]}{m}.$$

14. Calculate the improvement in goodness-of-fit, $\Delta = | \delta - \delta_{prev} |$.
15. If $\delta < \delta_{min}$, update $\delta_{min} = \delta$, and store the corresponding weights in a column vector *SW* with elements $sw_i = w_i$ for $i = 1, 2, \ldots, N$. Otherwise, proceed to Step 16.
16. Update the iteration number, $r = r + 1$.
17. If $\Delta > \varepsilon$ (a small positive number, e.g., $1 \times 10^{-4}$), go back to step 6. Otherwise, convergence has been achieved and a solution is obtained. The selected weights are stored in the column vector *SW* corresponding to the smallest absolute relative difference $\delta_{min}$.

The updated household weights are recorded in the column vector *SW*. It should be noted that Step 15 in the algorithm is critical because the $\delta$ value is not always strictly decreasing. As a result, it is necessary to ensure that weights corresponding to the minimum value of δ are retained at each iteration of the process. At the conclusion of the process outlined above, a perfect solution is obtained if it falls within the feasible range. If, however, the solution does not fall

within a feasible range, then additional steps may be warranted to choose the appropriate corner solution. Given the emphasis on matching household-level constraints in current practice, the additional steps in the procedure proceed to the corner solution to ensure that household constraints are met perfectly. The steps are:

18. Initialize a scalar $h = 1$, where $h = 1, 2, \ldots m_h$, where $m_h$ is the number of household constraints that need to be satisfied.
19. Retrieve the indices $s_{qh}$ of all the non-zero elements in the $h_{th}$ column stored in column vector $S_h$ of Step 5.
20. Calculate the adjustment $\rho$ for the $h_{th}$ constraint, $\rho = \frac{c_h}{\sum_q d_{s_{qh},h} \times w_{s_{qh}}}$
21. Update the weights with respect to the $h_{th}$ constraint as $sw_{sqh} = \rho\ sw_{sqh}$
22. Update $h = h + 1$.
23. If $h \leq m_h$, go back to Step 18; otherwise, a corner solution has been reached and the algorithm is terminated."

The IPU procedure considers reducing the inconsistency in person level distributions by adjusting the household level weights based on the person weights obtained from the IPF procedure. The process by which this can be accomplished is best illustrated with the help of a small numerical example. Figure 5 shows a frequency matrix where a row in the matrix corresponds to a single household record and provides data describing the composition of the household. For example, the first household is of type 1 and has one individual

each of person types 1, 2, and 3. There are two household types and three person types considered in this example. In this example, there are eight households with 23 individuals. All initial household weights are set to unity as shown in the Figure 5.

| Household ID | Weights | Household Type 1 | Household Type 2 | Person Type 1 | Person Type 2 | Person Type 3 | Weights 1 | Weights 2 | Weights 3 | Weights 4 | Weights 5 | Final weights |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 | 1 | 11.67 | 11.67 | 9.51 | 8.05 | 12.37 | 1.36 |
| 2 | 1 | 1 | 0 | 1 | 0 | 1 | 11.67 | 11.67 | 9.51 | 9.51 | 14.61 | 25.66 |
| 3 | 1 | 1 | 0 | 2 | 1 | 0 | 11.67 | 11.67 | 9.51 | 8.05 | 8.05 | 7.98 |
| 4 | 1 | 0 | 1 | 1 | 0 | 2 | 1.00 | 13.00 | 10.59 | 10.59 | 16.28 | 27.79 |
| 5 | 1 | 0 | 1 | 0 | 2 | 1 | 1.00 | 13.00 | 13.00 | 11.00 | 16.91 | 18.45 |
| 6 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 10.59 | 8.97 | 8.97 | 8.64 |
| 7 | 1 | 0 | 1 | 2 | 1 | 2 | 1.00 | 13.00 | 10.59 | 8.97 | 13.78 | 1.47 |
| 8 | 1 | 0 | 1 | 1 | 1 | 0 | 1.00 | 13.00 | 10.59 | 8.97 | 8.97 | 8.64 |
| Weighted Sum | | 3.00 | 5.00 | 9.00 | 7.00 | 7.00 | | | | | | |
| Constraints | | 35.00 | 65.00 | 91.00 | 65.00 | 104.00 | | | | | | |
| $\delta_0$ | | 0.9143 | 0.9231 | 0.9011 | 0.8923 | 0.9327 | | | | | | |
| Weighted Sum 1 | | 35.00 | 5.00 | 51.67 | 28.33 | 28.33 | | | | | | |
| Weighted Sum 2 | | 35.00 | 65.00 | 111.67 | 88.33 | 88.33 | | | | | | |
| Weighted Sum 3 | | 28.52 | 55.38 | 91.00 | 76.80 | 74.39 | | | | | | |
| Weighted Sum 4 | | 25.60 | 48.50 | 80.11 | 65.00 | 67.68 | | | | | | |
| Weighted Sum 5 | | 35.02 | 64.90 | 104.84 | 85.94 | 104.00 | | | | | | |
| $\delta_a$ | | 0.0006 | 0.0015 | 0.1521 | 0.3222 | 0.0000 | | | | | | |
| Final Weighted Sum | | 35.00 | 65.00 | 91.00 | 65.00 | 104.00 | | | | | | |

Figure 5: Example of the Iterative Proportional Updating (IPU) Algorithm

The row titled "weighted sum" represents the sum of each column weighted by the "weights" column. The "constraints" row provides the frequency distribution of the household and person types that must be matched. The rows titled $\delta_a$ and $\delta_0$ provide the absolute value of the relative difference between the weighted sum and the given constraints so that the "goodness of- fit" of the algorithm can be assessed at each stage of the algorithm and convergence criteria can be set. The data structure shown in the table can be used to formulate a mathematical

optimization problem in which one desires to calibrate weights such that the weighted sum equals or nearly equals the given frequency distribution.

### 3.2 Rounding Methods Used in PopGen

Since the IPF procedure can result in many cells with decimal values, the household level joint distributions obtained from the IPF procedure are rounded off to the nearest integer. Then, for each household type, households are drawn randomly from the set of PUMS households that belong to that particular category. The number of households randomly drawn is equal to the frequency of that household type in the rounded joint distribution table. The rounding procedures implemented in PopGen are arithmetic rounding, bucket rounding and stochastic rounding.

#### 3.2.1 Arithmetic Rounding

The arithmetic rounding procedure in PopGen accounts for the difference between the rounded frequency sum and the actual frequency sum. At first, the household type frequencies with decimal values greater than or equal to 0.50 are rounded up and decimal values less than 0.50 are rounded down. Next, frequencies with decimal values less than or equal to 0.50 are ranked in an order so that the closet value to 0.50 gets rank 1, the second closest is given rank 2 and the remaining values are ranked similarly; this provides a ranking for all of the values. Then, the difference between the actual frequency summation and the rounding frequency

summation is calculated. Finally, household frequencies are rounded according to the rank only as much as adjustments are required. Figure 6 shows an example of arithmetic rounding.

| Household Type | Frequency | Rounded Frequency | Difference | Ranking to Receive a Household | Adjustment | Adjusted Frequency |
|---|---|---|---|---|---|---|
| 1 | 64.85 | 65 | 0.15 | 16 | | 65 |
| 2 | 12.34 | 12 | -0.34 | 10 | | 12 |
| 3 | 10.36 | 10 | -0.36 | 9 | | 10 |
| 4 | 0.43 | 0 | -0.43 | 5 | 1 | 1 |
| 5 | 0.49 | 0 | -0.49 | 1 | 1 | 1 |
| 6 | 0.47 | 0 | -0.47 | 3 | 1 | 1 |
| 7 | 0.44 | 0 | -0.44 | 4 | 1 | 1 |
| 8 | 0.39 | 0 | -0.39 | 6 | | 0 |
| 9 | 0.48 | 0 | -0.48 | 2 | 1 | 1 |
| 10 | 0.10 | 0 | -0.10 | 15 | | 0 |
| 11 | 0.12 | 0 | -0.12 | 14 | | 0 |
| 12 | 0.20 | 0 | -0.20 | 13 | | 0 |
| 13 | 0.27 | 0 | -0.27 | 12 | | 0 |
| 14 | 0.28 | 0 | -0.28 | 11 | | 0 |
| 15 | 0.38 | 0 | -0.38 | 7 | | 0 |
| 16 | 0.37 | 0 | -0.37 | 8 | | 0 |
| ***Total*** | ***91.97*** | ***87*** | ***-4.97*** | | ***5*** | ***92*** |

Figure 6: Illustration of Arithmetic Rounding Procedure

### 3.2.2 Bucket Rounding

The bucket rounding procedure also ensures that the rounded frequency sum and the actual frequency sum are the same. Bucket rounding calculates the accumulated rounding error and adjusts a value to 1 when the accumulated sum is greater than or equal to 0.50. When a decimal value greater than or equal to 0.50 is rounded up the over estimation is deducted from the next fraction. With this, the accumulated rounding error is used to bias the rounding of the next frequency value. Figure 7 shows an example of bucket rounding.

| Household Type | Frequency | Integer Part | Calculations | Accumulated Difference | Adjustment | Adjusted Frequency |
|---|---|---|---|---|---|---|
| 1 | 64.85 | 64 | | 0.85 | 1 | 65 |
| 2 | 12.34 | 12 | *-0.15* + 0.34 | 0.19 | | 12 |
| 3 | 10.36 | 10 | 0.34 + 0.36 | 0.55 | 1 | 11 |
| 4 | 0.43 | 0 | *-0.45* + 0.43 | -0.02 | | 0 |
| 5 | 0.49 | 0 | -0.02 + 0.49 | 0.47 | | 0 |
| 6 | 0.47 | 0 | 0.47 + 0.47 | 0.94 | 1 | 1 |
| 7 | 0.44 | 0 | *-0.06* + 0.44 | 0.38 | | 0 |
| 8 | 0.39 | 0 | 0.38 + 0.39 | 0.77 | 1 | 1 |
| 9 | 0.48 | 0 | *-0.23* + 0.48 | 0.25 | | 0 |
| 10 | 0.10 | 0 | 0.25 + 0.10 | 0.35 | | 0 |
| 11 | 0.12 | 0 | 0.35 + 0.12 | 0.47 | | 0 |
| 12 | 0.20 | 0 | 0.47 + 0.20 | 0.67 | 1 | 1 |
| 13 | 0.27 | 0 | *-0.33* + 0.27 | -0.06 | | 0 |
| 14 | 0.28 | 0 | -0.06 + 0.28 | 0.22 | | 0 |
| 15 | 0.38 | 0 | 0.22 + 0.38 | 0.60 | 1 | 1 |
| 16 | 0.37 | 0 | *-0.40* + 0.37 | -0.03 | | 0 |
| ***Total*** | ***91.97*** | ***92*** | | | | ***92*** |

Figure 7: Illustration of Bucket Rounding Procedure

### 3.2.3 Stochastic Rounding

In the stochastic rounding procedure, frequencies are randomly rounded up or rounded down. This rounding procedure accounts for the difference between the rounded frequency sum and the actual frequency sum. The stochastic rounding procedure can be illustrated by the following example by Konduri et al. (2010).

- **“** Consider a household type frequency of 22.41
- It can be rounded up with a probability of 0.41 and rounded down with a probability of 0.59
- We randomly draw a number between 0 and 1 to decide which way the frequency gets rounded

- Say if the random number was 0.20, then 0.00 ≤*0.20*≤0.41, so the frequency gets rounded up to 23.00
- Alternatively if the random number was 0.78, then 0.41 < *0.78*≤ 1.00, so the frequency gets rounded down to 22.00"

### 3.3 Sample Data

In this study, the main sources of data are Census Bureau 2000 and American Community Survey (ACS) 2005-2009. The software PopGen requires three types of input records: marginal totals, sample records and geo-correspondence file. The significant household level constraints in the input data are household size, household income, household type, family type, presence of children, householder age, and householder race. Significant person level constraints are age, gender, race, and employment status.

Marginal data are three types, household, person and group quarter. The household file contains number of total households in each category of household variables for each block group, the person file contains number of total persons in each category of person variables for each block group, and the group quarter file contains number of total institutional and non-institutional group quarters in each block group. All files are followed by compulsory fields: state, county, tract, and block group. Figure 3, 4 and 5 show examples of household, person and group quarter marginal records.

| state | county | tract | bg | hhldinc1 | hhldinc2 | hhldinc3 | hhldinc4 | hhldinc5 | hhldinc6 | hhldinc7 | hhldinc8 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| int | int | bigint | int | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 4 | 13 | 10100 | 1 | 46 | 64 | 54 | 32 | 60 | 118 | 115 | 282 |
| 4 | 13 | 10100 | 3 | 57 | 14 | 23 | 0 | 30 | 26 | 0 | 24 |
| 4 | 13 | 20202 | 1 | 560 | 264 | 264 | 172 | 218 | 155 | 28 | 0 |
| 4 | 13 | 30304 | 1 | 81 | 77 | 58 | 57 | 27 | 86 | 0 | 0 |
| 4 | 13 | 30312 | 4 | 19 | 36 | 73 | 92 | 35 | 27 | 0 | 0 |
| 4 | 13 | 30315 | 1 | 181 | 155 | 212 | 111 | 120 | 94 | 0 | 0 |
| 4 | 13 | 30319 | 2 | 35 | 42 | 54 | 113 | 83 | 155 | 21 | 6 |
| 4 | 13 | 30322 | 6 | 97 | 27 | 141 | 53 | 221 | 92 | 12 | 17 |
| 4 | 13 | 30323 | 1 | 41 | 16 | 53 | 50 | 60 | 21 | 8 | 0 |
| 4 | 13 | 30323 | 2 | 25 | 48 | 48 | 50 | 134 | 104 | 21 | 14 |
| 4 | 13 | 30323 | 4 | 37 | 36 | 42 | 27 | 57 | 88 | 0 | 7 |
| 4 | 13 | 30325 | 1 | 13 | 20 | 13 | 36 | 90 | 74 | 100 | 32 |
| 4 | 13 | 30325 | 2 | 121 | 123 | 147 | 131 | 164 | 274 | 19 | 0 |
| 4 | 13 | 30327 | 1 | 31 | 11 | 3 | 52 | 67 | 146 | 8 | 0 |
| 4 | 13 | 30329 | 3 | 62 | 49 | 62 | 61 | 56 | 34 | 31 | 0 |
| 4 | 13 | 30333 | 2 | 59 | 63 | 79 | 78 | 119 | 379 | 195 | 196 |
| 4 | 13 | 30335 | 2 | 0 | 85 | 129 | 93 | 91 | 156 | 18 | 28 |
| 4 | 13 | 30337 | 2 | 27 | 72 | 89 | 66 | 96 | 200 | 37 | 0 |
| 4 | 13 | 30340 | 3 | 9 | 11 | 49 | 66 | 85 | 186 | 36 | 22 |
| 4 | 13 | 30341 | 2 | 43 | 129 | 142 | 70 | 91 | 178 | 82 | 31 |
| 4 | 13 | 30343 | 1 | 66 | 128 | 167 | 136 | 190 | 699 | 445 | 475 |
| 4 | 13 | 30344 | 1 | 0 | 27 | 31 | 62 | 53 | 281 | 116 | 46 |
| 4 | 13 | 30344 | 2 | 27 | 0 | 16 | 8 | 38 | 181 | 113 | 17 |
| 4 | 13 | 30344 | 3 | 9 | 10 | 15 | 13 | 78 | 215 | 146 | 26 |

Figure 8: Example of Household Marginal File

| state | county | tract | bg | agep1 | agep2 | agep3 | agep4 | agep5 | agep6 | agep7 | agep8 | agep9 | agep10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| int | int | bigint | int | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 4 | 13 | 10100 | 1 | 20 | 182 | 90 | 66 | 206 | 418 | 457 | 306 | 68 | 4 |
| 4 | 13 | 10100 | 3 | 0 | 37 | 50 | 60 | 111 | 50 | 35 | 0 | 0 | 0 |
| 4 | 13 | 20202 | 1 | 426 | 1117 | 782 | 517 | 703 | 457 | 453 | 374 | 300 | 64 |
| 4 | 13 | 30304 | 1 | 44 | 25 | 65 | 63 | 85 | 96 | 81 | 180 | 119 | 0 |
| 4 | 13 | 30312 | 4 | 0 | 0 | 0 | 0 | 0 | 0 | 19 | 126 | 249 | 45 |
| 4 | 13 | 30315 | 1 | 164 | 315 | 306 | 458 | 257 | 229 | 147 | 146 | 126 | 37 |
| 4 | 13 | 30319 | 2 | 144 | 307 | 180 | 247 | 327 | 175 | 73 | 41 | 20 | 17 |
| 4 | 13 | 30322 | 6 | 146 | 226 | 251 | 330 | 317 | 170 | 122 | 51 | 59 | 0 |
| 4 | 13 | 30323 | 1 | 103 | 264 | 76 | 205 | 144 | 92 | 18 | 15 | 0 | 0 |
| 4 | 13 | 30323 | 2 | 103 | 136 | 156 | 212 | 121 | 101 | 160 | 68 | 31 | 0 |
| 4 | 13 | 30323 | 4 | 104 | 172 | 107 | 205 | 167 | 164 | 18 | 0 | 4 | 0 |
| 4 | 13 | 30325 | 1 | 89 | 171 | 105 | 223 | 233 | 185 | 43 | 40 | 0 | 0 |
| 4 | 13 | 30325 | 2 | 202 | 307 | 285 | 508 | 391 | 323 | 280 | 109 | 117 | 29 |
| 4 | 13 | 30327 | 1 | 56 | 172 | 224 | 102 | 182 | 200 | 55 | 19 | 9 | 0 |
| 4 | 13 | 30329 | 3 | 7 | 26 | 110 | 442 | 517 | 222 | 178 | 237 | 114 | 7 |
| 4 | 13 | 30333 | 2 | 156 | 398 | 272 | 611 | 643 | 362 | 238 | 77 | 25 | 0 |
| 4 | 13 | 30335 | 2 | 175 | 217 | 176 | 447 | 232 | 197 | 37 | 91 | 0 | 0 |
| 4 | 13 | 30337 | 2 | 124 | 160 | 176 | 327 | 299 | 195 | 73 | 19 | 20 | 10 |
| 4 | 13 | 30340 | 3 | 162 | 309 | 235 | 282 | 270 | 202 | 63 | 49 | 25 | 0 |
| 4 | 13 | 30341 | 2 | 0 | 0 | 8 | 18 | 7 | 137 | 297 | 457 | 354 | 74 |
| 4 | 13 | 30343 | 1 | 200 | 621 | 317 | 382 | 906 | 1165 | 1093 | 627 | 237 | 28 |

Figure 9: Example of Person Marginal File

| state | county | tract | bg | groupquarter1 | groupquarter2 |
|---|---|---|---|---|---|
| int | int | bigint | int | bigint | bigint |
| 4 | 13 | 10100 | 1 | 0 | 0 |
| 4 | 13 | 10100 | 3 | 0 | 0 |
| 4 | 13 | 20202 | 1 | 39 | 11 |
| 4 | 13 | 30304 | 1 | 0 | 0 |
| 4 | 13 | 30312 | 4 | 0 | 0 |
| 4 | 13 | 30315 | 1 | 0 | 0 |
| 4 | 13 | 30319 | 2 | 0 | 17 |
| 4 | 13 | 30322 | 6 | 0 | 0 |
| 4 | 13 | 30323 | 1 | 0 | 0 |
| 4 | 13 | 30323 | 2 | 0 | 0 |
| 4 | 13 | 30323 | 4 | 0 | 0 |
| 4 | 13 | 30325 | 1 | 0 | 0 |
| 4 | 13 | 30325 | 2 | 0 | 0 |
| 4 | 13 | 30327 | 1 | 0 | 0 |
| 4 | 13 | 30329 | 3 | 1220 | 0 |

Figure 10: Example of Group Quarter Marginal File

The three types of sample data that are required for matching each type of marginal totals are household, person and group quarter records. The household sample file contains sample records of household in household level, the person sample file contains sample records of person in person level and the group quarter sample file contains sample records of group quarter in household level. All files are followed by compulsory fields state, pumano, hhid, and serial number. Figure 4, 5 and 6 show examples of household, person and group quarter sample records respectively.

| state | pumano | hhid | serialno | hhtype | hhldtype | hhldinc | hhldsize | hhchildpresence |
|---|---|---|---|---|---|---|---|---|
| int | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 4 | 101 | 35461 | 1419 | 1 | 5 | 3 | 1 | 2 |
| 4 | 101 | 35463 | 3229 | 1 | 5 | 1 | 1 | 2 |
| 4 | 101 | 35465 | 4659 | 1 | 1 | 3 | 2 | 2 |
| 4 | 101 | 35466 | 4837 | 1 | 1 | 7 | 2 | 2 |
| 4 | 101 | 35468 | 12815 | 1 | 1 | 6 | 2 | 2 |
| 4 | 101 | 35469 | 18892 | 1 | 1 | 5 | 2 | 2 |
| 4 | 101 | 35470 | 19063 | 1 | 5 | 1 | 1 | 2 |
| 4 | 101 | 35471 | 20690 | 1 | 1 | 3 | 2 | 2 |
| 4 | 101 | 35473 | 26233 | 1 | 5 | 3 | 1 | 2 |
| 4 | 101 | 35475 | 26760 | 1 | 1 | 8 | 2 | 2 |
| 4 | 101 | 35476 | 28945 | 1 | 5 | 3 | 1 | 2 |
| 4 | 101 | 35478 | 33088 | 1 | 1 | 5 | 3 | 2 |
| 4 | 101 | 35480 | 34671 | 1 | 5 | 4 | 1 | 2 |
| 4 | 101 | 35481 | 35136 | 1 | 1 | 4 | 2 | 2 |

Figure 11: Example of Household Sample File

| state | pumano | hhid | serialno | pnum | page | pgender | prace | employment |
|---|---|---|---|---|---|---|---|---|
| int | bigint | bigint | bigint | bigint | bigint | bigint | bigint | bigint |
| 4 | 101 | 35461 | 1419 | 1 | 8 | 2 | 1 | 4 |
| 4 | 101 | 35462 | 1957 | 1 | 7 | 2 | 1 | 3 |
| 4 | 101 | 35463 | 3229 | 1 | 8 | 2 | 1 | 4 |
| 4 | 101 | 35465 | 4659 | 1 | 7 | 1 | 1 | 2 |
| 4 | 101 | 35465 | 4659 | 2 | 7 | 2 | 1 | 2 |
| 4 | 101 | 35466 | 4837 | 1 | 8 | 2 | 1 | 4 |
| 4 | 101 | 35466 | 4837 | 2 | 8 | 1 | 1 | 4 |
| 4 | 101 | 35468 | 12815 | 1 | 4 | 1 | 1 | 2 |
| 4 | 101 | 35468 | 12815 | 2 | 4 | 2 | 1 | 2 |
| 4 | 101 | 35469 | 18892 | 1 | 8 | 1 | 1 | 4 |
| 4 | 101 | 35469 | 18892 | 2 | 8 | 2 | 1 | 4 |

Figure 12 Example of Person Sample File

| state | pumano | hhid | serialno | hhtype | groupquarter |
|---|---|---|---|---|---|
| int | bigint | bigint | bigint | bigint | bigint |
| 4 | 101 | 35462 | 1957 | 2 | 2 |
| 4 | 101 | 35484 | 39293 | 2 | 1 |
| 4 | 101 | 35489 | 62731 | 2 | 1 |
| 4 | 101 | 35546 | 229554 | 2 | 1 |
| 4 | 101 | 35586 | 353628 | 2 | 1 |
| 4 | 101 | 35592 | 375865 | 2 | 1 |
| 4 | 101 | 35614 | 430020 | 2 | 1 |
| 4 | 101 | 35694 | 657753 | 2 | 1 |
| 4 | 101 | 35698 | 665838 | 2 | 2 |

Figure 13: Example of Group Quarter Sample File

The geographic correspondence file provides the correspondence between the geography and the PUMA to which the geography belongs. All fields are compulsory here which contains information of county, tract, block group, state, puma number, state abbreviation, and county name.

| county | tract | bg | state | pumano | stateabb | countyname |
|---|---|---|---|---|---|---|
| int | bigint | int | int | bigint | text | text |
| 13 | 10100 | 3 | 4 | 106 | AZ | Maricopa |
| 13 | 10100 | 1 | 4 | 106 | AZ | Maricopa |
| 13 | 20202 | 1 | 4 | 106 | AZ | Maricopa |
| 13 | 30304 | 1 | 4 | 104 | AZ | Maricopa |
| 13 | 30312 | 4 | 4 | 101 | AZ | Maricopa |
| 13 | 30315 | 1 | 4 | 104 | AZ | Maricopa |
| 13 | 30319 | 2 | 4 | 104 | AZ | Maricopa |
| 13 | 30322 | 6 | 4 | 103 | AZ | Maricopa |
| 13 | 30323 | 2 | 4 | 103 | AZ | Maricopa |
| 13 | 30323 | 4 | 4 | 103 | AZ | Maricopa |
| 13 | 30323 | 1 | 4 | 103 | AZ | Maricopa |
| 13 | 30325 | 2 | 4 | 103 | AZ | Maricopa |

Figure 14: Example of Geographic Correspondence File

# 4. EXPERIMENT

This chapter contains three sections. First section represents the context of geographic locations, the second section contains the description of experimental design, and the third section details the goodness of fit measures used for the comparisons of several outputs in this thesis.

## 4.1 Context of Geographic Location

Maricopa County is the largest county in the state of Arizona which contains 57.63% of total households and 59.72% of the total population of Arizona (US Census Bureau 2010). Maricopa County consists of 663 tracts, which contain 2109 numbers of block groups.

Among the 2109 block groups of Maricopa County, 300 block groups were selected randomly for the analysis. This 300 block groups comprises 151,675 numbers of households and 411,414 numbers of persons which are 13.38% of the total households and 13.39% of total persons of Maricopa County.

## 4.2 Description of Experiments

### 4.2.1 Combination of Constraints

In this study, an approach has been taken to analyze the sensitivity of the synthetic population generation procedures by changing and varying the number of alternative constraints. Figure 15 and Figure 16 respectively show the household level and person level variables with their categories available in

Census 2000 dataset. A list was started taking four attributes from household level: household size, household income, household type and household children presence and four attributes from person level: person's age, person's gender, person's race and person's employment status. These attributes were selected for controlling while generating the synthetic population using PopGen. Household level variables, householder race and householder age was excluded from the list because these two variables are highly correlated with the person variables person's race and person's age. From the selected four household type and four person type attributes, a list of eleven potential combinations was developed by changing these numbers of controls. Figure 17 shows the List of Selected Combinations with Alternative Constraints in Household and Person Levels.

| Household Variables | Categories |
|---|---|
| Household Size | 1 - Households with 1 person |
| | 2 - Households with 2 persons |
| | 3 - Households with 3 persons |
| | 4 - Households with 4 persons |
| | 5 - Households with 5 persons |
| | 6 - Households with 6 persons |
| | 7 - Households with 7 or more persons |
| Household Income | 1- $0 - $24,999 |
| | 2- $25,000 - $34,999 |
| | 3- $35,000 - $44,999 |
| | 4- $45,000 - $59,999 |
| | 5- $60,000 - $99,999 |
| | 6- $100,000 - $149,999 |
| | 7- $100,000 - $149,999 |
| | 8- Over $ 150,000 |
| Household Type | 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 |
| Presence of Children | 1 - children present |
| | 2 - children not present |
| Householder Age | 1- Householder age < 65 years |
| | 2- Householder age is >= 65 years |
| Householder Race | 1-White |
| | 2-Black or African American |
| | 3-American Indian and Alaska Native |
| | 4-Asian |
| | 5-Native Hawaiian and Other Pacific Islander |
| | 6-Some other race |
| | 7-Two or more races |

Figure 15: Available Household Level Variables in Census 2000 Dataset

| Person Variables | Categories |
|---|---|
| Age | 1 - Age between 0 and 4 |
| | 2 - Age between 5 and 14 |
| | 3 - Age between 15 and 24 |
| | 4 - Age between 25 and 34 |
| | 5 - Age between 35 and 44 |
| | 6 - Age between 45 and 54 |
| | 7 - Age between 55 and 64 |
| | 8 - Age between 65 and 74 |
| | 9 - Age between 75 and 84 |
| | 10 - Age greater than equals 85 |
| Gender | 1 - Male |
| | 2 - Female |
| Race | 1-White |
| | 2-Black or African American |
| | 3-American Indian and Alaska Native |
| | 4-Asian |
| | 5-Native Hawaiian and Other Pacific Islander |
| | 6-Some other race |
| | 7-Two or more races |
| Employment Status | 1-Not Eligible (Under 16 years) |
| | 2-Employed |
| | 3-Unemployed |
| | 4-Not In Labor Force |

Figure 16: Available Person Level Variables in Census 2000 Dataset

| Combinations | No of Controls | | Household Constraints | | | | Person Constraints | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Household Level | Person Level | | | | | | | | |
| **1** | 4 | 4 | Size | Income | Type | Child Presence | Age | Gender | Race | Employment |
| **2** | 4 | 3 | Size | Income | Type | Child Presence | Age | Gender | Race | |
| **3** | 4 | 2 | Size | Income | Type | Child Presence | Age | Gender | | |
| **4** | 4 | 1 | Size | Income | Type | Child Presence | Age | | | |
| **5** | 4 | 0 | Size | Income | Type | Child Presence | - | | | |
| **6** | 3 | 0 | Size | Income | Type | | - | | | |
| **7** | 2 | 0 | Size | Income | | | - | | | |
| **8** | 1 | 0 | Size | | | | - | | | |
| **9** | 1 | 1 | Size | | | | Age | | | |
| **10** | 2 | 1 | Size | Income | | | Age | | | |
| **11** | 2 | 2 | Size | Income | | | Age | Gender | | |

Figure 17: List of Selected Combinations with Alternative Constraints in Household and Person Levels

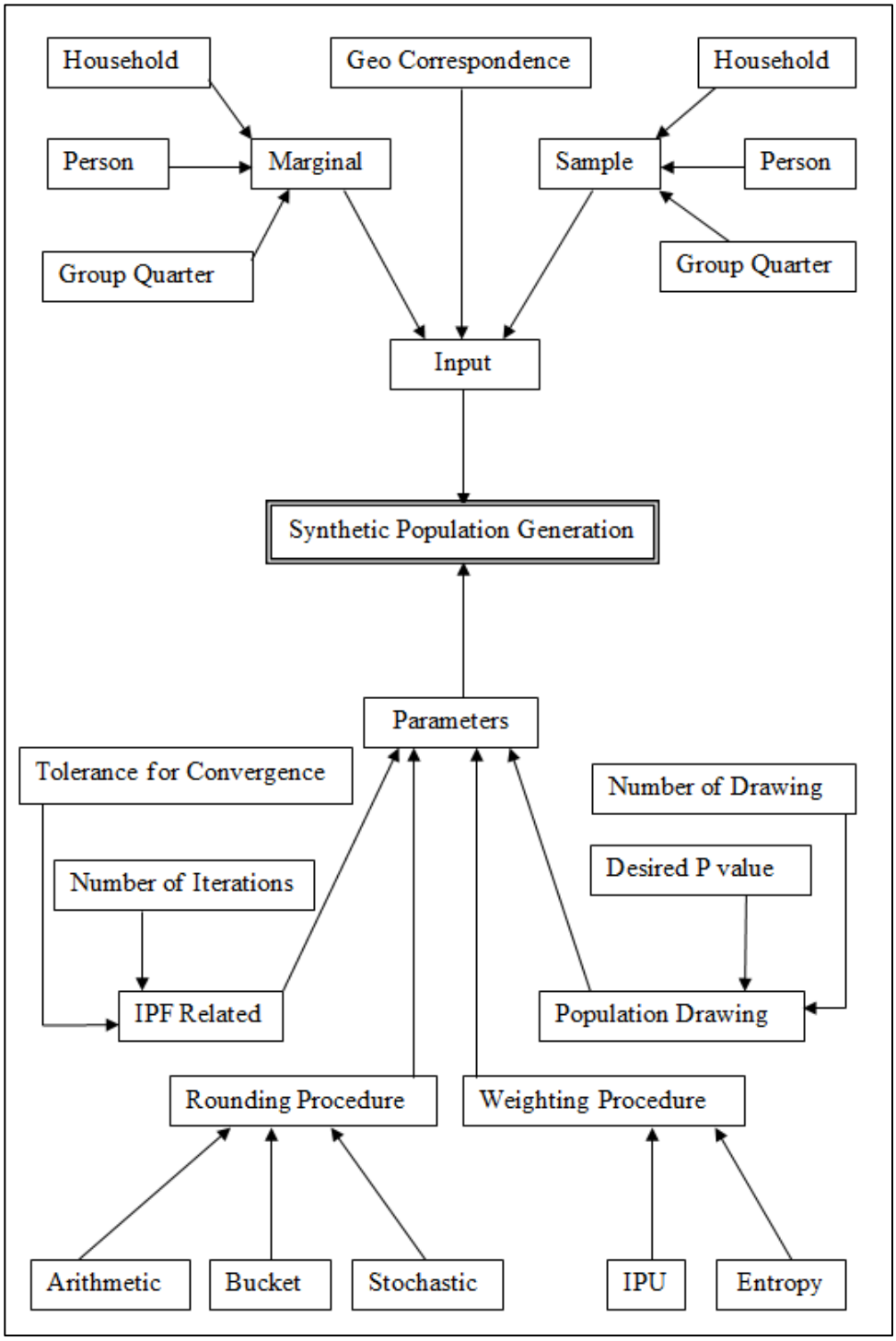


Figure 18: Framework for Analyzing Sensitivity of Population Generation

### 4.2.2 Alternative Rounding Methods

The second approach was to find out the combinations where the aggregate and disaggregate level synthetic population is less accurate in comparison to other combinations. Figure 18 shows parameters of synthetic population generations whose change affects the results. The motivation was to find out those parameters whose change affects the synthetic result significantly. One of these parameters is the rounding procedure (described in chapter 3). So the simulations were run for all the eleven combinations with arithmetic rounding and bucket rounding.

### 4.2.3 Alternative Sample Data

The final experiment was to investigate whether any change in accuracy is possible with a richer sample data where the software will have more flexibility in drawing households. To find this, two types of samples are used, the Census 2000 sample and a combined sample prepared from Census 2000 and ACS 2005-2009 sample datasets. Table 1 shows the total number of household, group quarter and person records available in Census 2000, ACS 2005-2009 and combined sample.

Table 1: Number of Household, Person and Group Quarter Records in Different Samples.

| Sample | Census | ACS | Combined |
|---|---|---|---|
| Household | 95,066 | 119278 | 214,344 |
| Person | 259,694 | 303402 | 563,096 |
| Group Quarter | 5,489 | 4573 | 10,062 |

### 4.3 Goodness of Fit Measurements

To measure the goodness of fit, three procedures were applied: chi-square value, normalized percent difference and the R-square value. The formulas for the above mentioned measures are as follows:

Chi-square Value:

$$X^2 = \Sigma \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

Normalized Percent Difference:

$$\text{Percent Difference} = \frac{|\text{Observed frequency} - \text{Expected frequency}|}{\text{Expected frequency}} \text{ X } 100\%$$

R-square value was computed by ordinary least square regression by taking marginal totals as independent variable and synthetic totals as dependent variable.

R-Square Value:

$$R^2 = 1 - \frac{SSE}{SST}$$

Where SSE = sum of squared errors and SST = sum of squared totals

The best combinations were selected depending on the lower value of chi-square and normalized percent difference and closest value of R-square to 1.

In summary, the methodology used for this study can be presented as follows:

- Investigate available meaningful household and person attributes
- Develop list of alternative constraints for Census 2000 dataset
- Synthesize population for the built alternatives using default arithmetic rounding
- Analyze the results in both aggregate and disaggregate levels
- Run analysis using bucket rounding for the critical combinations found in arithmetic rounding method
- Compare the difference in results for arithmetic rounding and bucket rounding
- Run analysis using combined sample as input for the critical combinations found in arithmetic rounding method
- Compare the difference in results for census sample and combined sample

## 5. ANALYSIS AND RESULTS

This chapter contains three major parts of the analysis and findings from different experimental combinations. First section analyzes the results for alternative constraints. Second section analyzes all combinations with two different rounding methods. Finally the third section analyses the variation in synthetic results for two different sample inputs: the Census 2000 sample and the combined sample prepared from the Census 2000 and the American Community Survey (ACS) 2005-2009 dataset.

### 5.1 Analysis with Alternative Constraints

In this section the potential set of eleven combinations of household and person attributes were compared. The Census 2000 sample and marginal records are used for synthesizing. Eleven numbers of simulations were run for the selected 300 block groups in Maricopa County.

Table 1 shows the comparison of the marginal and the synthetic households for eleven combinations of constraints. All of the synthetic household totals match perfectly with the marginal totals. As a test of goodness-of-fit, the result shows zero percent normalized difference, zero sum of chi-square values and a value of R-square equal to one. The conclusion from the test could be drawn that the synthetic household total matches perfectly with the marginal household total despite of number of controls or variation of control attributes.

Table 1: Comparison of Marginal and Synthetic Household Totals for Different Combinations

| Total Household Marginal: 151675 | | | | | | |
|---|---|---|---|---|---|---|
| Combinations | Household constraints | Person constraints | Synthetic Household total | Normalized difference | Chi SQ ($\chi^2$) | $R^2$ |
| 1 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race, Employment status | 151675 | 0.00% | 0 | 1.00 |
| 2 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race | 151675 | 0.00% | 0 | 1.00 |
| 3 | HH Size, HH Income, HH Type, Child Presence | Age, Gender | 151675 | 0.00% | 0 | 1.00 |
| 4 | HH Size, HH Income, HH Type, Child Presence | Age | 151675 | 0.00% | 0 | 1.00 |
| 5 | HH Size, HH Income, HH Type, Child Presence | - | 151675 | 0.00% | 0 | 1.00 |
| 6 | HH Size, HH Income, HH Type | - | 151675 | 0.00% | 0 | 1.00 |
| 7 | HH Size, HH Income | - | 151675 | 0.00% | 0 | 1.00 |
| 8 | HH Size | - | 151675 | 0.00% | 0 | 1.00 |
| 9 | HH Size | Age | 151675 | 0.00% | 0 | 1.00 |
| 10 | HH Size, HH Income | Age | 151675 | 0.00% | 0 | 1.00 |
| 11 | HH Size, HH Income | Age, Gender | 151675 | 0.00% | 0 | 1.00 |

Table 2: Comparison of Marginal and Synthetic Person Totals for Different Combinations

| Total Person Marginal: 411414 | | | | | | |
|---|---|---|---|---|---|---|
| Combinations | Household constraints | Person constraints | Synthetic person total | Normalized Difference | Chi SQ ($\chi^2$) | $R^2$ |
| 1 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race, Employment status | 402595 | 2.14% | 648.07 | 0.9977 |
| 2 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race | 402771 | 2.10% | 633.51 | 0.9977 |
| 3 | HH Size, HH Income, HH Type, Child Presence | Age, Gender | 402529 | 2.16% | 652.03 | 0.9976 |
| 4 | HH Size, HH Income, HH Type, Child Presence | Age | 402982 | 2.05% | 622.08 | 0.9978 |
| 5 | HH Size, HH Income, HH Type, Child Presence | - | 402616 | 2.14% | 1879.91 | 0.9921 |
| 6 | HH Size, HH Income, HH Type | - | 402527 | 2.16% | 1888.26 | 0.9921 |
| 7 | HH Size, HH Income | - | 406406 | 1.22% | 1744.36 | 0.9921 |
| 8 | HH Size | - | 406682 | 1.15% | 1719.73 | 0.9922 |
| 9 | HH Size | Age | 410014 | 0.34% | 392.36 | 0.9979 |
| 10 | HH Size, HH Income | Age | 409952 | 0.36% | 394.36 | 0.9980 |
| 11 | HH Size, HH Income | Age, Gender | 409855 | 0.38% | 397.95 | 0.9979 |

Table 2 shows the comparison of the marginal and the synthetic persons for eleven combinations of constraints. From this test, a potential set of constraints are selected for further analysis based on the chi-square value, normalized percent difference and R-square values as goodness-of-fit measurement. Combination 9, 10 and 11 show the least percent difference and chi-square values and best R-square values. The table shows that fewer number of controls in both household and person levels yield better match of person totals.

### 5.1.1 Comparison of Household Attributes

In this section each individual household attribute is compared to understand how closely the combinations generate the synthetic households in different categories and how the procedure is being influenced by different controls.

Table 3 shows the comparison of different combinations in matching the household controlled variable: household size. For combination 7, 8, 9, 10 and 11 the synthetic households match very closely with the marginal values which means, least number of controls in both the household and the person level, yields better match.

Table 3: Comparison of Household Controlled Variable: Household Size for Different Combinations

| HH Size | 1 | 2 | 3 | 4 | 5 | 6 | 7+ | Sum of $\chi^2$ |
|---|---|---|---|---|---|---|---|---|
| Actual | 38189 | 49674 | 23550 | 20522 | 10447 | 5274 | 4019 | - |
| Comb_1 | 39086 | 50431 | 23540 | 20244 | 9959 | 4774 | 3641 | 142.13 |
| Comb_2 | 39086 | 50431 | 23540 | 20244 | 9959 | 4774 | 3641 | 142.13 |
| Comb_3 | 39086 | 50431 | 23540 | 20244 | 9959 | 4774 | 3641 | 142.13 |
| Comb_4 | 39086 | 50431 | 23540 | 20244 | 9959 | 4774 | 3641 | 142.13 |
| Comb_5 | 38694 | 50108 | 23480 | 20355 | 10196 | 5045 | 3797 | 40.27 |
| Comb_6 | 38694 | 50108 | 23480 | 20355 | 10196 | 5045 | 3797 | 40.27 |
| Comb_7 | 38181 | 49711 | 23564 | 20518 | 10428 | 5255 | 4018 | 0.14 |
| Comb_8 | 38189 | 49674 | 23550 | 20522 | 10447 | 5274 | 4019 | 0.00 |
| Comb_9 | 38189 | 49674 | 23550 | 20522 | 10447 | 5274 | 4019 | 0.00 |
| Comb_10 | 38181 | 49711 | 23564 | 20518 | 10428 | 5255 | 4018 | 0.14 |
| Comb_11 | 38181 | 49711 | 23564 | 20518 | 10428 | 5255 | 4018 | 0.14 |

Note: see page 42 for explanation of combinations

Table 4 shows comparison of household income for different combinations. For combination 8 and 9, the chi-square values are 3,135.72 and 1,361.54 compare to 12.49 for other combinations. These two values are very high compare to others, because only for these two combinations, household income variable was not used as a control. So we can conclude that we should take control variables according to our interest to get better results.

Table 4: Comparison of Household Income for Different Combinations

| HH Income | $0 - $14,999 | $15,000 - $24,999 | $25,000 - $34,999 | $35,000 - $44,999 | $45,000 - $59,999 | $60,000 - $99,999 | $100,000 - $149,999 | Over $ 150,000 | Sum of $\chi^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Actual | 18255 | 19307 | 19842 | 19048 | 21892 | 33782 | 12224 | 7325 | - |
| Comb_1 | 18187 | 19328 | 19881 | 19104 | 21938 | 34086 | 12044 | 7107 | 12.49 |
| Comb_2 | 18187 | 19328 | 19881 | 19104 | 21938 | 34086 | 12044 | 7107 | 12.49 |
| Comb_3 | 18187 | 19328 | 19881 | 19104 | 21938 | 34086 | 12044 | 7107 | 12.49 |
| Comb_4 | 18187 | 19328 | 19881 | 19104 | 21938 | 34086 | 12044 | 7107 | 12.49 |
| Comb_5 | 18244 | 19330 | 19872 | 19084 | 21959 | 33978 | 12063 | 7145 | 8.03 |
| Comb_6 | 18244 | 19330 | 19872 | 19084 | 21959 | 33978 | 12063 | 7145 | 8.03 |
| Comb_7 | 18242 | 19311 | 19856 | 19039 | 21910 | 33787 | 12223 | 7307 | 0.08 |
| Comb_8 | 23521 | 21497 | 21375 | 18558 | 21452 | 28319 | 10573 | 6380 | 3135.72 |
| Comb_9 | 22359 | 20101 | 19943 | 17251 | 21639 | 31421 | 11375 | 7586 | 1361.54 |
| Comb_10 | 18242 | 19311 | 19856 | 19039 | 21910 | 33787 | 12223 | 7307 | 0.08 |
| Comb_11 | 18242 | 19311 | 19856 | 19039 | 21910 | 33787 | 12223 | 7307 | 0.08 |

Note: see page 42 for explanation of combinations

Table 5: Comparison of Household Type for Different Combinations

| HH Type | Family: Married Couple | Family: Male Householder, No Wife | Family: Female Householder, No Husband | Non-family: Householder Alone | Non-family: Householder Not Alone | Sum of $\chi^2$ |
|---|---|---|---|---|---|---|
| Actual | 78973 | 7648 | 15419 | 38189 | 11446 | - |
| Comb_1 | 80818 | 6278 | 14559 | 38486 | 11534 | 339.47 |
| Comb_2 | 80818 | 6278 | 14559 | 38486 | 11534 | 339.47 |
| Comb_3 | 80818 | 6278 | 14559 | 38486 | 11534 | 339.47 |
| Comb_4 | 80818 | 6278 | 14559 | 38486 | 11534 | 339.47 |
| Comb_5 | 80310 | 6986 | 15058 | 38061 | 11260 | 91.84 |
| Comb_6 | 80310 | 6986 | 15058 | 38061 | 11260 | 91.84 |
| Comb_7 | 82013 | 6703 | 15122 | 22936 | 24901 | 22148.30 |
| Comb_8 | 80345 | 7010 | 16511 | 22642 | 25167 | 22931.86 |
| Comb_9 | 80899 | 6339 | 6956 | 29257 | 28224 | 31599.04 |
| Comb_10 | 81965 | 6267 | 6520 | 29202 | 27721 | 30754.97 |
| Comb_11 | 43964 | 11247 | 22123 | 34739 | 39602 | 89700.63 |

Note: see page 42 for explanation of combinations

Table 5, 6, 7, and 8 shows comparison of marginal and synthetic households for household type, household children presence, householder race, and householder age respectively. The synthetic household type is close with the marginal totals for those combinations where household type variable was taken as a control. Householder age and householder race are uncontrolled for all combinations. However in combination 1 and 2 householder race shows better result. This is because in person level persons' race was controlled. For the categories of householder age stochastic behavior is observed.

Table 6: Comparison of Household Children Presence for Different Combinations

| | Children Present | Children Not Present | Sum of $\chi^2$ |
|---|---|---|---|
| Actual | 51427 | 100248 | - |
| Comb_1 | 51804 | 99871 | 4.18148 |
| Comb_2 | 51804 | 99871 | 4.18148 |
| Comb_3 | 51804 | 99871 | 4.18148 |
| Comb_4 | 51804 | 99871 | 4.18148 |
| Comb_5 | 49108 | 102567 | 158.215 |
| Comb_6 | 49121 | 102554 | 156.446 |
| Comb_7 | 49368 | 102307 | 124.727 |
| Comb_8 | 50302 | 101373 | 37.2351 |
| Comb_9 | 39264 | 112411 | 4352.4 |
| Comb_10 | 39898 | 111777 | 3910.48 |
| Comb_11 | 35861 | 115814 | 7128.55 |

Table 7: Comparison of Householder Age for Different Combinations

| Householder Age | < 65 years | >= 65 years | Sum of $\chi^2$ |
|---|---|---|---|
| Actual | 122163 | 29512 | - |
| Comb_1 | 124295 | 27380 | 191.227 |
| Comb_2 | 124333 | 27342 | 198.105 |
| Comb_3 | 124321 | 27354 | 195.92 |
| Comb_4 | 123286 | 28389 | 53.0561 |
| Comb_5 | 121832 | 29843 | 4.60926 |
| Comb_6 | 121502 | 30173 | 18.3814 |
| Comb_7 | 118555 | 33120 | 547.657 |
| Comb_8 | 117184 | 34491 | 1042.94 |
| Comb_9 | 123096 | 28579 | 36.6217 |
| Comb_10 | 123458 | 28217 | 70.553 |
| Comb_11 | 121980 | 29695 | 1.40889 |

Table 8: Comparison of Householder Race for Different Combinations

| Householder Race | White | Black or African American | American Indian and Alaska Native | Asian | Native Hawaiian and other Pacific Islander | Some Other Race | Two or More Races | Sum of $\chi^2$ |
|---|---|---|---|---|---|---|---|---|
| Actual | 126170 | 5394 | 1716 | 2697 | 138 | 12264 | 2568 | - |
| Comb_1 | 125835 | 5602 | 1948 | 2940 | 228 | 12157 | 2965 | 183.2 |
| Comb_2 | 126509 | 5546 | 1736 | 2832 | 204 | 11921 | 2927 | 103.5 |
| Comb_3 | 124909 | 4058 | 4891 | 2386 | 132 | 11755 | 3544 | 6646.2 |
| Comb_4 | 125446 | 4040 | 5167 | 2508 | 107 | 11246 | 3161 | 7525.9 |
| Comb_5 | 124437 | 4003 | 5800 | 2232 | 114 | 11903 | 3186 | 10345.9 |
| Comb_6 | 124403 | 4020 | 5739 | 2280 | 99 | 11896 | 3238 | 10067.6 |
| Comb_7 | 125118 | 3929 | 5609 | 2143 | 114 | 11735 | 3027 | 9461.3 |
| Comb_8 | 123489 | 4032 | 6256 | 2216 | 96 | 12388 | 3198 | 12666.7 |
| Comb_9 | 126285 | 3934 | 4771 | 2650 | 93 | 10401 | 3541 | 6501.3 |
| Comb_10 | 127063 | 3863 | 4492 | 2551 | 103 | 10161 | 3442 | 5606.5 |
| Comb_11 | 124138 | 5836 | 4676 | 2670 | 73 | 10464 | 3818 | 6078.3 |

Note: see page 42 for explanation of combinations

Table 9 shows the comparison of the marginal and the synthetic totals for group quarters. All of the chi-square values are zero. It can be seen that the estimation of institutionalized population and non-institutionalized population is exactly the same as actual values i.e. no change is observed in group quarters' synthetic populations for change in controls. Therefore, for further tests group quarters are being dropped out.

Table 9: Comparison of Group Quarter Population for Different Combinations

| Group Quarter | Institutionalized Population | Non-Institutionalized Population | Total | Sum of $\chi^2$ |
|---|---|---|---|---|
| Actual | 2120 | 3899 | 6019 | - |
| Comb_1 | 2120 | 3899 | 6019 | 0 |
| Comb_2 | 2120 | 3899 | 6019 | 0 |
| Comb_3 | 2120 | 3899 | 6019 | 0 |
| Comb_4 | 2120 | 3899 | 6019 | 0 |
| Comb_5 | 2120 | 3899 | 6019 | 0 |
| Comb_6 | 2120 | 3899 | 6019 | 0 |
| Comb_7 | 2120 | 3899 | 6019 | 0 |
| Comb_8 | 2120 | 3899 | 6019 | 0 |
| Comb_9 | 2120 | 3899 | 6019 | 0 |
| Comb_10 | 2120 | 3899 | 6019 | 0 |
| Comb_11 | 2120 | 3899 | 6019 | 0 |

### 5.1.2 Comparison of Person Attributes

In this section each individual person attribute is compared to understand how closely the combinations generate the synthetic persons in different categories, and how the procedure is being influenced by different controls.

Table 10 shows the comparison of different combinations in matching persons' age variable. It can be seen that chi-squares are less for combinations 1 to 4 and 9 to 11. For combinations 5 to 8 chi-square values are very high because of not controlling any person variables. Again combinations 9 to 11 have least sum of chi-squares because the total number of controls is the least for these combinations.

Table 10: Comparison of Person Age for Different Combinations

| Age | Under 5 | 5 to 14 | 15 to 24 | 25 to 34 | 35 to 44 | 45 to 54 | 55 to 64 | 65 to74 | 75 to 84 | 85 and more | Sum of $\chi^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Actual | 32145 | 62405 | 57638 | 65716 | 64755 | 49401 | 31043 | 23853 | 17642 | 6816 | - |
| Comb_1 | 31376 | 60545 | 53530 | 64084 | 64405 | 49604 | 30767 | 23965 | 17881 | 6438 | 437.06 |
| Comb_2 | 30924 | 60762 | 54121 | 63562 | 64506 | 49675 | 31185 | 23803 | 17784 | 6449 | 398.98 |
| Comb_3 | 31197 | 59818 | 54756 | 64314 | 64004 | 49397 | 31025 | 23873 | 17762 | 6383 | 346.28 |
| Comb_4 | 31826 | 60316 | 54987 | 64047 | 63772 | 49060 | 30878 | 23683 | 17663 | 6750 | 257.44 |
| Comb_5 | 30388 | 62632 | 53750 | 56349 | 63538 | 52755 | 36093 | 26905 | 16004 | 4202 | 4311.47 |
| Comb_6 | 30345 | 62348 | 53977 | 56702 | 62756 | 53231 | 35533 | 27168 | 16132 | 4335 | 4070.89 |
| Comb_7 | 30891 | 63443 | 54054 | 56015 | 61727 | 52687 | 36674 | 28273 | 17744 | 4898 | 4462.04 |
| Comb_8 | 31826 | 64883 | 54943 | 55741 | 60625 | 50260 | 35777 | 29099 | 18476 | 5052 | 4391.65 |
| Comb_9 | 31974 | 62031 | 56944 | 65744 | 64628 | 49468 | 31086 | 23724 | 17644 | 6771 | 12.91 |
| Comb_10 | 32091 | 61996 | 57280 | 65733 | 64679 | 49219 | 31031 | 23615 | 17570 | 6738 | 9.32 |
| Comb_11 | 31886 | 62251 | 57340 | 65310 | 64582 | 49461 | 31142 | 23817 | 17370 | 6696 | 13.73 |

Note: see page 42 for explanation of combinations

Table 11 shows the comparison of marginal and synthetic totals for persons' gender. Combination 11 shows the best result because of two possible reasons: least number of control variables and control of gender variable. Table 12 shows the comparison of marginal and synthetic totals for persons' employment status. The results show stochastic behavior from where no conclusion could be drawn.

Table 11: Comparison of Persons' Gender for Different Combinations

| Gender | Male | Female | Sum of $\chi^2$ |
|---|---|---|---|
| Actual | 203903 | 207511 | - |
| Comb_1 | 204212 | 198383 | 401.99 |
| Comb_2 | 204245 | 198526 | 389.61 |
| Comb_3 | 204634 | 197895 | 448.22 |
| Comb_4 | 215252 | 187730 | 2517.30 |
| Comb_5 | 213043 | 189573 | 1960.33 |
| Comb_6 | 212865 | 189662 | 1929.18 |
| Comb_7 | 200918 | 205488 | 63.42 |
| Comb_8 | 199835 | 206847 | 83.28 |
| Comb_9 | 213814 | 196200 | 1098.28 |
| Comb_10 | 214804 | 195148 | 1319.34 |
| Comb_11 | 202377 | 207478 | 11.43 |

Note: see page 42 for explanation of combinations

Table 12: Comparison of Persons' Employment Status for Different Combinations

| Employment Status | Not Eligible | Employed | Unemployed | Not In Labor Force | Sum of $\chi^2$ |
|---|---|---|---|---|---|
| Actual | 99936 | 193358 | 10221 | 107899 | - |
| Comb_1 | 97156 | 190245 | 9627 | 105567 | 212.37 |
| Comb_2 | 97114 | 186877 | 9974 | 108806 | 310.51 |
| Comb_3 | 96279 | 186129 | 9889 | 110232 | 465.32 |
| Comb_4 | 97172 | 186589 | 10343 | 108878 | 323.75 |
| Comb_5 | 98868 | 183800 | 10001 | 109947 | 527.49 |
| Comb_6 | 98736 | 184126 | 10100 | 109565 | 482.35 |
| Comb_7 | 100447 | 182801 | 9979 | 113179 | 843.11 |
| Comb_8 | 102874 | 175832 | 10585 | 117391 | 2522.92 |
| Comb_9 | 95141 | 191957 | 10199 | 112717 | 455.40 |
| Comb_10 | 95346 | 195086 | 9846 | 109674 | 269.22 |
| Comb_11 | 94969 | 200998 | 8254 | 105634 | 974.83 |

Note: see page 42 for explanation of combinations

Table 13 shows the comparison of marginal and synthetic totals for persons' race. Combination 1 and 2 show better results because race variable was controlled in the person level for these two combinations. All other combinations show very high chi square values.

Table 13: Comparison of Persons' Race for Different Combinations

| Race | White | Black or African American | American Indian and Alaska Native | Asian | Native Hawaiian and other Pacific Islander | Some Other Race | Two or More Races | Sum of $\chi^2$ |
|---|---|---|---|---|---|---|---|---|
| Actual | 319265 | 15898 | 5886 | 8830 | 649 | 49797 | 11089 | - |
| Comb_1 | 315626 | 14870 | 5394 | 8533 | 633 | 47204 | 10335 | 345.75 |
| Comb_2 | 316692 | 14824 | 5054 | 8355 | 582 | 47023 | 10241 | 462.74 |
| Comb_3 | 310731 | 10273 | 15898 | 7723 | 407 | 45169 | 12328 | 20046.18 |
| Comb_4 | 312825 | 10830 | 17956 | 7936 | 354 | 41120 | 11961 | 28301.70 |
| Comb_5 | 305141 | 10335 | 21244 | 7082 | 347 | 46108 | 12359 | 43549.47 |
| Comb_6 | 305092 | 10270 | 20730 | 7014 | 401 | 46593 | 12427 | 40892.70 |
| Comb_7 | 309457 | 10291 | 21362 | 6852 | 364 | 45933 | 12147 | 43938.72 |
| Comb_8 | 304003 | 10710 | 23173 | 6916 | 334 | 48620 | 12926 | 54093.87 |
| Comb_9 | 322291 | 10881 | 16368 | 8585 | 344 | 39506 | 12039 | 22636.89 |
| Comb_10 | 324605 | 10796 | 15349 | 8033 | 337 | 38587 | 12245 | 19806.41 |
| Comb_11 | 314091 | 17353 | 14847 | 7880 | 332 | 41505 | 13847 | 16183.22 |

Note: see page 42 for explanation of combinations

### 5.2 Analysis with Alternative Rounding Methods

This section discusses about the results of two rounding methods: Arithmetic and Bucket rounding. The analysis was done for all the eleven combinations. Variation was found in generating person totals by two different rounding methods. Table 14 shows the comparison of marginal and synthetic person totals for two different rounding methods for the selected eleven combinations of household and person constraints. Based on chi-square value and normalized percent difference as goodness-of-fit measurements, it can be seen that in total person comparison, bucket rounding method generates synthetic population more accurate compare to arithmetic rounding method.

Table 14: Comparison of Person Totals for Arithmetic and Bucket Rounding Methods

Marginal Person Total: 411,414

| Combinations | Household constraints | Person constraints | Arithmetic | | | Bucket | | |
|---|---|---|---|---|---|---|---|---|
| | | | Total Persons | Chi SQ ($\chi^2$) | Normalized Difference | Total Persons | Chi SQ ($\chi^2$) | Normalized Difference |
| 1 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race, Employment | 402595 | 648 | 2.14% | 409680 | 400 | 0.42% |
| 2 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race | 402771 | 634 | 2.10% | 409791 | 398 | 0.39% |
| 3 | HH Size, HH Income, HH Type, Child Presence | Age, Gender | 402529 | 652 | 2.16% | 409563 | 405 | 0.45% |
| 4 | HH Size, HH Income, HH Type, Child Presence | Age | 402982 | 622 | 2.05% | 410060 | 386 | 0.33% |
| 5 | HH Size, HH Income, HH Type, Child Presence | - | 402616 | 1880 | 2.14% | 406597 | 1737 | 1.17% |
| 6 | HH Size, HH Income, HH Type | - | 402527 | 1888 | 2.16% | 406599 | 1740 | 1.17% |
| 7 | HH Size, HH Income | - | 406406 | 1744 | 1.22% | 406424 | 1746 | 1.21% |
| 8 | HH Size | - | 406682 | 1720 | 1.15% | 406384 | 1734 | 1.22% |
| 9 | HH Size | Age | 410014 | 392 | 0.34% | 410062 | 387 | 0.33% |
| 10 | HH Size, HH Income | Age | 409952 | 394 | 0.36% | 409927 | 395 | 0.36% |
| 11 | HH Size, HH Income | Age, Gender | 409855 | 398 | 0.38% | 409893 | 399 | 0.37% |

### 5.2.1 Comparison of Household Attributes

As from the previous section it can be seen that, with higher number of controls like in combination 1, 2 and 3 the accuracy for generating synthetic population by arithmetic rounding is lower compare to other combinations with fewer number of controls. So, an attempt was taken to see how the results improve in household and person level attribute with bucket rounding. Table 15 shows the comparison of bucket and arithmetic rounding for combinations 1, 2 and 3. The comparison is based on sum of chi-square value where lower sum of chi-square value represents a better fit to actual number of households to each categories of that household attribute. It can be seen that bucket rounding lowered the sum of chi-square in a significant amount for the attributes household size, household income, household type, and household children presence. This means, the number of synthetic households generated by bucket rounding method is very accurate to the actual number of households in each category of these household attributes. For the attribute householder age, bucket rounding is more accurate compare to arithmetic rounding as well, but the improvement is smaller compared to other variables. Figure 19 and Figure 20 shows the comparison of bucket and arithmetic rounding for controlled variable: household size, and uncontrolled variable: householder age respectively.

Table 15: Comparison of Household Attributes for Bucket and Arithmetic Rounding

| Household Variables | Sum of $\chi^2$ for Combination_1 | | Sum of $\chi^2$ for Combination_2 | | Sum of $\chi^2$ for Combination_3 | |
|---|---|---|---|---|---|---|
| | Bucket | Arithmetic | Bucket | Arithmetic | Bucket | Arithmetic |
| HH Size | 0.1 | 142.1 | 0.1 | 142.1 | 0.1 | 142.1 |
| HH Income | 0.0 | 12.5 | 0.0 | 12.5 | 0.0 | 12.5 |
| HHld Type | 0.2 | 339.5 | 0.2 | 339.5 | 0.2 | 339.5 |
| Child Presence | 0.0 | 4.2 | 0.0 | 4.2 | 0.0 | 4.2 |
| HHldr Age | 190.0 | 191.2 | 197.3 | 198.1 | 187.5 | 195.9 |

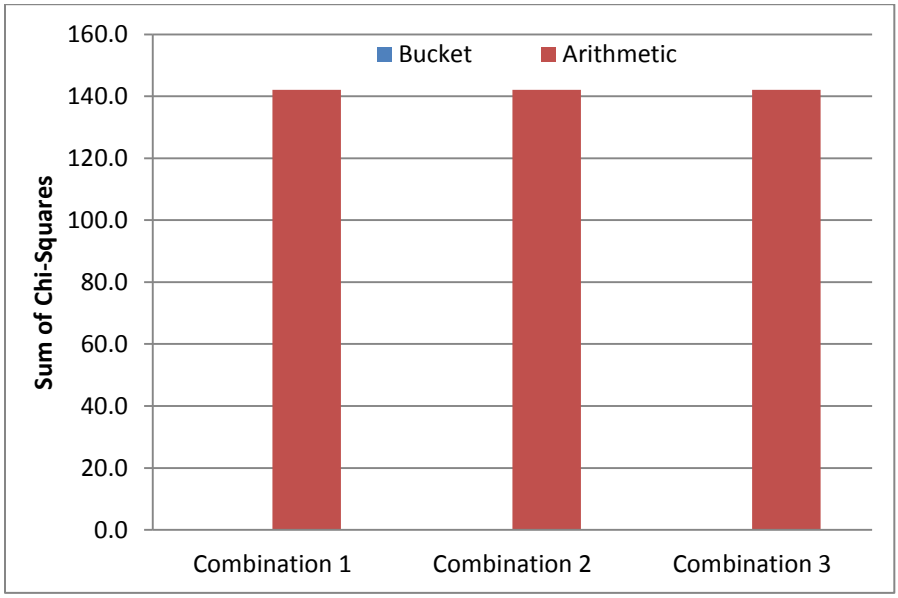


Figure 19: Comparison of Controlled Variable: Household Size for Bucket and Arithmetic Rounding

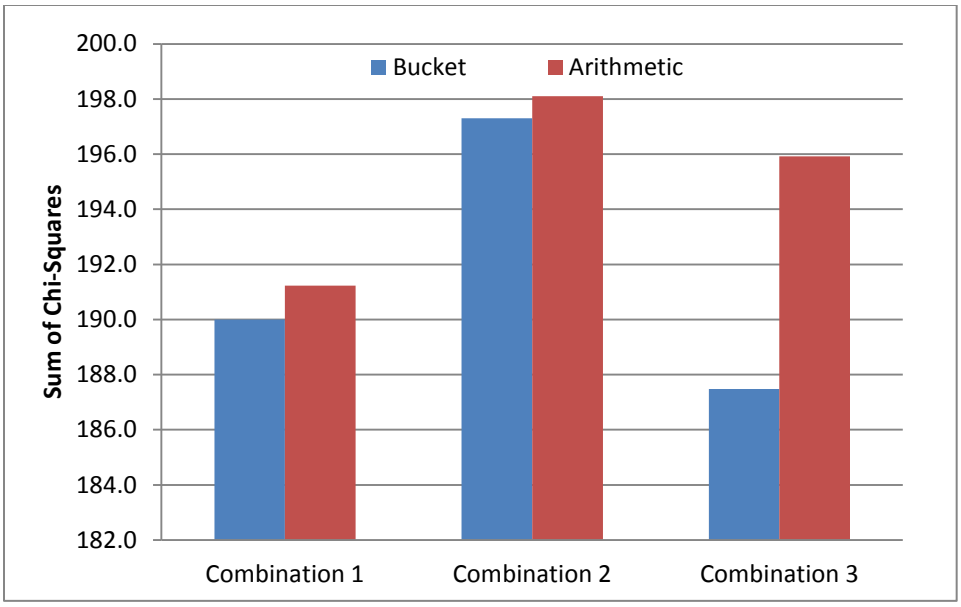


Figure 20: Comparison of Uncontrolled Variable: Householder Age for Bucket and Arithmetic Rounding

### 5.2.2 Comparison of Person Attributes

For this section, again, combinations with higher number of controls, i.e. combination 1, 2, and 3, where accuracy for generating synthetic population by arithmetic rounding is lower compare to the other combinations with fewer numbers of controls, were selected for individual person level comparison. A comparison between the bucket and arithmetic rounding was made to see how the bucket rounding improves the accuracy of the numbers of persons in each person category. Table 16 shows the comparison of the person attributes for bucket and arithmetic rounding for combinations 1, 2 and 3. The comparison is based on sum of chi-square value where lower sum of chi-square value represents a better fit to actual number of persons to each categories of that person attribute. It can be seen that bucket rounding lowered the sum of chi-square in a significant amount for the

attribute person age. This means, the number of synthetic persons generated by bucket rounding method is closer to the actual number of persons in each category of this person attribute compared to arithmetic rounding. For the attributes person's gender and employment status, bucket rounding is again, more accurate compare to arithmetic rounding, but the improvement is smaller for the variable gender compared to the variable employment status. Figure 21, Figure 22 and Figure 23 shows the comparison of bucket and arithmetic rounding for person attributes age, gender and employment status.

Table 16: Comparison of Person Attributes for Bucket and Arithmetic Rounding

| Person Variables | Sum of $\chi^2$ for Combination_1 | | Sum of $\chi^2$ for Combination_2 | | Sum of $\chi^2$ for Combination_3 | |
|---|---|---|---|---|---|---|
| | Bucket | Arithmetic | Bucket | Arithmetic | Bucket | Arithmetic |
| Age | 60.6 | 437.1 | 37.2 | 399.0 | 49.5 | 346.3 |
| Gender | 244.2 | 402.0 | 319.8 | 389.6 | 300.5 | 448.2 |
| Employment | 41.1 | 212.4 | 144.5 | 310.5 | 291.9 | 465.3 |

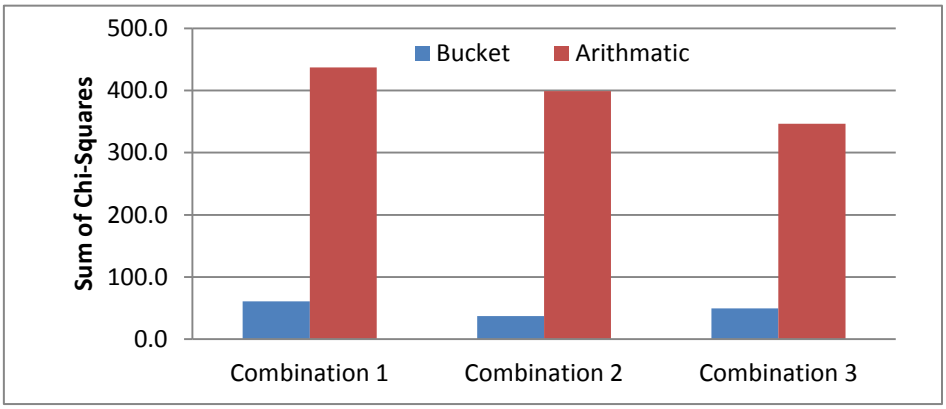


Figure 21: Comparison of Person's Age for Bucket and Arithmetic Rounding

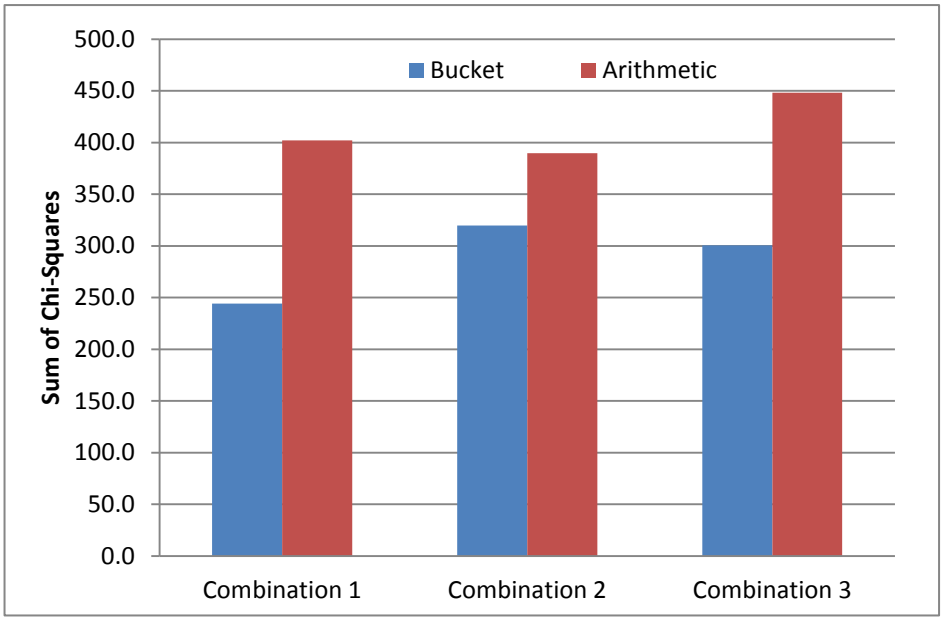


Figure 22: Comparison of Person's Gender for Bucket and Arithmetic Rounding

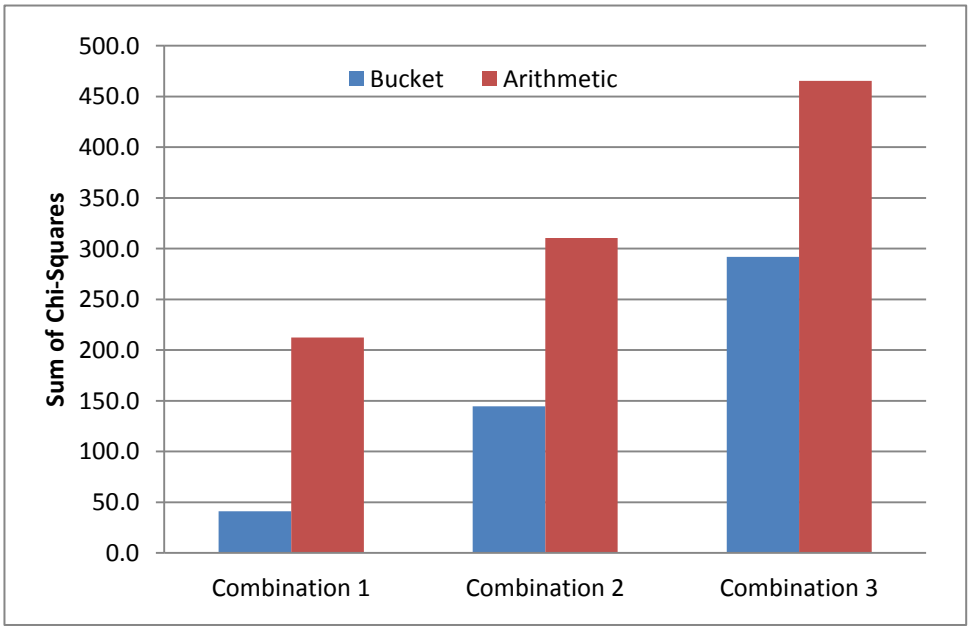


Figure 23: Comparison of Employment Status for Bucket and Arithmetic Rounding

### 5.3 Analysis with Alternative Sample Data

For this section two different sample inputs were prepared from Census 2000 and ACS 2005-2009 dataset, one is only census sample and the other one is combined Census and ACS sample. Simulation was run for the combinations 1, 2, and 3 i.e. combinations with higher number of controls for both household and person. The motivation was to investigate whether any change in accuracy is possible with a richer sample data where the software will have more flexibility in drawing households.

Table 17 shows the comparison of marginal and synthetic person totals generated by combined sample input and census sample input for the selected three combinations of household and person constraints. Based on chi-square value and normalized percent difference as goodness of fit measures, it can be seen that accuracy for generating synthetic person totals for combined sample is better compared to census sample.

Table 17: Comparison of Person Totals for Combined and Census Samples

| Marginal Person Total: 411,414 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Combinations | Household constraints | Person constraints | Census Sample | | | Combined Sample | | |
| | | | Total Persons | Chi SQ ($\chi^2$) | Normalized Difference | Total Persons | Chi SQ ($\chi^2$) | Normalized Difference |
| 1 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race, employment | 402595 | 648 | 2.14% | 402995 | 617 | 2.05% |
| 2 | HH Size, HH Income, HH Type, Child Presence | Age, Gender, Race | 402771 | 634 | 2.10% | 403022 | 611 | 2.04% |
| 3 | HH Size, HH Income, HH Type, Child Presence | Age, Gender | 402529 | 652 | 2.16% | 402920 | 622 | 2.06% |

### 5.3.1 Comparison of Household Attributes

As from the first section of this chapter, it can be seen that, with higher number of controls the accuracy for generating synthetic population by census sample input is lower compare to other combinations with fewer number of controls, an attempt was taken to see how the results improve in household and person level attribute with combined sample input. Table 18 shows the comparison of combined and census sample inputs for combinations 1, 2 and 3. The comparison is based on sum of chi-square value where lower sum of chi-square value represents a better fit to actual number of households to each categories of that household attribute. It can be seen that combined sample input lowered the sum of chi-square value for the attributes household size, household income, household type, and household children presence. This means, the number of synthetic households generated by combined sample input is slightly closer to the actual number of households in each category of these household attributes. For the attribute householder age, combined sample input was found to be significantly lowering the sum of chi-square value. So, for all household attributes, combined sample input generates more accurate synthetic population than census sample input. Figure 24, Figure 25 and Figure 26 show the comparison of combined and census sample inputs for controlled variables: household size, and household income and uncontrolled variable: householder age respectively.

Table 18: Comparison of Household Attributes for Combined and Census Samples

| Household Variables | Sum of $\chi^2$ for Combination_1 | | Sum of $\chi^2$ for Combination_2 | | Sum of $\chi^2$ for Combination_3 | |
|---|---|---|---|---|---|---|
| | Combined | Census | Combined | Census | Combined | Census |
| HH Size | 111.2 | 142.1 | 111.2 | 142.1 | 111.2 | 142.1 |
| HH Income | 10.4 | 12.5 | 10.4 | 12.5 | 10.4 | 12.5 |
| HHld Type | 304.1 | 339.5 | 304.1 | 339.5 | 304.1 | 339.5 |
| Child Presence | 1.2 | 4.2 | 1.2 | 4.2 | 1.2 | 4.2 |
| HHldr Age | 9.5 | 191.2 | 7.4 | 198.1 | 9.3 | 195.9 |

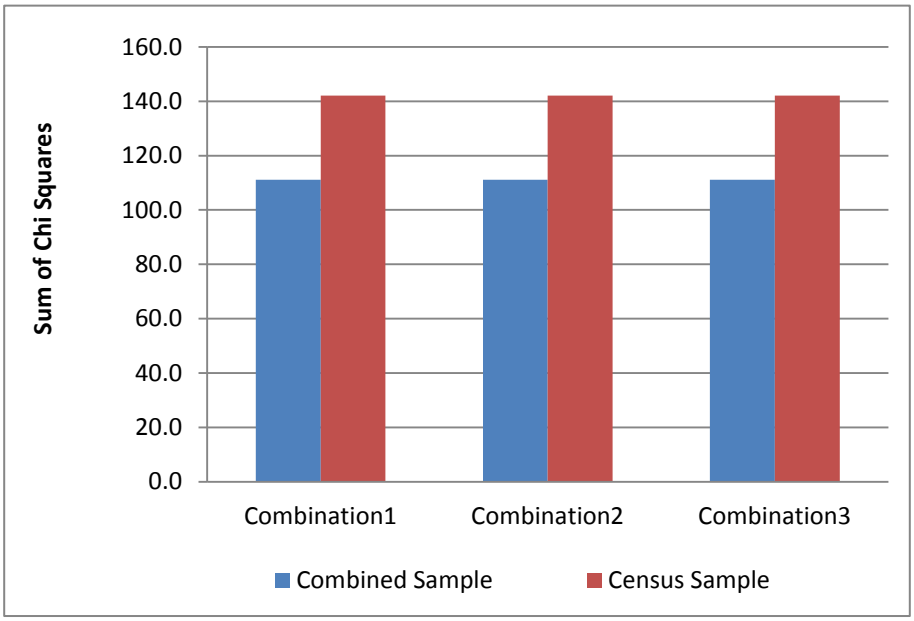


Figure 24: Comparison of Controlled Variable: Household Size for Combined and Census Samples

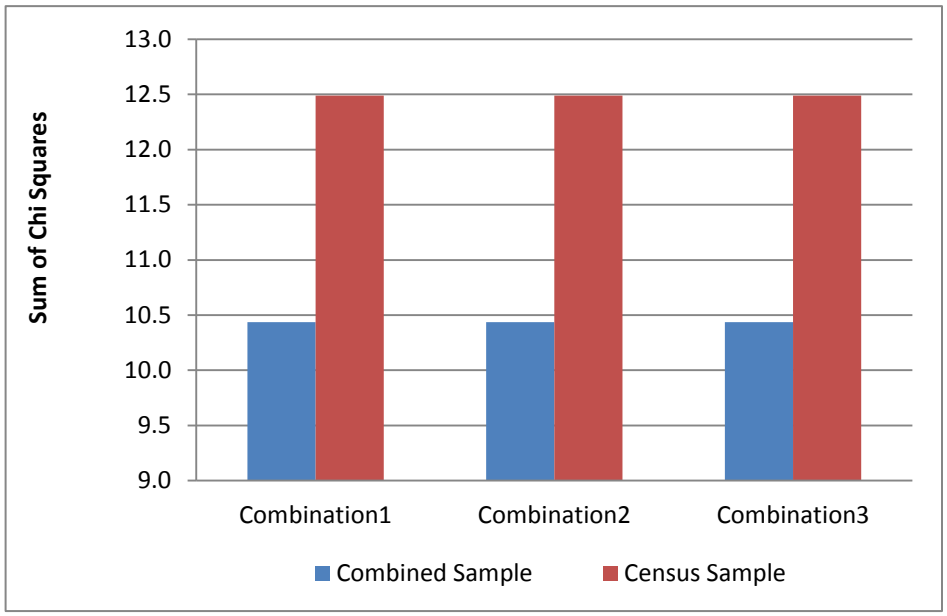


Figure 25: Comparison of Controlled Variable: Household Income for Combined and Census Samples

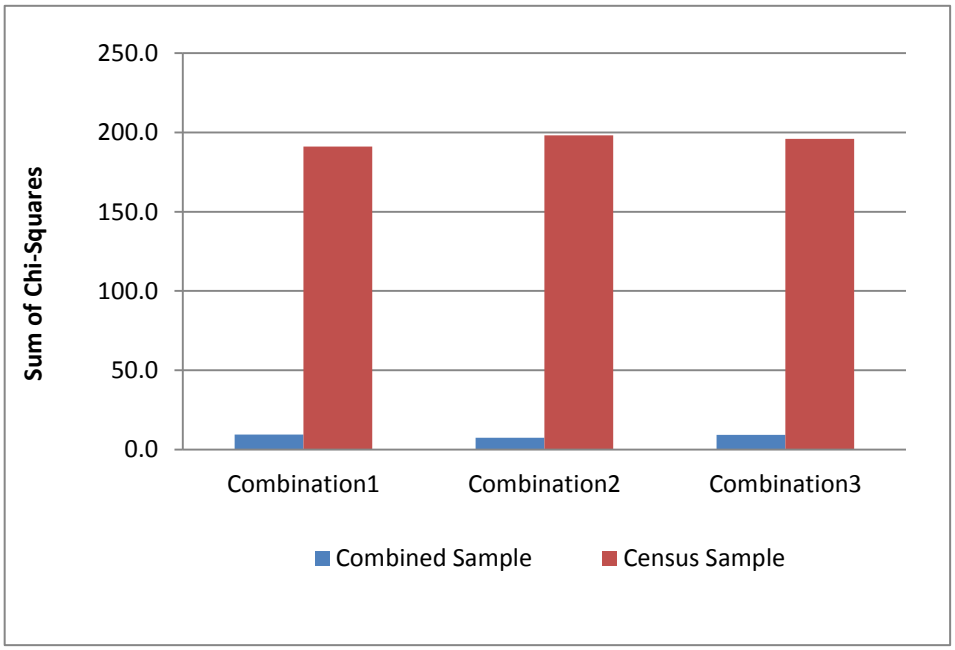


Figure 26: Comparison of Uncontrolled Variable: Householder Age for Combined and Census Samples

### 5.3.2 Comparison of Person Attributes

For this section, again, combinations with higher number of controls, i.e. combination 1, 2, and 3 were selected for individual person level comparison. Accuracy for generating synthetic population by census sample is lower in these combinations compare to the other combinations with fewer numbers of controls. A comparison between the combined and census samples was made to see how the result in each person category improves while a rich sample record is used. Table 19 shows the comparison of the person attributes for combined and census samples for combinations 1, 2 and 3. Again, the comparison is based on sum of chi-square value where lower sum of chi-square value represents a better fit to actual number of persons to each categories of that person attribute. It can be seen that combined sample lowered the sum of chi-square for person attributes age, gender and employment status. This means, the number of synthetic persons generated by combined sample is closer to the actual number of persons in each category of this person attribute compared to census sample. Figure 27, Figure 28 and Figure 29 shows the comparison of bucket and arithmetic rounding for person attributes age, gender and employment status.

Table 19: Comparison of Person Attributes for Combined and Census Samples

| Person Variables | Combination_1 | | Combination_2 | | Combination_3 | |
|---|---|---|---|---|---|---|
| | Combined | Census | Combined | Census | Combined | Census |
| Age | 437 | 437.1 | 387 | 399 | 314.3 | 346.3 |
| Gender | 191.5 | 402 | 195.7 | 389.6 | 190.2 | 448.2 |
| Employment | 211.1 | 212.4 | 304.5 | 310.5 | 463.5 | 465.3 |

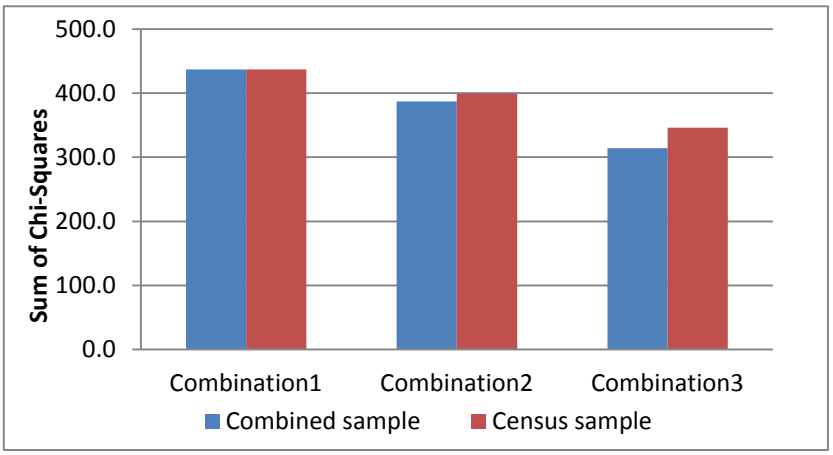


Figure 27: Comparison of Person's Age for Combined and Census Samples

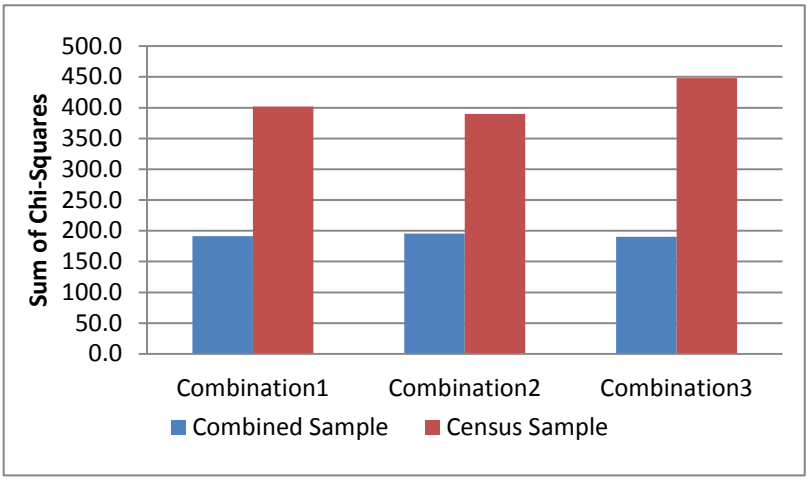


Figure 28: Comparison of Person's Gender for Combined and Census Samples

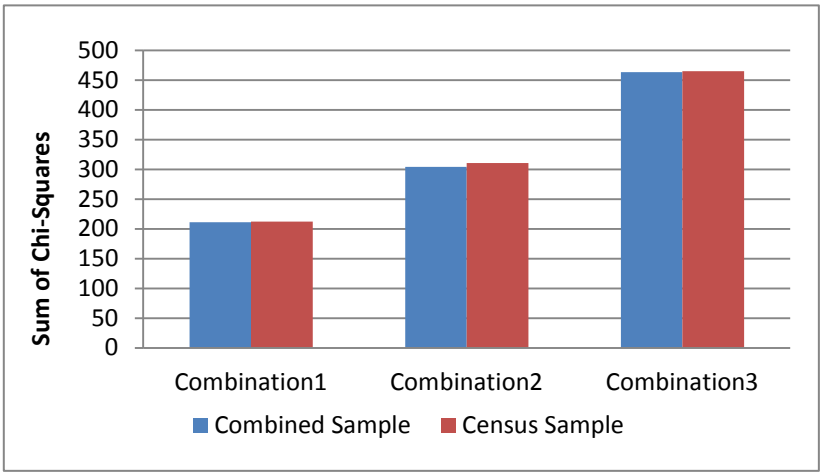


Figure 29: Comparison of Employment Status for Combined and Census Samples

## 5. CONCLUSIONS

The IPU algorithm provides a flexible mechanism for generating a synthetic population where both household and person-level attribute distributions can be matched very closely. The IPU algorithm works with joint distributions of household and person attributes derived using the IPF method, and then iteratively adjusts and reallocates weights across households such that both household and person-level attribute distributions are matched as closely as possible. The algorithm is flexible in that it can accommodate a multitude of household and person level variables of interest and meets dual household and person level constraints with reasonable computational time.

Across all the tests it was found that household level marginal totals are always matched perfectly, as the IPU algorithm is designed so. However person totals were always underestimated compare to the actual or marginal totals. From the eleven combinations of alternative constraints it was found that a fewer number of controls both in household and person levels yields a better match of person totals and it is always better to use both household and person controls than to use only household controls for more accurate synthetic person totals.

Further investigation on categorized household totals for both controlled and uncontrolled household variables shows that household size which was the only common controlled variable for all combinations matches better when the number of controls both in household and person level is fewer. Household income, which was a controlled variable for nine combinations, was matched better in those nine combinations where it was controlled compared to the other

two combinations where household income was not controlled. However in practice, a higher number of attributes need to be controlled since in most of the transportation models comprise with several number of person and household attributes.

The second test was done to improve the accuracy of the synthetic population with higher number of controlled variables both in household and person level by changing the rounding method from arithmetic to bucket. First three combinations were selected from the alternative constraints list where number of controls is higher. Bucket rounding was found to improve the accuracy in generating synthetic population significantly. It estimates more accurate synthetic populations for the comparisons of total persons as well as categorized controlled and uncontrolled variables both in household and person levels.

Finally, two different types of sample inputs were computed from the Census 2000 and the ACS 2005-2009 dataset. These samples were feed into PopGen with the Census 2000 marginal data. Again first three combinations were selected from the alternative constraint list with higher number of controls. The combined sample provides more flexibility and a rich seed matrix to PopGen in drawing households. Combined sample was found to improve the accuracy in generating synthetic population significantly. It generates synthetic populations more closely matched with both aggregate total and disaggregate level for both household and person level attributes.

## REFERENCES


Abraham, J. E., K. J. Stefan, and J. D. Hunt (2012) Population Synthesis Using Combinatorial Optimization at Multiple Levels. Presented at the *91st Annual Meeting of Transportation Research Board*, Washington, D.C.

Arentze, T., H. J. Timmermans, and F. Hofman (2007) Creating Synthetic Household Populations: Problems and Approach. In *Transportation Research Record: Journal of Transportation Research Board, No. 2014*, pp. 85-91.

Auld, J. and A. K. Mohammadian (2010) An Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. Presented at the *89th Annual Meeting of Transportation Research Board,* Washington, D.C.

Bar-Gera, H, K.C. Konduri, B. Sana, X. Ye, and R. M. Pendyala (2009) Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. Presented at the *88th Annual Meeting of Transportation Research Board,* Washington, D.C.

Beckman, R. J., K. A. Baggerly, and M. D. McKay (1996) Creating Synthetic Baseline Populations. *Transportation Research Part A*, Vol. 30, No. 6, pp. 415-429

Deming, W. E., and F. F. Stephan (1940) On a Least Squares Adjustment of a Sampled Frequency Table when the Expected Marginal Totals are Known. In *The Annals of Mathematical Statistics*, Vol. 11, No. 4, pp. 427-444.

Dial, R. B. (1971) A probabilistic Multipath Traffic Assignment Algorithm which Obviates Path Enumeration, *Transportation Research 5*, pp. 83-111.

Evers, L., and D. Santapaola (2007) On the Use of the IPF Algorithm for Combining Traffic Count Data with Missing Dimensions. In *Transportation Research Record: Journal of Transportation Research Board, No. 1993,* Washington, D.C., pp. 95-100.

Fang, S. C., and H. -S. J. Tsao (1995) Linearly-Constrained Entropy Maximization Problem with Quadratic Cost and Its Applications to Transportation Planning Problems. *Transportation Science 29(4),* pp. 353-365.

Guo, J. Y., and C. P. Bhat (2007) Population Synthesis for Microsimulating Travel Behavior. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2014*, pp. 92-101

Huang, Z. and P. Williamson (2001) Comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata, *Working Paper*, 2001/2, Department of Geography, University of Liverpool, UK.

Jornsten, K. O., and J. T. Lundgren (1989) An Entropy-Based Modal Split Model. *Transportation Research Part B 23(5),* pp. 345-349.

Kao, S. Chieh., H. K. Kim, C. Liu, X. Cui, and B. L. Bhaduri (2012) A Dependence-Preserving Approach in Synthesizing Household Characteristics. Presented at the *91st Annual Meeting of the Transportation Research Board,* Washington, D.C.

Konduri, K. C., B. Sana, X. Ye, and R. M. Pendyala (2010) Synthetic Population Generation for Travel Demand Forecasting. Presented at the *PopGen 1.1 Training Workshop*, Tempe, AZ.

Li, Y., T. Ziliaskopoulos, and D. Boyce (2002) Combined Model for Time-dependent Trip Distribution and Traffic Assignment, *Transportation Research Record 1783*, pp. 98-110.

Mohannadian, A., M. Javanmardi, and Y. Zhang (2010) Synthetic Household Travel Survey Data Simulation. In *Transportation Research Part C*, Article in Press

Müller, K. and K. W. Axhausen (2010) Population synthesis for microsimulation: State of the art. Presented at the *Swiss Transport Research Conference*

Oppenheim, N. (1995) Urban Travel Demand Modeling: From Individual Choices to General Equilibrium, *Wiley-Interscience Publication*.

Ryan, J., H. Maoh and P. Kanaroglou (2009) Population synthesis: Comparing the major techniques using a small, complete population of firms, *Geographical Analysis*, 41 (2) pp.181–203

Rossi, T. F., S. McNeil, and C. Hendrickson (1989) Entropy Model for Consistent Impact Fee Assessment. *Journal of Urban Planning and Development 115(2), American Society of Civil Engineers*, pp.51-63.

Srinivasan, S. and L. Ma (2009) Synthetic population generation: A heuristic data-fitting approach and validations, paper presented at the *12th International Conference on Travel Behaviour Research (IATBR)*, Jaipur.

Wheaton, W. D., J. C. Cajka, B. M. Chasteen, D. K. Wagener, P. C. Cooley, and L. Ganapathi (2009) Synthesized Population Databases: A US Geospatial Database for Agent-Based Models. *RTI Press Method* Report No. MR-0010-0905.

Wilson, A. G. (1969) The Use of Entropy Maximizing Models in the Theory of Trip Distribution, Mode Split and Route Split. *Journal of Transport Economics and Policy 3,* pp. 108-126.

Ye, X., K.C. Konduri, R.M. Pendyala, B. Sana, and P. Waddell (2009) A Methodology to Match Distributions of both Household and Person Attributes in the Generation of Synthetic Populations. Presented at the *88th Annual Meeting of Transportation Research Board,* Washington, D.C.