UPDATES TO THE BMC POPULATION SYNTHESIS MODEL: INCORPORATING CONTROLS AT MULTIPLE GEOGRAPHIC RESOLUTIONS

FINAL REPORT: JUNE 2015

Prepared By

Karthik C Konduri
Assistant Professor
University of Connecticut
Storrs, CT 06269-3037, USA.
Phone: 860-486-2733
Email: kkonduri@engr.uconn.eduri

Ram M. Pendyala
Frederick R. Dickerson Chair and Professor of Transportation
Georgia Institute of Technology
Atlanta, GA 30332-0355, USA
Phone: 404-385-3754
Fax: 404-894-2278

TABLE OF CONTENTS

CHAPTER 1: BACKGROUND

A key input to the implementation of an activity-based travel demand model is disaggregate data about socio-economic and demographic characteristics of the entire population in a region. Disaggregate socio-economic and demographic data at the individual household and person level for an entire regional population is not readily available. Analytic tools in the form of synthetic population generators (SPG) are used to synthesize a population based on readily available data in the form of aggregate marginal distributions of population characteristics, and disaggregate data about a sample of the population in the region. BMC currently employs a synthetic population generator called PopGen-BMC for use in their existing four-step travel demand modeling process and also plans to use it in the activity-based travel demand model development effort that is currently underway.

In PopGen-BMC, only a limited set of marginal control distributions available at the level of the Traffic Analysis Zone (TAZ) are utilized to control the generation of the synthetic population. The variables currently controlled include household-level variables of number of persons, income, and worker count, and a person-level variable of employment. However, BMC also maintains a number of marginal distributions for other variables at other levels of spatial resolution. For example, person-level marginal distributions of age, gender, and race, and household-level marginal distributions of age of householder are available at the county-level or planning district level. PopGen-BMC currently is not unable to fully utilize information contained in these additional marginal distributions available at a level of resolution different than the traffic analysis zone. This inability in PopGen-BMC may potentially lead to issues in the synthetic population generation including mismatch against known marginal distributions of population characteristics at other levels of spatial resolution and also potential inaccuracies in the representation of the underlying population due to controlling for only a limited set of TAZ-level data.

In consultation with the technical staff of the Transportation Division at Baltimore Metropolitan Council, the study team embarked on a study effort to enhance the existing synthetic population generator namely PopGen-BMC. The primary objective of the study was to extend the existing implementation of the PopGen-BMC to be able to accommodate marginal distributions at different levels of spatial resolution. This report details enhanced approaches for accommodating multilevel marginal distributions developed in this study effort. Further, the sensitivity and applicability of the approaches are demonstrated using data from the real world including data readily available from Census and data provided by BMC. The remaining report is organized as follows. In Chapter 2, the enhanced approaches are described along with an illustration of the approaches using a simple numerical example. In Chapter 3, a sensitivity analysis is presented using Census 2000 data. In Chapter 4, results from applying the enhanced procedures to generate a synthetic population for the 2012 model year are presented. In addition to developing the methodologies, the study team also implemented the procedures into a stand-alone software packaged dubbed PopGen 2.0. A brief overview of the software is presented in Chapter 5. Installation instructions for PopGen 2.0 are presented in Appendix A and instructions for using the software are described in Appendix B.

CHAPTER 2: ENHANCED POPULATION SYNTHESIS APPROACHES FOR INCORPORATING MULTILEVEL CONTROLS

In PopGen-BMC there are two key steps for synthetic population generation. First, household sample weights are generated that closely match the marginal distribution of household and person attributes of interest. In the second step, estimated sample weights are used to probabilistically draw households (and subsequently the associated person units) to create a synthetic population for the given geographic unit. In order to generate a synthetic population for the entire model region, the above two steps are carried out for each geographic unit (e.g. TAZ) independently. By independently generating a synthetic population for each geographic unit, PopGen-BMC is not capable of controlling for available marginal distributions of household- and person-level attributes at higher levels of spatial resolutions (e.g. county, state). In order to overcome this limitation of PopGen-BMC, the study team has developed two enhanced approaches to population synthesis that are capable of accommodating marginal control distributions at multiple spatial resolutions. In the next section, an overview of the enhanced approaches is presented. In the following section, the population synthesis approaches are illustrated using a simple example. In the last section, conclusions are presented.

Enhanced Population Synthesis Approaches

In the enhanced approaches to population synthesis, the two steps namely household sample weight generation and household drawing remain. However, the methodology for estimating the sample household weights for each individual geographic unit and the order in which the population synthesis steps are carried out is significantly different from PopGen-BMC. In PopGen-BMC, individual geographic units provide a clear separation for performing the two steps for synthesizing a population. As a result, sample household estimation and household drawing steps are performed independently and sequentially for each geographic unit. However, in order to control marginal distributions at higher spatial resolutions, individual geographic units cannot be used as a separator for performing the steps for generating a synthetic population. Therefore, in an effort to overcome this limitation, the highest level spatial resolution (e.g. county, or state) at which controls are provided is used to separate the steps in the enhanced approaches. This allows the enhanced approaches to accommodate marginal distributions of control variables at all spatial resolutions from the highest to the lowest.

The choice of spatial resolution for separating the steps of population synthesis also necessitates a change in the order of the steps for generating a synthetic population for any given geographic unit. The two steps are no longer applied sequentially for a geographic unit. For a given highest level spatial unit, household sample weights for all geographic units that belong to the highest level spatial unit are simultaneously estimated while controlling for marginal distributions both at the individual geographic unit level and also at higher-level spatial resolutions. After estimating the sample weights for all geographic units, the household drawing process is carried out to generate a synthetic population for each geographic unit. The household drawing step is the same even in the enhanced population synthesis therefore the discussion in the remaining sections of this chapter is focused on the approaches for estimating the sample household weights. Information regarding the drawing procedure can be obtained from Ye et al. (2009).

Two different approaches were developed for estimating the household sample weights such that available marginal distributions at multiple spatial resolutions can be controlled. These

approaches build on existing formulations for estimating the household sample weights for a given geography while accommodating household- and person-level constraints developed by the study team. The first approach builds on the Iterative Proportional Updating (IPU) algorithm proposed by Ye et al. (2009). IPU is a heuristic approach that proceeds by first initializing sample household weights and subsequently reweighting/adjusting sample household weights in an iterative manner until all the constraints (including the household- and person-level marginal distributions) are satisfied. The second approach extends the entropy-based optimization procedure for estimating weights proposed by Bar-Gera et al. (2009). It must be noted if a solution(s) for the sample household weights exists, then IPU will always lead to a feasible solution but the solution may not always be optimal – this is because IPU is a heuristic procedure. On the other hand, Entropy procedure will always lead to an optimal solution that maximizes the entropy. If a solution for the sample household weights doesn't exist then both approaches will settle on a corner solution. The biggest difference between earlier implementations of the IPU and Entropy approaches and those developed in this effort are that the new implementations can control for marginal distributions at multiple spatial resolutions.

In the rest of the report, it is assumed that marginal distributions at only two levels of spatial resolutions will be provided when generating a synthetic population – this assumption is consistent with the main use case of the methodology namely to control for marginal distributions at TAZ and county levels when simulating a synthetic population for the BMC model area. However, it must be noted that the number of spatial resolutions at which marginal distributions can be provided is not limited to two in the enhanced approaches. The enhanced approaches can easily be extended to accommodate marginal distributions for any number of spatial resolutions. From this point forward, highest level spatial resolution is referred to as "region" and lower level spatial resolution will be referred to as "geographic unit". The term controls, constraints, and marginal distributions are used interchangeably in the remaining text – it must be noted that all these terms refer to restrictions that must be satisfied by the estimated sample household weights and subsequently by the synthetic population that is generated. Also, in the remaining report, the enhanced IPU and Entropy approaches will be referred to as just IPU and Entropy for the sake of brevity.

Illustration of the Enhanced Approaches for Estimating Sample Weights

As noted in the previous section, for each region, sample weights for all geographic units in that region are estimated simultaneously such that marginal distributions at all spatial resolutions are satisfied. Both IPU and Entropy approaches employ iterative procedures for estimating the sample weights. The iterative procedures share common steps as shown in Table 1.

Table 1: Overview of the steps for the IPU and Entropy approaches

Steps	Description
1	Initialize sample weights for all geographic units belonging to the given highest level spatial unit
2	Reweight (adjust) the sample weights for all geographic units simultaneously so that the household- and person-level controls for a given region are matched
3	For each geographic unit, reweight (adjust) the sample weights to match the household- and person-level controls for the geographic unit

Both approaches start by assigning initial weights to each sample household for all the geographic units under consideration. In the second step, sample weights for all geographic units under consideration are adjusted such that the household- and person-level marginal distributions at the region level are matched. In the third step, sample weights are then readjusted for each geographic unit such that the household- and person-level marginal distributions at the level of resolution of the geographic unit are satisfied. The general steps between IPU and Entropy approaches are similar but there is a key difference in how the adjustment factors are calculated in Steps 2 and 3 between the two approaches.

If the marginal distributions result in a unique solution for the sample household weights then the order of Steps 2 and 3 can be interchanged and both approaches will result in the same set of final weights. However, in most real-world applications, marginal distributions do not always lead to a unique solution. In such cases, changing the order of Steps 2 and 3 may result in a different solution using the IPU and Entropy approaches. The solutions from the IPU and Entropy approaches while different still belong to the set of feasible and optimal solutions respectively that satisfy the given constraints. The IPU and Entropy approaches are illustrated in this section with the help of a simple numerical example.

Assume there is a model area with a single region. Further, assume there are exactly two geographic units (referenced as Geo 1 and Geo 2) within the region. Marginal distributions are available for each individual geographic unit for one household- and one person-level attribute. Additionally, marginal distribution for one household-level attribute is also available at the region level. More specifically, the household-level attribute is a two category marginal distribution of a household type and the person-level variable is a three category marginal distribution of person type. Also, household-level attribute at the region level is a three category marginal distribution of household type. Household type (person type) represents attributes about households (persons) and can be defined either based on a single household characteristic e.g. household income (e.g. age of person) or by combining multiple household characteristics e.g. household income by household size (e.g. age of person by gender). Typically, the household type attribute at the individual geographic unit and the household type attribute at the region level are different. If they are not different then the controls are essentially redundant and the additional regional controls do not add value for generating a more representative synthetic population for the model region. The sample dataset comprises of 8 households and 24 persons. The primary objective of the sample household weight estimation step using the IPU and Entropy approaches then is to estimate sample household weights such that the individual geography level and region level marginal distributions are satisfied.

As mentioned earlier, both IPU and Entropy procedures follow the same general steps shown in Table 1. The main difference between the IPU and Entropy is in the approach used to calculate the adjustment factors in Steps 2 and 3. In the following discussion, IPU approach is illustrated using a numerical example. Entropy procedure can be implemented by replacing the approach for estimating the adjustment described below in Steps 2 and 3 with the steps outlined in Bar-Gera et al. (2009).

Step 1: Initialize household sample weights

The sample weight estimation process begins by assigning an initial set of weights to all sample households for all geographic units. Weights are typically assigned to sample households and not sample persons. Unit weights are assigned to each sample household to start the weighting process. Table 2 shows the initial set of weights for Geo 1 and Geo 2. Table 1 also shows the given marginal distributions at both the region and geographic unit levels. Marginal distributions

are referenced as household-type (person-type) constraints in the table because the marginal distribution values also serve as controls when a single household-level (person-level) marginal distribution is provided. If more than one household-level (person-level) marginal distributions are provided, Iterative Proportional Fitting (IPF) procedure (see Beckman et al. 1996) is applied to estimate the household-type (person-type) constraints to be matched during the sample weight estimation process.

Sample data for each geographic unit is presented in the form a frequency matrix in Table 2. A row in the frequency matrix defines the household in terms of the household- and person-types at both the region level and geographic unit level. For a given row in the frequency matrix, each column represents the contribution of the sample household to the corresponding household- and/or person-types. Values in the frequency matrix for household types generally assume a value of zero or one because a household can only belong to one household type. On the other hand, the columns under person types can assume any value between zero and maximum number of people in the household because each household can have multiple people and depending on the number of people and their types a value of more than 1 can be assumed by the person type column. For example, for sample household with hid = 1, the value under region household type 2 column indicates that the household belongs to that type. Also the household belongs to household type 1 at the geographic unit level. The sample household also has three individuals each belonging to each of the person type categories at the geographic unit level.

Table 2 contains additional information namely "weighted sum", and a deviation measure "δ" that shows how well the weights satisfy the constraints. The "weighted sum" row provides the sum of the values in a given household-type (person-type) column weighted by the "weight" column. δ value provides an estimate of the difference between the weighted sum and the corresponding constraint value and is used to measure how well the weights satisfy the different constraints. It is calculated by first taking the absolute difference between the weighted sum and constraint, and then dividing that value by the constraint. The value of this measure is always greater than or equal to zero. A δ value of zero indicates a perfect match with the corresponding constraint and a value greater than zero indicates a deviation from the constraint with the deviation increasing with increasing value of δ. As expected, the weighted sums resulting from the initial set of weights do not match the given constraints well because the initial weight values are chosen arbitrarily. In Steps 2, and 3 these initial weights are adjusted iteratively until the given constraints are satisfied.

Step 2: Adjust the household sample weights to match region level constraints

In this step, sample household weights for all geographic units are adjusted such that the marginal distributions at the region level are satisfied. For a given unit at the region level, the process proceeds by adjusting the sample household weights for all geographic units that belong to the region unit such that given constraints for the region unit are satisfied. In the numerical example, this process begins by first adjusting the sample household weights for the two geographic units to match the region household type 1 column. The adjustment factor is calculated by dividing the constraint by the weighted sum for the corresponding column i.e. 86/4 = 21.50 (as shown in Table 2). The weights of all sample households for both geographic units that contribute to the region household type 1 column are scaled by a factor of 21.25. The adjusted sample household weights along with updated values for the weighted sum and δ value

are presented in Table 3. It can be seen that weighted sum value for the region household type 1 column now perfectly matches the constraint for that column.

After adjusting the sample household weights with respect to a constraint, updated sample weights are carried forward and adjusted with respect to the next constraint. This process continues until the sample household weights are adjusted with respect to all household- and person-type constraints at the region level. In the numerical example, the sample household weights in Table 3 are adjusted next with respect to region household type 2 by applying adjustment factor of 10.17 (constraint/weighted sum for region household type 2 i.e. 61/6). The sample household weights are then adjusted with respect to region household type 3 column by a factor of 13.67 (i.e. 82/6). The sample household weights after adjustment with respect to region household type 2 and region household type 3 constraints are shown in Tables 4 and 5 respectively.

It can be seen in Table 5 that all sample household weights are now different from the initial values assumed in Step 1. Also, it can be seen that that the sample household weights now result in weighted sums that match the household type constraints at the region level perfectly. While interesting, it must be noted that the perfect match observation is possible only when the weights are last adjusted with respect to household-type constraints. If weights are adjusted with person-type constraints then it is unlikely that this observation will hold. Now observing the weighted sum values for the household- and person-type constraints at the individual geography level, it can be seen that the sample household weights do not satisfy those constraints well. The weights are next adjusted to satisfy the household- and person-type constraints for the individual geographic units in Step 3.

Step 3: Adjust the household sample weights to match constraints available for each geographic unit

In Step 3, sample household weights are adjusted such that the household- and person-type constraints for each geographic unit are satisfied. Unlike the previous step where sample household weights for all geographic units for a given region are adjusted together for each constraint, the sample household weights in this step are adjusted one geographic unit at a time to satisfy the constraint for the corresponding geographic unit. Going back to the numerical example, the sample household weights at the end of Step 2 for Geographic unit 1 (Geo 1) are first adjusted with respect household type 1 constraint. The adjustment is calculated by taking the ratio of the constraint and the weighted sum i.e. 46/45.33 = 1.01. Weights of all sample households for Geo 1 that contribute to household type 1 are then adjusted by this factor resulting in the sample household weights for Geo 1 as shown in Table 6. The weighted sum for household type 1 now satisfies the constraint perfectly for Geo 1. As expected, this adjustment with respect to household type 1 constraint for Geo 1 does not update the weighted sum values for the second geographic unit (Geo 2) since only sample household weights for Geo 1 were adjusted. Further, the weighted sum values for the region household type constraints have changed because of the changes in the sample household weights for the first geographic unit.

The sample household weights for Geo 1 are then adjusted with respect to household type 2 constraint and subsequently with respect to the three person type constraints. The adjustment process is then repeated similarly to update the sample household weights for Geo 2 so that the corresponding household type constraints and person type constraints are matched. The sample household weights for the two geographic units after one pass of the adjustments with respect to household type and person type constraints are shown in Table 7.

One round of the adjustments described in Steps 2 and 3 comprises an iteration of the IPU procedure. After the first iteration, match between the sample household weights has improved as is evidenced by the close match in values of the weighted sum when compared to the constraints as shown in Table 7. However, there are still some differences between the weighted sums and constraints so the adjustment process described above in Steps 2 and 3 is repeated. The weights are iteratively adjusted until there is no further improvement in the match with respect to the different constraints. Figure 1 shows the average δ value across constraints at the geographic unit level for Geo 1 and Geo 2. It can be seen that as iterations progress, the average δ value is approaching zero indicating that the sample weights for the two geographic units almost perfectly match the geographic unit level constraints.

As indicated above the IPU and entropy approaches are iterative procedures. Therefore, a heuristic for stopping the iterative process must be defined. The process is typically stopped when the improvement in the average δ value drops below a predefined threshold – this heuristic is reasonable because it ensures that the algorithm automatically stops when it cannot find a better solution. The resulting solution can be a feasible solution wherein all constraints are almost perfectly matched or a corner solution wherein a subset of the constraints are almost perfectly matched whereas the remaining constraints are only closely matched. Whenever constraints are consistent, both algorithms always result in a feasible solution. Further, Entropy approach will always result in an optimal solution that maximizes the entropy. On the other hand, if the constraints are inconsistent then both the algorithms will settle on a corner solution.

Figure 2 shows the improvement in average δ value across iterations for Geo 1 and Geo 2. Assuming 1e-6 as the value of the threshold for improvement in average δ, the iterative process reaches a solution at around iteration 200. At iteration 200, the iterative process for estimating weights is said to have achieved convergence. At this iteration, it can be seen from Figure 1 that the value of average δ for both geographic units is close to zero indicating that a feasible solution and not a corner solution was obtained. In Figures 1 and 2, the measures are shown for iterations beyond 200 to illustrate that there is no further improvement in the estimated sample weights with additional iterations. Performance of the sample weights in matching the given constraints at the region level and for the individual geographic units is discussed below.

Table 8 shows the sample household weights for the two geographic units at the end of 1000 iterations. It can be seen that the sample household weights in Table 8 almost perfectly satisfy the different constraints both at the region level and at the individual geographic unit level. Table 9 shows the sample household weights for the two geographic units at the end of 1000 iterations if no region-level household type constraints were controlled. Results presented in Table 9 emulate the sample weight estimation algorithm implemented in the existing PopGen-BMC. It can be seen that when control variables are not provided at the region level, the match between the estimated sample household weights and the corresponding region level constraints is poor. While the average δ value ranges from 0 to 0.011 for the three regional household type constraints in Table 8, the δ values for the same three household type constraints in Table 9 range from 0.017 to 0.161 – a significant difference. While this is consistent with expectations that controls that are not considered cannot be accommodated, this demonstrates the value of additional control variables. Failure to account for additional controls even if they may only be at a higher level of spatial resolution can lead to a synthetic population that is not representative of the underlying population in the model region.

As mentioned above, as long as the constraints are consistent, the iterative approach will always lead to a feasible solution for sample household weights that will satisfy all the constraints. However, when working with real world datasets, often constraints suffer from inconsistencies and as a result, a feasible solution for the sample household weights does not exist. The iterative procedure then estimates sample household weights that satisfy a subset of constraints and not all constraints. The resulting solution for the sample household weights is referred to as a corner solution. There are many potential corner solutions for a given set of inconsistent constraints. Each solution corresponds to a solution for sample household weights where a subset of constraints are perfectly satisfied while remaining constraints are only closely satisfied. From the different corner solutions, the enhanced population synthesis implemented in this study currently selects the one where household constraints at the geographic unit level are perfectly satisfied. In this corner solution, person level constraints at the geographic unit level and the region-level constraints are also satisfied but with small deviations. The choice of this particular corner solution ensures that the sample household weights are consistent with the unit of sampling namely households in the drawing step of the population synthesis process.

So far the discussion was focused on the IPU procedure both for illustrating the iterative algorithm and for describing the performance of the estimated sample household weights when multilevel controls are provided. In the remaining portion of the subsection, results from the Entropy procedure are presented and similarities and differences between the IPU and Entropy approaches are discussed. The biggest difference between the IPU and Entropy procedure is in the calculation of the adjustment factor during the iterative process for estimating sample household weights. In the Entropy procedure, a polynomial equation is solved to estimate the adjustment factor for a given constraint. More details regarding the adjustment factor calculation in the Entropy procedure can be found in Bar-Gera et al. (2009). Table 10 shows the sample household weights for the two geographic units at the end of 1000 iterations using the Entropy procedure. The estimated sample household weights perfectly match the different constraints at both the region level and geographic unit level. It is interesting to note that in the IPU procedure differences even if only very small were observed for some constraints whereas in the Entropy procedure all the constraints are matched perfectly. However, this minor difference in performance of the sample household weights doesn't appear to affect the fit of the synthetic population that is generated (as discussed below). This observation alludes to the algorithmic superiority of the Entropy procedure in finding a feasible solution. It is also interesting to note that IPU and Entropy procedures result in slightly different solutions for the sample household weights both of which satisfy the given constraints at the region level and geographic unit level. This lends support to the notion that the problem of estimating sample household weight estimation given constraints has many feasible solutions and doesn't have a single unique solution. Each of the two approaches appears to be selecting one of the many feasible solutions that satisfy the given constraints.

Tables 11 and 12 present performance summary of the synthetic population generation using the IPU and entropy procedures respectively. Performance summaries for both the sample household weights estimation and household drawing steps of the synthetic population generation process are presented. As noted above, Entropy process is resulting in sample household weights that perfectly match the given constraints whereas in the IPU procedure the same household weights are near perfect with small deviations. The small differences however do not seem to affect the household drawing step and appear to result in synthetic populations with comparable performance. In both cases, the total number of synthetic households is exactly

matched. However, when the distributions of household- and person-level variables are compared between the given values and the synthetic population, small differences can be observed. Differences between the synthetic population and given constraints in the IPU procedure range from -3 to +2 for different region level and geographic unit level constraints. On the other hand for the synthetic population generated using the Entropy procedure the range of differences is from -4 to +3.

Conclusions

In this chapter, two approaches for estimating sample household weights namely IPU and Entropy procedures are described. These approaches are able to accommodate marginal distributions at multiple spatial resolutions. The approaches were illustrated using a simple numerical example in this chapter. Results point to the plausibility of the approaches. Sensitivity of the enhanced methodologies is presented in Chapter 3 and applicability of the enhanced approaches for population synthesis is demonstrated in Chapter 4 respectively.

Table 2: Table showing the frequency matrix and initial sample household weights

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	1.00	0	0	1	1	0	1	1	1
	2	1.00	1	0	0	1	0	1	0	1
	3	1.00	0	1	0	1	0	2	1	0
	4	1.00	1	0	0	0	1	1	0	2
	5	1.00	0	1	0	0	1	0	2	1
	6	1.00	0	0	1	0	1	1	1	0
	7	1.00	0	1	0	0	1	2	1	2
	8	1.00	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	**Weighted Sum**		3.000	5.000	9.000	8.000	7.000
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.935	0.902	0.902	0.909	0.917
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	1.00	0	0	1	1	0	1	1	1
	2	1.00	1	0	0	1	0	1	0	1
	3	1.00	0	1	0	1	0	2	1	0
	4	1.00	1	0	0	0	1	1	0	2
	5	1.00	0	1	0	0	1	0	2	1
	6	1.00	0	0	1	0	1	1	1	0
	7	1.00	0	1	0	0	1	2	1	2
	8	1.00	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		3.000	5.000	9.000	8.000	7.000
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.909	0.949	0.935	0.934	0.933
Match in constraints for Region		**Weighted Sum**	4.000	6.000	6.000					
		Constraint	86.000	61.000	82.000					
		δ	0.953	0.902	0.927					

Table 3: Table showing the frequency matrix and sample household weights after adjustment with respect to the region household type 1 constraint using IPU Procedure

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	1.00	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	1.00	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	1.00	0	1	0	0	1	0	2	1
	6	1.00	0	0	1	0	1	1	1	0
	7	1.00	0	1	0	0	1	2	1	2
	8	1.00	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	**Weighted Sum**		23.500	25.500	50.000	8.000	68.500
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.489	0.500	0.457	0.909	0.185
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	1.00	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	1.00	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	1.00	0	1	0	0	1	0	2	1
	6	1.00	0	0	1	0	1	1	1	0
	7	1.00	0	1	0	0	1	2	1	2
	8	1.00	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		23.500	25.500	50.000	8.000	68.500
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.288	0.742	0.638	0.934	0.341
Match in constraints for Region		**Weighted Sum**	86.000	6.000	6.000					
		Constraint	86.000	61.000	82.000					
		δ	0.000	0.902	0.927					

Table 4: Table showing the frequency matrix and sample household weights after adjustment with respect to the region household type 2 constraint using IPU Procedure

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	1.00	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	10.17	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	10.17	0	1	0	0	1	0	2	1
	6	1.00	0	0	1	0	1	1	1	0
	7	10.17	0	1	0	0	1	2	1	2
	8	1.00	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	**Weighted Sum**		32.667	43.833	86.667	44.667	96.000
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.290	0.141	0.058	0.492	0.143
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	1.00	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	10.17	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	10.17	0	1	0	0	1	0	2	1
	6	1.00	0	0	1	0	1	1	1	0
	7	10.17	0	1	0	0	1	2	1	2
	8	1.00	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		32.667	43.833	86.667	44.667	96.000
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.010	0.557	0.372	0.634	0.077
Match in constraints for Region		**Weighted Sum**	86.000	61.000	6.000					
		Constraint	86.000	61.000	82.000					
		δ	0.000	0.000	0.927					

Table 5: Table showing the frequency matrix and sample household weights after adjustment with respect to the region household type 3 constraint using IPU Procedure

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	13.67	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	10.17	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	10.17	0	1	0	0	1	0	2	1
	6	13.67	0	0	1	0	1	1	1	0
	7	10.17	0	1	0	0	1	2	1	2
	8	13.67	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	**Weighted Sum**		45.333	69.167	124.667	95.333	108.667
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.014	0.356	0.355	0.083	0.294
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	13.67	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	10.17	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	10.17	0	1	0	0	1	0	2	1
	6	13.67	0	0	1	0	1	1	1	0
	7	10.17	0	1	0	0	1	2	1	2
	8	13.67	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		45.333	69.167	124.667	95.333	108.667
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.374	0.301	0.097	0.219	0.045
Match in constraints for Region		**Weighted Sum**	86.000	61.000	82.000					
		Constraint	86.000	61.000	82.000					
		δ	0.000	0.000	0.000					

Table 6: Table showing the frequency matrix and sample household weights after adjustment with respect to all the region household type constraints and household type 1 constraint for Geo 1 using IPU Procedure

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	13.87	0	0	1	1	0	1	1	1
	2	21.82	1	0	0	1	0	1	0	1
	3	10.32	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	10.17	0	1	0	0	1	0	2	1
	6	13.67	0	0	1	0	1	1	1	0
	7	10.17	0	1	0	0	1	2	1	2
	8	13.67	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	**Weighted Sum**		46.000	69.167	125.483	95.684	109.184
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.000	0.356	0.364	0.087	0.300
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	13.67	0	0	1	1	0	1	1	1
	2	21.50	1	0	0	1	0	1	0	1
	3	10.17	0	1	0	1	0	2	1	0
	4	21.50	1	0	0	0	1	1	0	2
	5	10.17	0	1	0	0	1	0	2	1
	6	13.67	0	0	1	0	1	1	1	0
	7	10.17	0	1	0	0	1	2	1	2
	8	13.67	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		45.333	69.167	124.667	95.333	108.667
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.374	0.301	0.097	0.219	0.045
Match in constraints for Region		**Weighted Sum**	86.316	61.150	82.201					
		Constraint	86.000	61.000	82.000					
		δ	0.004	0.002	0.002					

Table 7: Table showing frequency matrix and sample household weights after one iteration using IPU Procedure (i.e. after one round of adjustment with respect to all region household type constraints and with respect to household type and person type constraints for two geographic units)

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	14.74	0	0	1	1	0	1	1	1
	2	17.94	1	0	0	1	0	1	0	1
	3	11.43	0	1	0	1	0	2	1	0
	4	13.04	1	0	0	0	1	1	0	2
	5	9.30	0	1	0	0	1	0	2	1
	6	11.17	0	0	1	0	1	1	1	0
	7	7.97	0	1	0	0	1	2	1	2
	8	11.17	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	**Weighted Sum**		44.120	52.643	106.869	86.249	84.000
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.041	0.032	0.162	0.020	0.000
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	8.03	0	0	1	1	0	1	1	1
	2	12.29	1	0	0	1	0	1	0	1
	3	7.53	0	1	0	1	0	2	1	0
	4	24.17	1	0	0	0	1	1	0	2
	5	11.86	0	1	0	0	1	0	2	1
	6	19.89	0	0	1	0	1	1	1	0
	7	11.74	0	1	0	0	1	2	1	2
	8	19.89	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		27.844	87.550	122.800	110.679	104.000
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.156	0.116	0.110	0.093	0.000
Match in constraints for Region	**Weighted Sum**		67.444	59.825	84.888					
	Constraint		86.000	61.000	82.000					
	δ		0.216	0.019	0.035					

Figure 1: Figure showing the fit values across iterations in the IPU procedure for Geo 1 and Geo 2

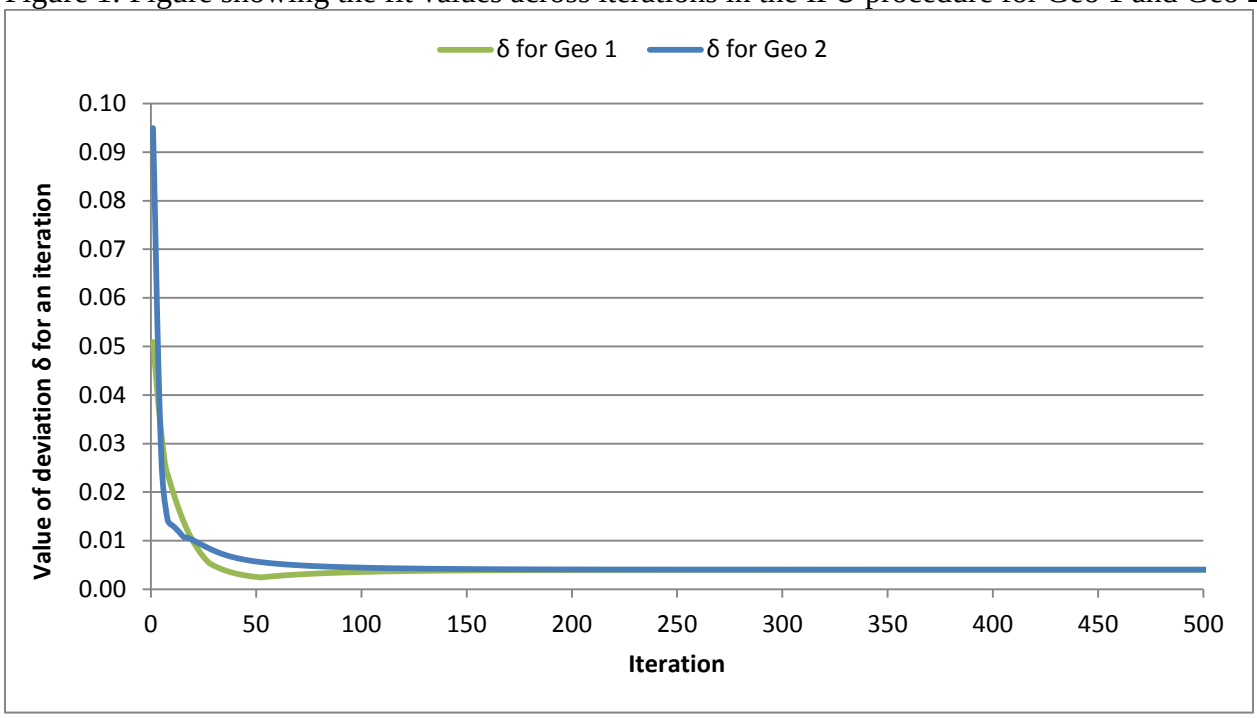

Figure 2: Figure showing the change in fit values across iterations in the IPU procedure for Geo 1 and Geo 2

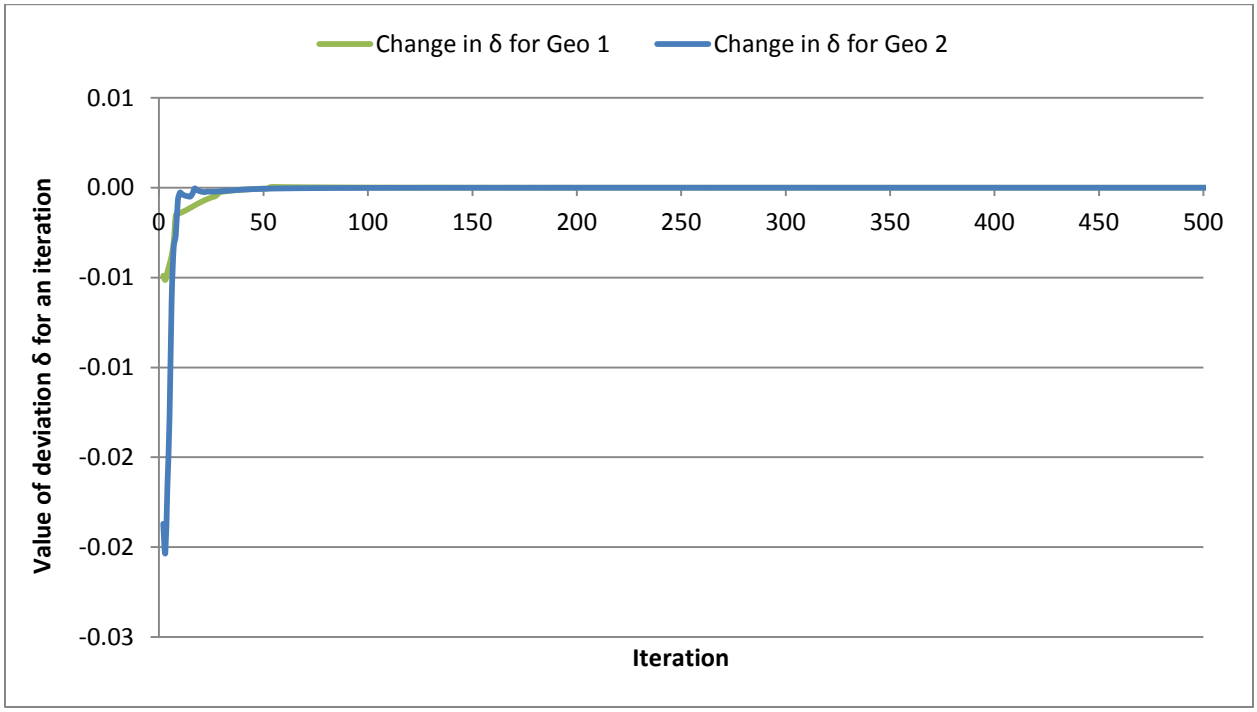

Table 8: Table showing the frequency matrix and sample household weights after one thousand iterations using IPU Procedure

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	8.33	0	0	1	1	0	1	1	1
	2	25.71	1	0	0	1	0	1	0	1
	3	12.19	0	1	0	1	0	2	1	0
	4	12.19	1	0	0	0	1	1	0	2
	5	20.02	0	1	0	0	1	0	2	1
	6	8.22	0	0	1	0	1	1	1	0
	7	2.78	0	1	0	0	1	2	1	2
	8	8.22	0	0	1	0	1	1	2	0
		Match in constraints for Geo 1		**Weighted Sum**		46.227	51.434	92.604	88.000	84.000
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.005	0.009	0.007	0.000	0.000
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	4.46	0	0	1	1	0	1	1	1
	2	17.71	1	0	0	1	0	1	0	1
	3	11.00	0	1	0	1	0	2	1	0
	4	30.39	1	0	0	0	1	1	0	2
	5	10.31	0	1	0	0	1	0	2	1
	6	26.85	0	0	1	0	1	1	1	0
	7	5.38	0	1	0	0	1	2	1	2
	8	26.85	0	0	1	0	1	1	2	0
		Match in constraints for Geo 2		**Weighted Sum**		33.171	99.767	139.004	122.000	104.000
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.005	0.008	0.007	0.000	0.000
Match in constraints for Region		**Weighted Sum**	86.000	61.682	82.916					
		Constraint	86.000	61.000	82.000					
		δ	0.000	0.011	0.011					

Table 9: Table showing the frequency matrix and sample household weights after one thousand iterations using IPU Procedure when no region-level household type constraints are controlled

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	15.81	0	0	1	1	0	1	1	1
	2	22.23	1	0	0	1	0	1	0	1
	3	7.96	0	1	0	1	0	2	1	0
	4	11.39	1	0	0	0	1	1	0	2
	5	16.37	0	1	0	0	1	0	2	1
	6	11.59	0	0	1	0	1	1	1	0
	7	3.41	0	1	0	0	1	2	1	2
	8	8.24	0	0	1	0	1	1	2	0
			Match in constraints for Geo 1	Weighted Sum		46.000	51.000	92.000	88.000	84.000
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.000	0.000	0.000	0.000	0.000
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	7.59	0	0	1	1	0	1	1	1
	2	15.52	1	0	0	1	0	1	0	1
	3	9.89	0	1	0	1	0	2	1	0
	4	24.66	1	0	0	0	1	1	0	2
	5	13.11	0	1	0	0	1	0	2	1
	6	34.93	0	0	1	0	1	1	1	0
	7	9.23	0	1	0	0	1	2	1	2
	8	17.08	0	0	1	0	1	1	2	0
			Match in constraints for Geo 2	**Weighted Sum**		33.000	99.000	138.000	122.000	104.000
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.000	0.000	0.000	0.000	0.000
Match in constraints for Region	**Weighted Sum**		73.801	59.963	95.236					
	Constraint		86.000	61.000	82.000					
	Deviation		0.142	0.017	0.161					

Table 10: Table showing the frequency matrix and sample household weights after one thousand iterations using Entropy Procedure

			Region Household Type			Household Type		Person Type		
For first geographic unit (Geo 1)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	8.88	0	0	1	1	0	1	1	1
	2	27.27	1	0	0	1	0	1	0	1
	3	9.84	0	1	0	1	0	2	1	0
	4	11.61	1	0	0	0	1	1	0	2
	5	18.10	0	1	0	0	1	0	2	1
	6	6.25	0	0	1	0	1	1	1	0
	7	3.26	0	1	0	0	1	2	1	2
	8	11.78	0	0	1	0	1	1	2	0
		Match in constraints for Geo 1		**Weighted Sum**		46.000	51.000	92.000	88.000	84.000
				Constraint		46.000	51.000	92.000	88.000	84.000
				δ		0.000	0.000	0.000	0.000	0.000
			Region Household Type			**Household Type**		**Person Type**		
For second geographic unit (Geo 2)	**hid**	**weight**	**1**	**2**	**3**	**1**	**2**	**1**	**2**	**3**
	1	3.07	0	0	1	1	0	1	1	1
	2	18.88	1	0	0	1	0	1	0	1
	3	11.06	0	1	0	1	0	2	1	0
	4	28.24	1	0	0	0	1	1	0	2
	5	11.90	0	1	0	0	1	0	2	1
	6	26.81	0	0	1	0	1	1	1	0
	7	6.84	0	1	0	0	1	2	1	2
	8	25.22	0	0	1	0	1	1	2	0
		Match in constraints for Geo 2		**Weighted Sum**		33.000	99.000	138.000	122.000	104.000
				Constraint		33.000	99.000	138.000	122.000	104.000
				δ		0.000	0.000	0.000	0.000	0.000
Match in constraints for Region		**Weighted Sum**	86.000	61.000	82.000					
		Constraint	86.000	61.000	82.000					
		Deviation	0.000	0.000	0.000					

Table 11: Summary of the fit for the estimated weights and the synthetic population against the given constraints using IPU procedure

Constraint Name		Category	Constraint	Performance for Step 1: Household sample weight generation			Performance for Step 2: Household drawing		
				Weighted Sum	Difference	δ	Synthesized	Difference	δ
For the first geographic unit (Geo 1)	Household Type	1	46	46.2	0.2	0.0049	46	0	0.0000
		2	51	51.4	0.4	0.0085	51	0	0.0000
	Person Type	1	92	92.6	0.6	0.0066	91	-1	-0.0109
		2	88	88.0	0.0	0.0000	85	-3	-0.0341
		3	84	84.0	0.0	0.0000	84	0	0.0000
For the first geographic unit (Geo 2)	Household Type	1	33	33.2	0.2	0.0052	33	0	0.0000
		2	99	99.8	0.8	0.0077	99	0	0.0000
	Person Type	1	138	139.0	1.0	0.0073	137	-1	-0.0072
		2	122	122.0	0.0	0.0000	124	2	0.0164
		3	104	104.0	0.0	0.0000	102	-2	-0.0192
For Region	Region Household Type	1	86	86.0	0.0	0.0000	87	1	0.0116
		2	61	61.7	0.7	0.0112	59	-2	-0.0328
		3	82	82.9	0.9	0.0112	83	1	0.0122

Table 12: Summary of the fit for the estimated weights and the synthetic population against the given constraints using Entropy procedure

Constraint Name		Category	Constraint	Performance for Step 1: Household sample weight generation			Performance for Step 2: Household drawing		
				Weighted Sum	Difference	δ	Synthesized	Difference	δ
For the first geographic unit (Geo 1)	Household Type	1	46	46.0	0.0	0.0000	46	0	0.0000
		2	51	51.0	0.0	0.0000	51	0	0.0000
	Person Type	1	92	92.0	0.0	0.0000	91	-1	-0.0109
		2	88	88.0	0.0	0.0000	88	0	0.0000
		3	84	84.0	0.0	0.0000	86	2	0.0238
For the first geographic unit (Geo 2)	Household Type	1	33	33.0	0.0	0.0000	33	0	0.0000
		2	99	99.0	0.0	0.0000	99	0	0.0000
	Person Type	1	138	138.0	0.0	0.0000	141	3	0.0217
		2	122	122.0	0.0	0.0000	121	-1	-0.0082
		3	104	104.0	0.0	0.0000	100	-4	-0.0385
For Region	Region Household Type	1	86	86.0	0.0	0.0000	85	-1	-0.0116
		2	61	61.0	0.0	0.0000	63	2	0.0328
		3	82	82.0	0.0	0.0000	81	-1	-0.0122

CHAPTER 3: SENSITIVITY ANALYSIS OF THE ENHANCED APPROACHES TO POPULATION SYNTHESIS

In this chapter, a sensitivity of the IPU and entropy approaches for estimating sample household weights is presented. The primary objectives of the sensitivity analysis are twofold namely 1) to evaluate the ability of the two approaches for accommodating multilevel controls using real world data and 2) to assess the value of the additional controls at higher levels of spatial resolution. To this end, a county-level population synthesis was carried out using data from Census 2000 for the state of Maryland. Similar to the numerical example in the previous chapter, the focus again was on the sample household weight estimation step of synthetic population generation – drawing households step of population synthesis was not carried out in this analysis. Marginal distributions at both county level and state level served as the multilevel controls in the analysis. The remaining chapter is organized as follows. The input data and the different scenario runs are described in the next section. In the following section, results are presented starting with the convergence properties of the IPU and entropy procedures for estimating county-level sample household weights. In the next section, performance results of the estimated sample household weights are presented. Some concluding thoughts are presented in the last section.

Description of Sample Weight Estimation Scenarios

Data from Census 2000 for the state of Maryland was used in the analysis. Public Use Microdata Sample (PUMS) was used to generate the sample files and Census Summary Files were used to prepare the marginal distribution files. County and state level marginal distributions for both household- and person-level attributes of interest served as the multilevel controls. Sample household weights were then generated for each of the 24 counties while accounting for both county- and state-level marginal distributions. The choice of county as the geographic unit (lower level spatial resolution) was made to keep the runtimes manageable because there are only 24 counties in the state of Maryland. Three different scenarios as described below were completed to evaluate the sensitivity of the IPU and entropy approaches and also to demonstrate the value of the additional controls at higher level of spatial resolution during the population synthesis process.

- Estimating Sample Household Weights without Multilevel Controls using IPU (Scenario 1): In this synthesis run, three household attributes – household type (hhldtype), household size (hhldsize), household income (hhldinc), and two person attributes – person age (page), and person gender (pgender) were controlled at the individual county level. It must be noted that no variables were controlled at the state level. Additionally the IPU algorithm was employed to estimate the weights for the sample households. This setup reflects the existing population synthesis process and serves as the baseline for comparing the results from the enhanced IPU and Entropy approaches wherein marginal distributions at multiple spatial resolutions are accommodated. This run will be referred to as Scenario 1 from this point forward.
- Estimating Sample Household Weights with Multilevel Controls using IPU (Scenario 2): In this synthesis run, the same household and person attributes were controlled as in Scenario 1 at the county level. Additionally, presence of children in the household – a household attribute and race – a person attribute were controlled at the state level. The enhanced IPU approach was used in this scenario to estimate the sample household weights. This setup represents the first step of a population synthesis process with multilevel controls. This will be referred to as Scenario 2 in the remaining discussion.

- Estimating Sample Household Weights with Multilevel Controls using Entropy (Scenario 3): The setup in this scenario is same as Scenario 2 except for the fact that the enhanced Entropy procedure was used to estimate the sample household weights.

Results: Monitoring Convergence

As noted earlier, the two approaches employ iterative procedures for estimating the sample household weights that match the given controls. Therefore, it is important to evaluate the convergence properties of the approaches to ensure that the best solution for the sample household weights is obtained. Two measures namely 1) average adjustment value across all constraints and 2) entropy were monitored across iterations to assess the convergence properties of the IPU and Entropy approaches for estimating sample weights. The convergence measures were monitored for four randomly selected counties and for the entire state. In the analysis, the number of iterations was limited to fifty in the interest of run times (as will be noted below, the number of iterations were enough to ensure convergence). Figure 3 shows the convergence properties for both the average adjustment and entropy value for Scenario 1. Figures 4 and 5 illustrate the convergence properties for Scenario 2 and 3 respectively. It can be seen from the three figures that after significant improvements in the values of the convergence measures in the first few iterations, the values seem to plateau as the iteration count increases. This shows that the two approaches are generally progressing towards convergence for all the scenarios. Additionally, it is also interesting to note that even though only 50 iterations were performed (for the sake of computational convenience), the iterative process was able to attain convergence and a valid solution for the sample household weights was obtained.

Results: Performance of the Sample Household Weights

As noted earlier, the primary objective of the study effort was to develop a synthetic population generation procedure that is capable of satisfying marginal distributions at different spatial resolutions. An important first step in the synthetic population generation is the estimation of sample household weights. The weights are subsequently used to draw the households that form the synthetic population. It is therefore important that the estimated weights satisfy the different marginal distributions provided as controls. The focus of the chapter is to evaluate performance of the IPU and Entropy procedures for estimating household weights in matching multilevel controls. Further, sample household weights are compared between scenarios where multilevel controls are present versus scenarios where no multilevel controls are provided. Aggregate and disaggregate analysis of the sample household weights estimated in the three scenarios are presented below.

Aggregate Comparison of the Sample Household Weights

Table 12 presents results from comparison of the scenarios in matching the given household and person totals. As expected, the household totals are perfectly matched at the regional level by the sample household weights estimated in all three scenarios. This is reasonable and consistent with expectations. In the first scenario, even though no regional controls are provided, the regional control totals are implied/controlled by the household-level variables specified for the individual county-level geographic units. Additionally, in Scenario 2 and Scenario 3, the sample household weights match the given household totals because regional marginal distributions are explicitly provided as controls. The estimated sample household weights also result in a close match of the person totals. The observation of a close match in person totals and not a perfect match is reasonable because marginal distributions often suffer from inconsistencies as a result feasible

solution sets (that satisfy all constraints) do not exist and only corner solutions (wherein a subset of the constraints are satisfied) are likely. In both IPU and Entropy approaches for estimating household sample weights, the corner solution where the household-level constraints are perfectly matched is chosen. In this corner solution, a better match in household-level marginal distributions comes at the expense of a deviation in the person-level marginal distributions. Typically, the corner solution where the household-level marginal distributions are perfectly matched is chosen because households are the basic sampling units in the drawing step of population synthesis. It can be seen that despite the choice of the corner solution (favoring match in household-level marginal distribution), the IPU and Entropy procedures appear to result in a good match of the given person totals. It is interesting to note that the match in person totals between Scenario 1 and Scenario 2 is negligible. This is reasonable because the total count of persons at the region level is implied/controlled by the total count of persons at the individual geographies. The match in person totals is much better in Scenario 3 where Entropy procedure is employed possibly alluding to the superiority of the Entropy procedure in matching controls compared to the IPU procedure.

In an effort to further analyze the sample household weights, the household- and person-level totals for four counties (with ids 1, 3, 5 and 9) were compared for the three scenarios. The match in person totals is similar at the individual geographies with entropy procedure performing the best (Scenario 3) followed by Scenarios 1 and 2 with very comparable performance. It is interesting to note that while the household totals are matched perfectly in Scenario1, there is a small deviation in the household totals for Scenarios 2 and 3. This is reasonable because in Scenarios 2 and 3, the iterative procedures for estimating the household weights are trading off matching household totals at the county level for matching household totals at the state level. It is also interesting to note that the deviation in household totals is higher for Scenario 3 employing the Entropy procedure with multilevel controls and lower for Scenario 2 employing the IPU procedure with multilevel controls. This observation combined with earlier observation of a better performance in matching the person totals more closely for Scenario 3 indicates that while there is a better fit in person totals for Scenario 3, this comes at the cost of additional deviation in the household totals.

Comparison of the Sample Household Weights in Matching Given Marginal Distributions

Table 13 present results from the comparison of estimated sample household weights in matching the marginal distributions of various controlled and uncontrolled household-level variables of interest. Unlike Table 12, the results in the table compare the estimated weights against given control totals at a more disaggregate level namely the marginal distributions of various household-level variables. As noted earlier, in each of the scenarios, three variables were controlled at the individual county level including household type, household size, and household income. Additionally, one variable was controlled at the state level (representing the multilevel control) namely presence of children in the household in Scenarios 2 and 3. Results from the comparison of marginal distribution of these four variables and one additional uncontrolled variable are presented to help evaluate the sample household weighting procedures and the scenario setups. Given column in the table represents the marginal distribution values that are provided as inputs and weighted sum represents the marginal distribution values implied by the estimated sample household weights. As expected the marginal distributions of controlled variables are almost perfectly matched with very small deviations observed in each of the three scenarios. Also, it is interesting to note that the marginal distribution of one of the controlled variables is perfectly matched in all three scenarios. In Scenario 1, the household income

variable is perfectly matched and in Scenarios 2 and 3, the presence of children in the household variable is perfectly matched. This is reasonable because these control variables represent the last matched variable in the Iterative Proportional Fitting (IPF) procedure (Beckman et al. 1996) that is used to estimate the household-type constraints before sample household weight estimation is carried out. As a result, the estimated sample weights perfectly match the marginal distributions for these variables. If order of control variables is modified, then the marginal distribution for the variable that is last controlled will be perfectly matched.

For household-level variables that were controlled in all three scenarios, it can be seen that the estimated weights are able to match the corresponding given values very closely. Between the different scenarios, the performance of Scenarios 1 and 2 is very comparable in matching the marginal distributions of controlled variables. This similarity in the performance may be attributable to the use of IPU procedure in the two scenarios. Scenario 3 performs better than the other scenarios in matching the household type variable but performs poorly compared to the other scenarios in matching the household size and household income variables. As expected, there are deviations between the estimated weights and the marginal distribution for the presence of children in the household variable in Scenario 1. This is reasonable because in Scenario1, this variable was not controlled. On the other hand in Scenarios 2 and 3, the marginal distribution for the presence of children is matched perfectly because 1) the variable was controlled at the state level and 2) it was also the last variable among the household-level variables that were controlled. It is interesting to note that the match in the marginal distribution for the uncontrolled variable namely householder age is best for Scenario 3 followed by Scenario 1 and lastly Scenario 2.

In addition to analyzing the performance of the estimated weights in matching the given household-level marginal distributions, the performance of the estimated weights in matching the given person-level marginal distributions were also analyzed. Table 14 present results from the comparison of the estimated sample household weights in matching the marginal distributions of various controlled and uncontrolled person-level variables of interest. As noted earlier, in each of the scenarios, two person-level variables were controlled at the individual county level including age and gender. Additionally, one variable was controlled at the state level (representing the multilevel control) namely race of the person. Similar to the comparison of household-level marginal distributions, the marginal distributions of all controlled variables are closely matched. The deviations in the person-level marginal distributions are generally higher than the household-level marginal distributions because in the iterative procedures for estimating weights (including IPU and Entropy), the marginal distributions are often inconsistent and a corner solution where household-level constraints are matched is chosen.

For person-level variables that were controlled in all three scenarios, it can be seen that the estimated weights match the given marginal distributions very closely. Scenario 3 performs best with comparable performance for Scenarios 1 and 2. Further, as expected, the performance of Scenario 1 is least in matching the marginal distribution of the race variable because the variable is not controlled in this scenario. On the other hand in Scenarios 2 and 3, the match in the race variable is very close because the variable is explicitly controlled at the state level. The match in uncontrolled variable was also compared to evaluate performance of the procedures for sample household weighting and also to assess the scenario setups. Scenario 3 performs best followed by comparable performance for Scenarios 1 and 2.

Lastly, it is interesting to note that the match in person-level marginal distributions is consistently better for Scenario 3 for all controlled and uncontrolled variables. This observation

combined with the reasonable performance of Entropy procedure in matching household-level marginal distribution of interest possibly allude to the potentially better performance of the Entropy iterative procedure for estimating weights compared to the IPU procedure. This observation is also consistent with the aggregate comparison of household and person total presented in the previous subsection.

Comparison of the Sample Household Weights in Matching Given Multiway Distributions

In an effort to evaluate the performance of the sample household weighting procedures and also the scenario setups, the performance of the estimated weights in matching multiway distributions was analyzed. The multiway distributions are derived from the given marginal distributions by applying the IPF procedure. Figures 6 through 8 show the match between the weighted sum (calculated from the estimated weights) and constraints (estimated from the given marginal distributions) qualitatively for the three scenarios respectively. From the figures, it can be seen that the weighted sums match both household- and person-level multiway constraints closely at both the county and state spatial resolutions in all the three scenarios as they follow the 45 degree line closely. However, closer inspection of the actual deviation values reveals subtle but important differences in performance across the different Scenarios. Table 15 presents the minimum and maximum values of the deviation between the weighted sums and the constraints for the three Scenarios. For each scenario, the deviation between estimated weights and the multiway constraints is generally higher at the person-level compared to the household-level. This is expected because in the iterative procedures settle on the corner solution where household-level constraints are closely matched.

At the state level only one household- and one person-level attribute was controlled. Thus, the constraints at the region level are nothing but the marginal distributions for the respective variables. The deviation values are the highest in Scenario 1 for multiway constraints at the state level. This is expected because the constraints are not controlled in Scenario 1. Between Scenarios 2 and 3, the deviation in the household-level attribute namely presence of children in the household is comparable across the scenarios. However, the deviation in the person-level attribute namely race of the person is lower for Scenario 3 employing the Entropy procedure compared to Scenario 2 employing the IPU procedure. This is again consistent with earlier observations of a better match in person-level variables for the Entropy procedure compared to the IPU procedure in the presence of the multilevel controls.

At the county level multiple household- and person-level attributes were controlled. Thus, the constraints represent cells in the multiway distributions defined by the respective attributes. It can be seen that the deviation values for the different constraints at the county-level are comparable between Scenarios 1 and 2 with a slightly superior performance for Scenario 2 with the multilevel constraints. It is also interesting to note that for all constraints, the deviation values are less extreme for Scenario 3 compared to Scenarios 1 and 2.

Conclusions

In this chapter, the two approaches for estimating sample household weights namely IPU and Entropy procedures were applied to estimate county-level sample household weights for state of all 24 counties in the state of Maryland. Data from Census 2000 was used to prepare the sample and marginal distribution inputs. The focus of the analysis was twofold namely 1) to evaluate the ability of the two approaches for accommodating multilevel controls and 2) to assess the value of the additional controls at higher levels of spatial resolution. Following are the key takeaways from the analysis that was conducted.

- The enhanced IPU and Entropy procedures are both capable of accommodating multilevel controls.
- As expected (and also confirmed by the scenario analysis), not accounting for the household- and person-level control variables of interest at higher spatial resolution that are readily available will cause discrepancies between the estimated sample weights (and subsequently synthetic population generated) and known marginal distributions of interest. Therefore, it is desirable to accommodate additional controls at higher spatial resolution when data is available to ensure that a representative synthetic population is generated.
- The performance of the estimated weights in matching the additional controls improved as expected when the multilevel controls were provided. However, the difference in performance in matching control variables at lower spatial resolution and also the performance in matching uncontrolled variables is only marginal with the introduction of multilevel controls.
- Entropy procedure appears to consistently provide a better fit in person-level controls compared to the IPU procedure. However, the better fit in person-level controls comes at the expense of a slightly poor fit in the household-level controls. It may be desirable to apply Entropy procedure when a better match in person-level controls is desired. Another consideration that must go into the choice of the sample weighting process is the computational overhead. It was observed that the processing time for applying IPU procedure (Scenario 2) is approximately an order of magnitude lower than the processing time for applying Entropy procedure (Scenario 3).

Figure 3: Convergence Properties for Scenario 1 - Synthetic Population Generation without Multilevel Controls using IPU

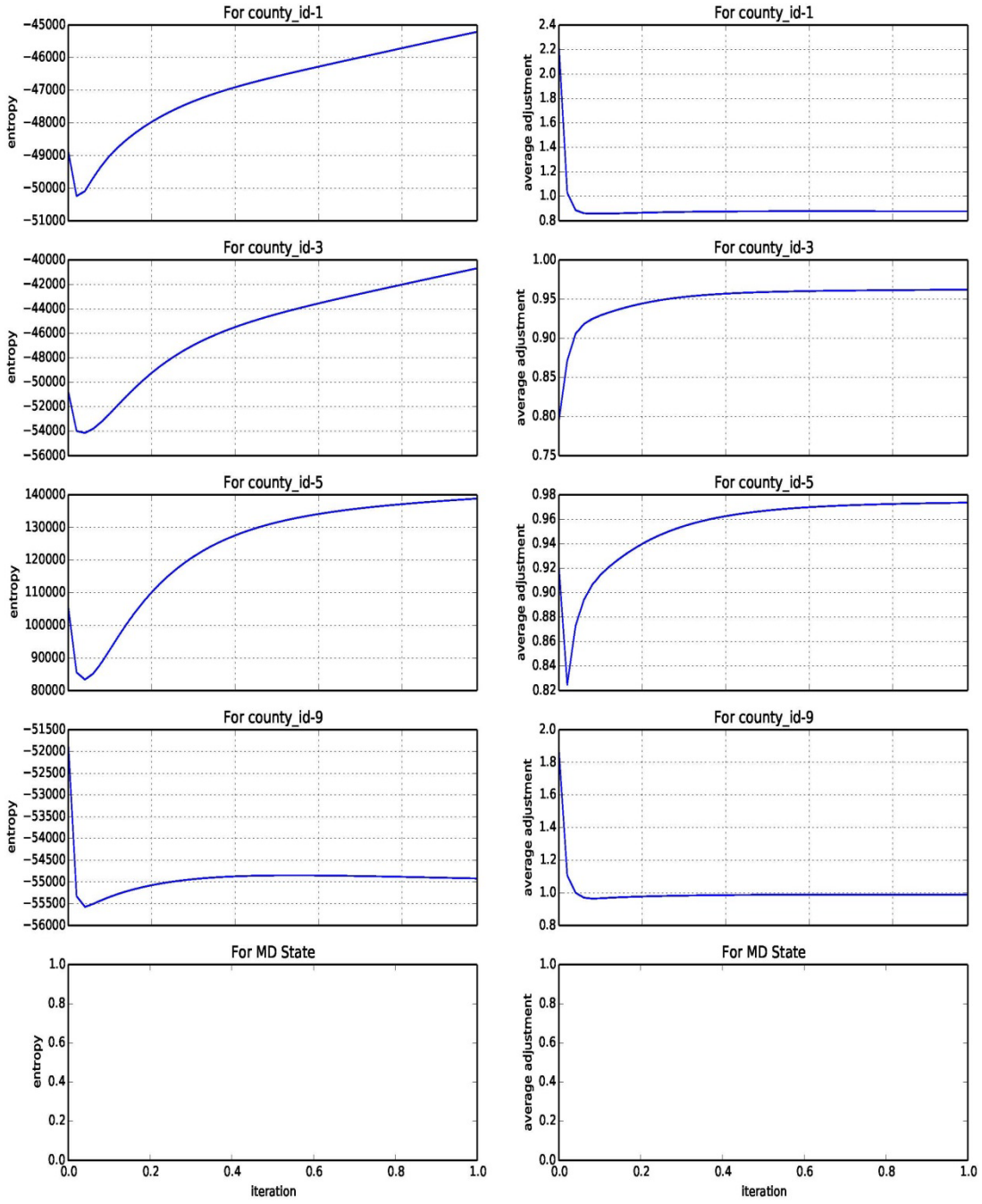

Figure 4: Convergence Properties for Scenario 2 - Synthetic Population Generation with Multilevel Controls using IPU

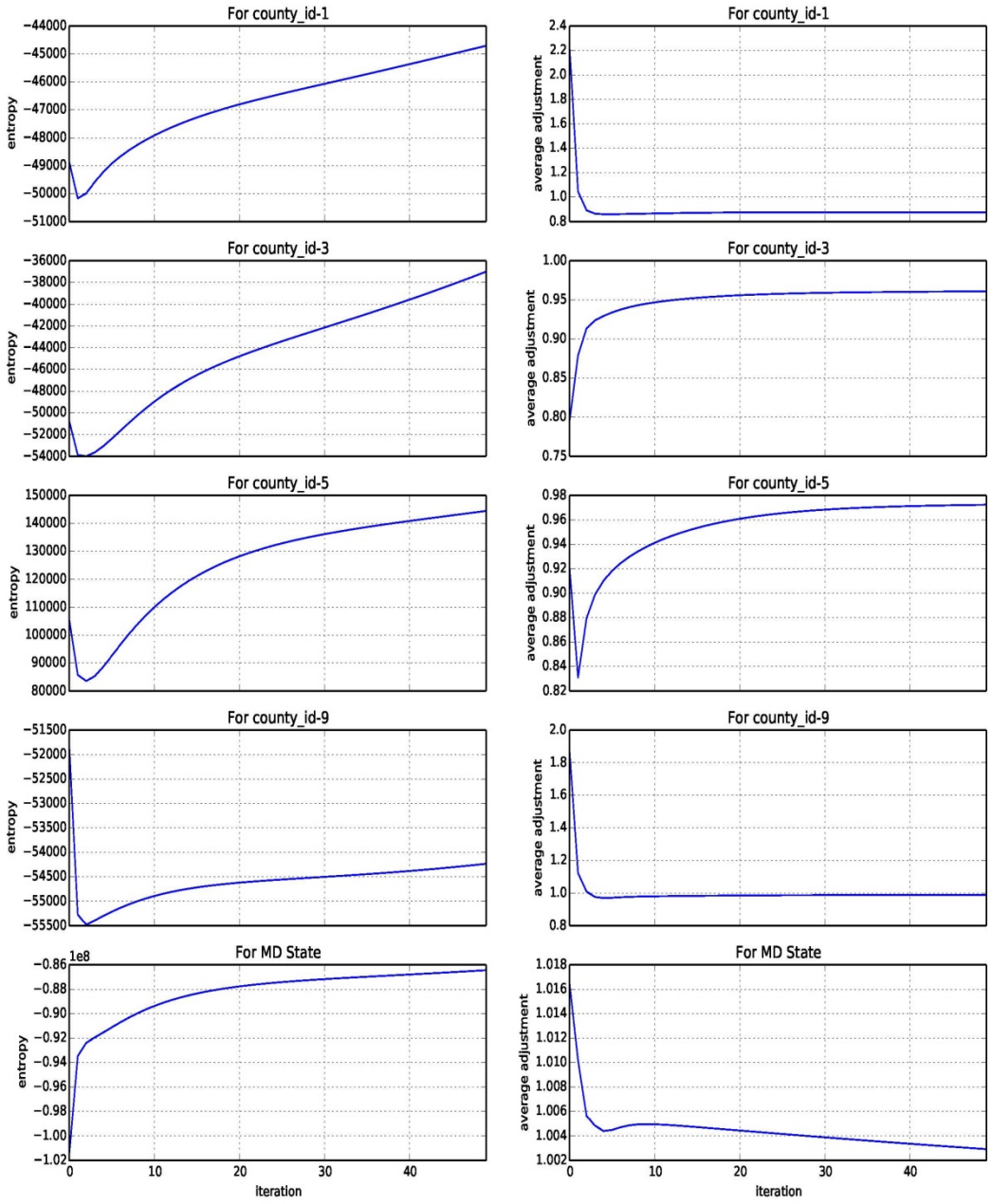

Figure 5: Convergence Properties for Scenario 3 - Synthetic Population Generation with Multilevel Controls using Entropy

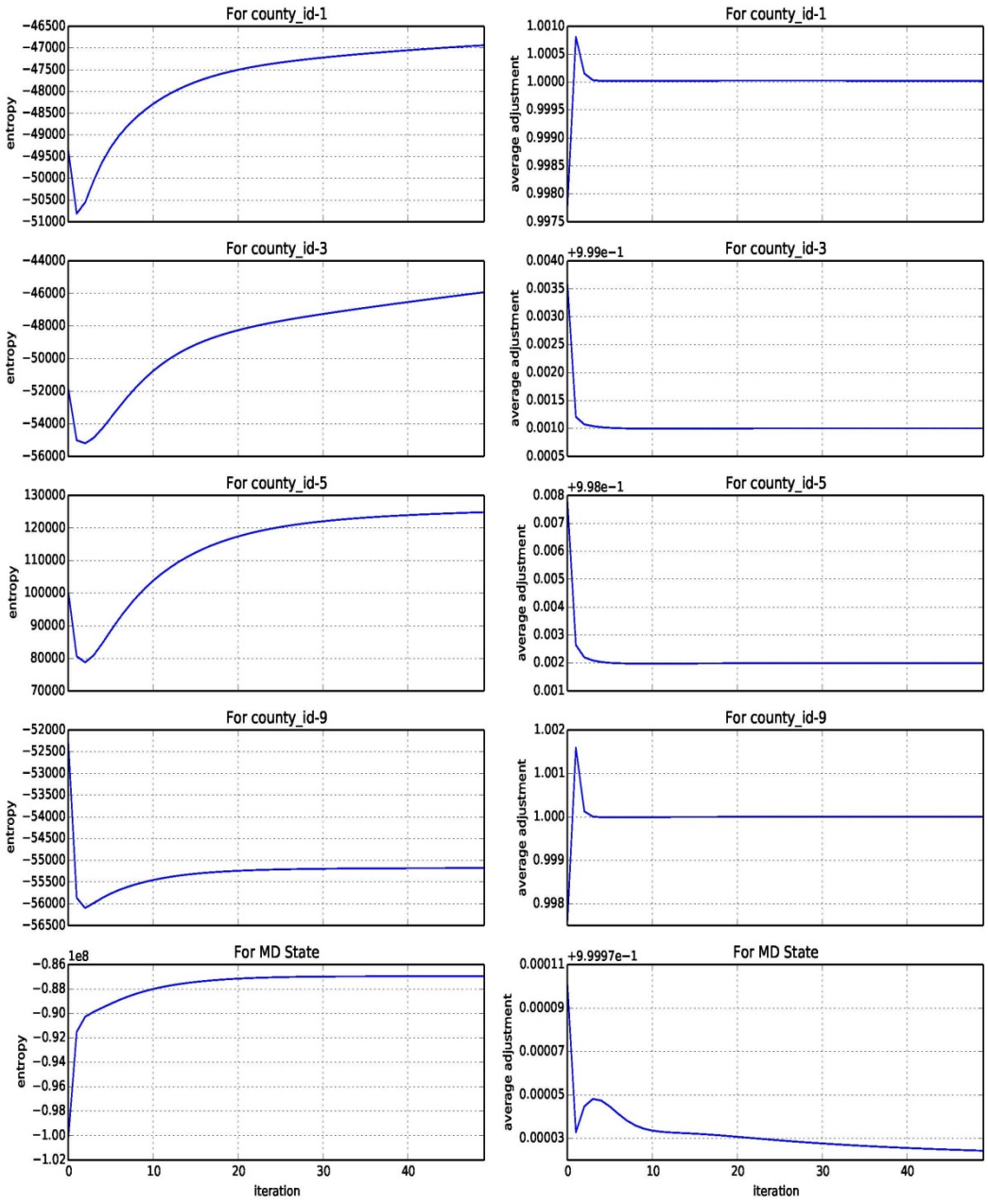

Table 12: Aggregate Comparison of the Sample Household Weights

Spatial Resolution of Comparison	Measure of Comparison	Given	Scenario 1			Scenario 2			Scenario 3		
			Weighed Sum	Diff	% Diff	Weighted Sum	Diff	% Diff	Weighted Sum	Diff	% Diff
Region	Person total	5296486.0	5202069.1	-94416.9	-1.8	5201050.7	-95435.3	-1.8	5247223.4	-49262.6	-0.9
	Household total	1981795.0	1981795.0	0.0	0.0	1981795.0	0.0	0.0	1981795.0	0.0	0.0
County 1	Person total	74930.0	69172.0	-5758.0	-7.7	69153.3	-5776.7	-7.7	69504.8	-5425.2	-7.2
	Household total	29350.0	29350.0	0.0	0.0	29344.6	-5.4	0.0	29318.5	-31.5	-0.1
County 3	Person total	489656.0	477569.8	-12086.2	-2.5	477535.9	-12120.1	-2.5	482413.0	-7243.0	-1.5
	Household total	178754.0	178754.0	0.0	0.0	178767.5	13.5	0.0	178790.9	36.9	0.0
County 5	Person total	754292.0	742073.8	-12218.2	-1.6	741938.7	-12353.3	-1.6	747976.8	-6315.2	-0.8
	Household total	300020.0	300020.0	0.0	0.0	300007.0	-13.0	0.0	299843.9	-176.1	-0.1
County 9	Person total	74563.0	73984.4	-578.6	-0.8	73990.4	-572.6	-0.8	74234.2	-328.8	-0.4
	Household total	25428.0	25428.0	0.0	0.0	25430.3	2.3	0.0	25455.3	27.3	0.1

Notes:

1) Scenario 1 - Synthetic Population without Multilevel Controls using IPU

2) Scenario 2 - Synthetic Population with Multilevel Controls using IPU

3) Scenario 3 - Synthetic Population with Multilevel Controls using Entropy

Table 13: Comparison of the Sample Household Weights in Matching Given Household-level Marginal Distributions

Variable Name	Category	Resolution at which Variable is Controlled	Given	Scenario 1		Scenario 2		Scenario 3	
				Weighted Sum	% Diff	Weighted Sum	% Diff	Weighted Sum	% Diff
Household Type	1	Controlled at individual county level	1015033	1011624	-0.3	1011824	-0.3	1014103	-0.1
	2		82569	82296	-0.3	82301	-0.3	82484	-0.1
	3		271045	270155	-0.3	270067	-0.4	270740	-0.1
	4		494959	498524	0.7	498433	0.7	495948	0.2
	5		118189	119196	0.9	119170	0.8	118520	0.3
Household Size	1	Controlled at individual county level	494959	493685	-0.3	493336	-0.3	490587	-0.9
	2		627558	627764	0.0	627937	0.1	626173	-0.2
	3		349932	350238	0.1	350308	0.1	350774	0.2
	4		297941	298403	0.2	298415	0.2	300008	0.7
	5		133238	133424	0.1	133473	0.2	134631	1.0
	6		51859	51929	0.1	51969	0.2	52586	1.4
	7		26308	26351	0.2	26357	0.2	27035	2.8
Household Income	1	Controlled at individual county level	220527	220527	0.0	220404	-0.1	219907	-0.3
	2		188104	188104	0.0	188020	0.0	187706	-0.2
	3		212135	212135	0.0	212058	0.0	211831	-0.1
	4		210424	210424	0.0	210382	0.0	210225	-0.1
	5		283274	283274	0.0	283293	0.0	283337	0.0
	6		508027	508027	0.0	508180	0.0	508729	0.1
	7		230285	230285	0.0	230401	0.1	230789	0.2
	8		129019	129019	0.0	129057	0.0	129271	0.2
Household Child Presence Indicator	1	Controlled at state level (where applicable)	675659	643607	-4.7	675659	0.0	675659	0.0
	2		1306136	1338188	2.5	1306136	0.0	1306136	0.0
Householder Age	1	Uncontrolled	415068	464297	11.9	465252	12.1	455438	9.7
	2		913212	910865	-0.3	914213	0.1	919742	0.7
	3		278383	267837	-3.8	266154	-4.4	267817	-3.8
	4		375132	338796	-9.7	336176	-10.4	338798	-9.7

Notes:
1) Scenario 1 - Synthetic Population without Multilevel Controls using IPU
2) Scenario 2 - Synthetic Population with Multilevel Controls using IPU
3) Scenario 3 - Synthetic Population with Multilevel Controls using Entropy

Table 14: Comparison of the Sample Household Weights in Matching Given Household-level Marginal Distributions

Variable Name	Category	Given	Resolution at which Variable is Controlled	Scenario 1		Scenario 2		Scenario 3	
				Weighted Sum	% Diff	Weighted Sum	% Diff	Weighted Sum	% Diff
Person Age	1	351443	Controlled at individual county level	344241	-2.0	344272	-2.0	347748	-1.1
	2	785407		770407	-1.9	769865	-2.0	777701	-1.0
	3	664041		654846	-1.4	654711	-1.4	658435	-0.8
	4	744251		731353	-1.7	731370	-1.7	737828	-0.9
	5	930256		908585	-2.3	908330	-2.4	919422	-1.2
	6	753808		741337	-1.7	741308	-1.7	747526	-0.8
	7	469276		461557	-1.6	461519	-1.7	465295	-0.8
	8	322605		318577	-1.2	318515	-1.3	320298	-0.7
	9	210255		207420	-1.3	207418	-1.3	208610	-0.8
	10	65144		63745	-2.1	63743	-2.2	64360	-1.2
Person Gender	1	2554588	Controlled at individual county level	2511110	-1.7	2510875	-1.7	2532082	-0.9
	2	2741898		2690959	-1.9	2690175	-1.9	2715141	-1.0
Person Race	1	3391021	Controlled at state level (where applicable)	3371603	-0.6	3330814	-1.8	3359462	-0.9
	2	1468243		1378974	-6.1	1441194	-1.8	1454464	-0.9
	3	15651		14895	-4.8	15367	-1.8	15518	-0.9
	4	209713		214066	2.1	205797	-1.9	207777	-0.9
	5	2030		1995	-1.7	1993	-1.8	2013	-0.8
	6	96773		107052	10.6	94942	-1.9	95997	-0.8
	7	113055		113484	0.4	110943	-1.9	111993	-0.9
Person Employment	1	1210544	Uncontrolled	1195804	-1.2	1200225	-0.9	1212403	0.2
	2	2640623		2612777	-1.1	2605002	-1.3	2614917	-1.0
	3	128902		118978	-7.7	117881	-8.6	121036	-6.1
	4	1316417		1274511	-3.2	1277943	-2.9	1298867	-1.3

Notes:

1) Scenario 1 - Synthetic Population without Multilevel Controls using IPU

2) Scenario 2 - Synthetic Population with Multilevel Controls using IPU

3) Scenario 3 - Synthetic Population with Multilevel Controls using Entropy

Figure 6: Performance of the Estimated Weights in Matching the Given Multiway Distributions for Scenario 1 - Synthetic Population Generation without Multilevel Controls using IPU

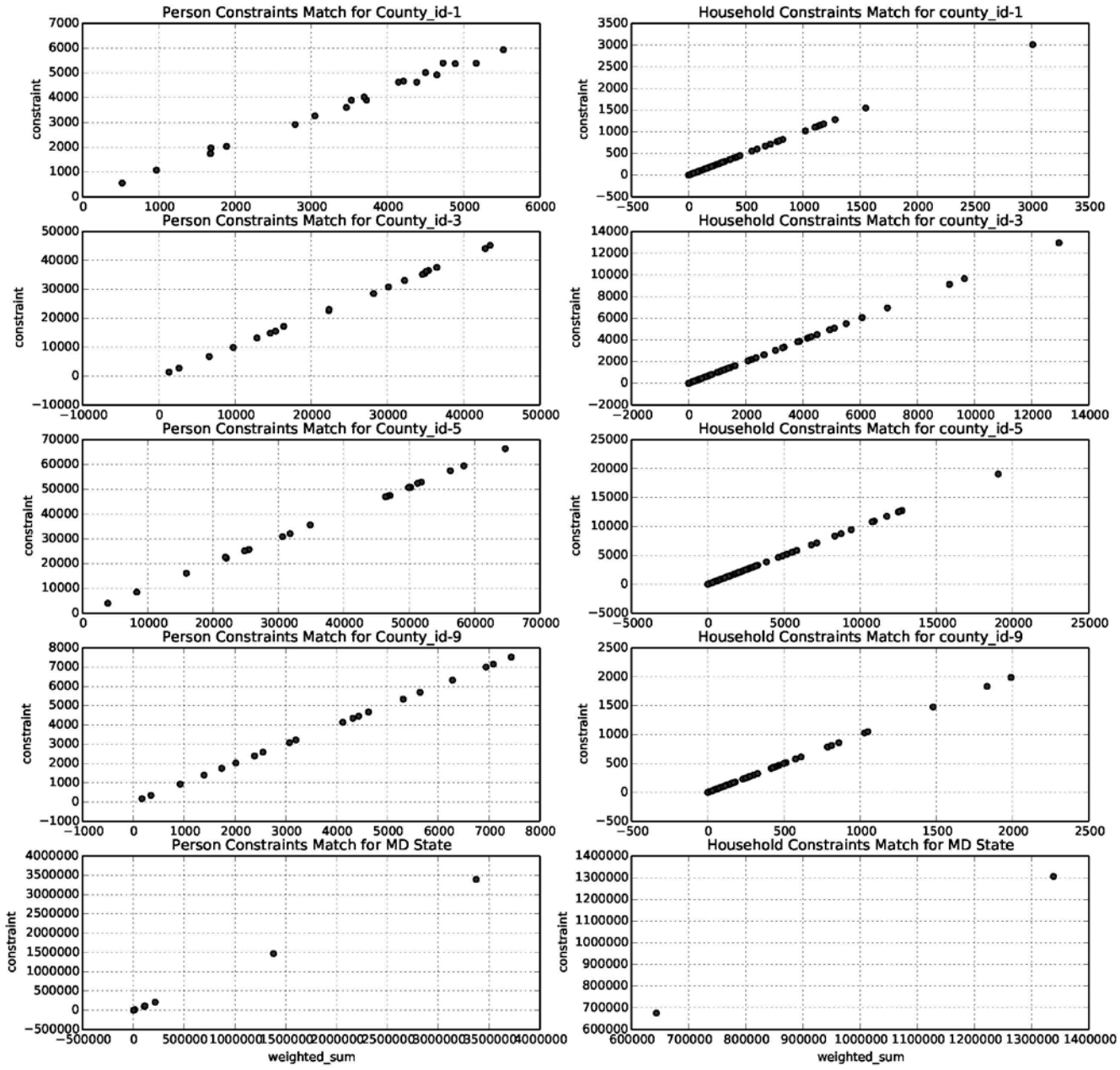

Figure 7: Performance of the Estimated Weights in Matching the Given Multiway Distributions for Scenario 2 - Synthetic Population Generation with Multilevel Controls using IPU

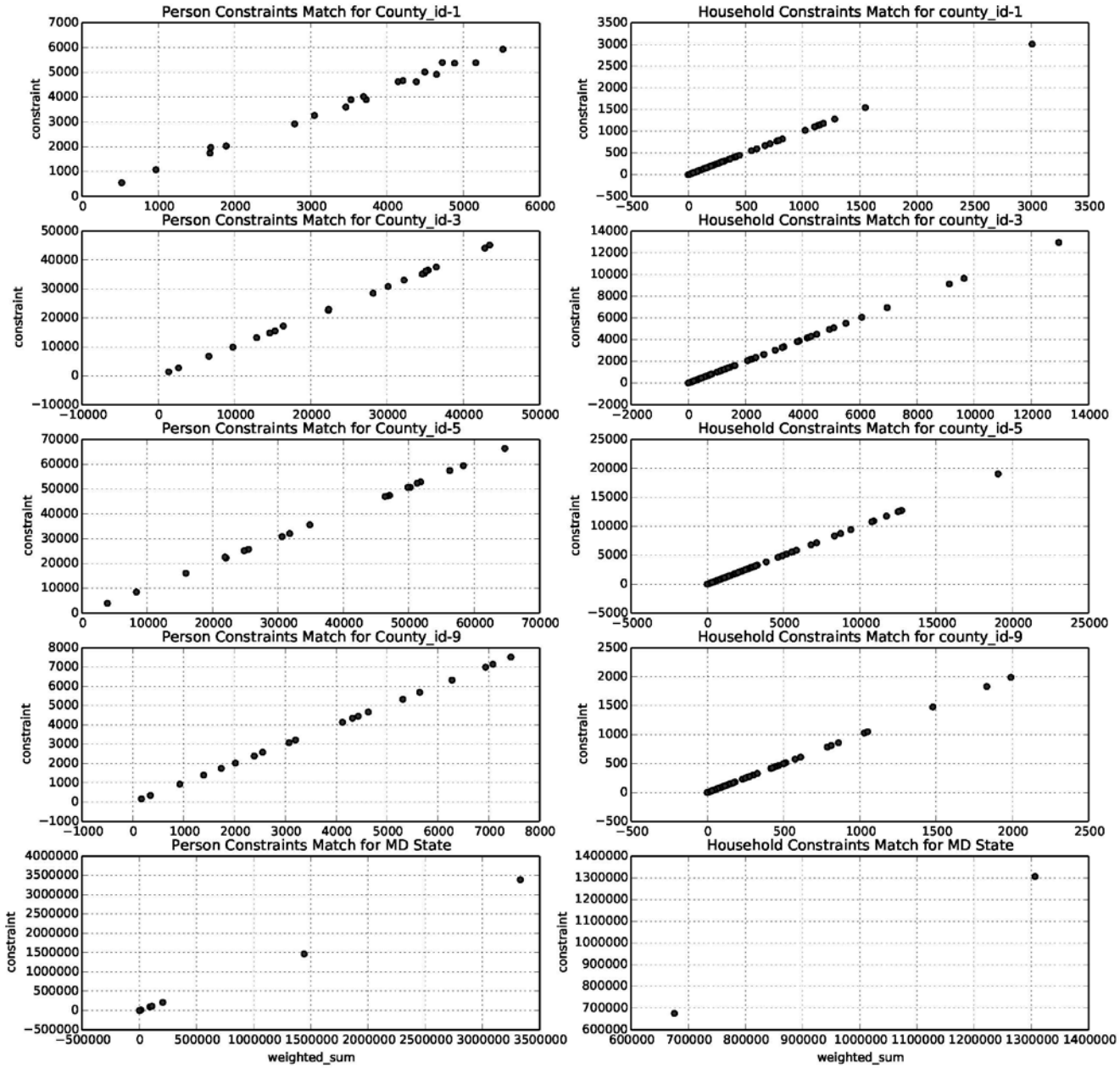

Figure 8: Performance of the Estimated Weights in Matching the Given Multiway Distributions for Scenario 3 - Synthetic Population Generation with Multilevel Controls using Entropy

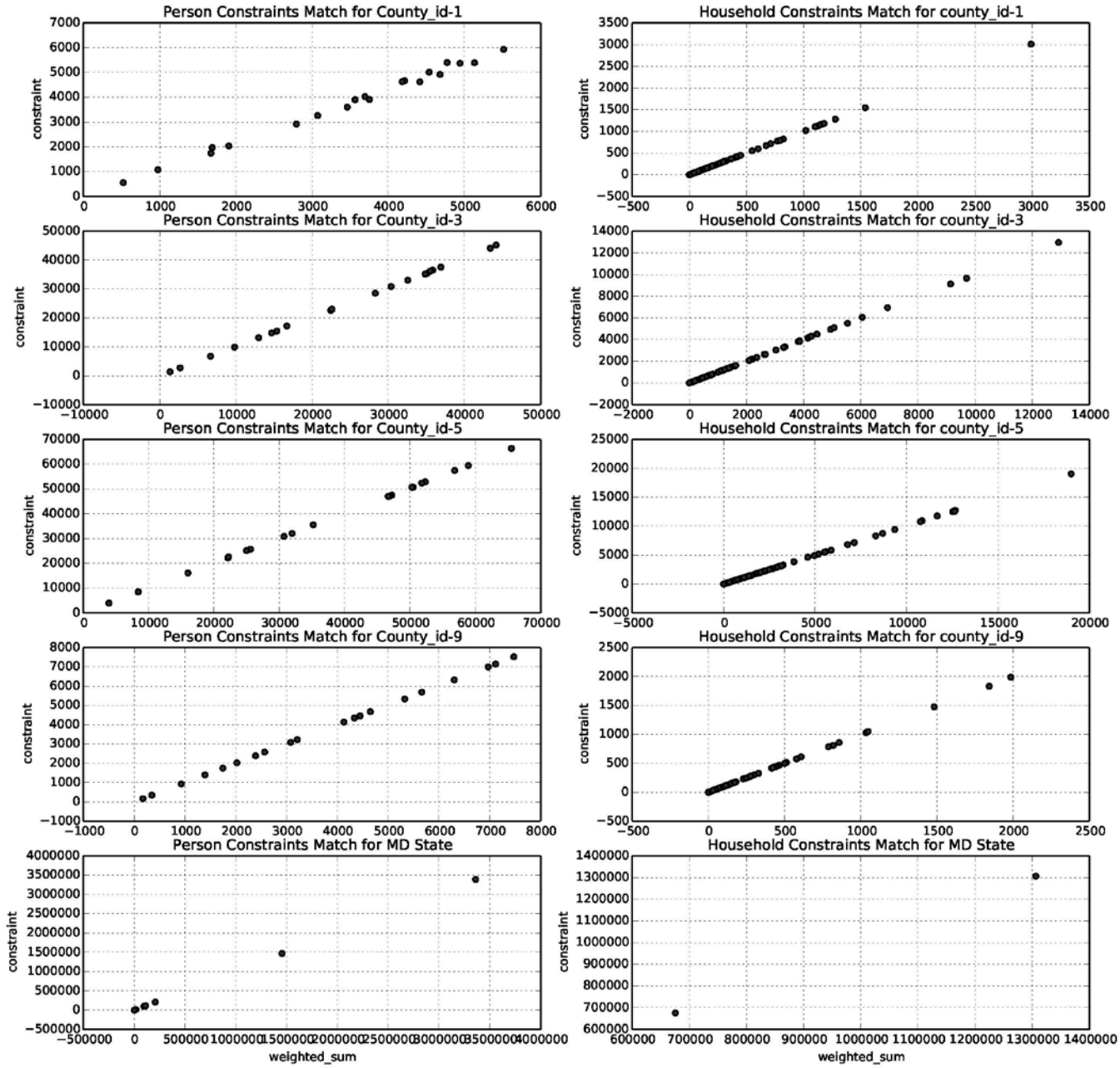

Table 15: Comparison of the Sample Household Weights in Matching the Given Multiway Distributions

Spatial Resolution of Comparison	Multiway Distribution of Comparison	Scenario 1		Scenario 2		Scenario 3	
		Min Diff	Max Diff	Min Diff	Max Diff	Min Diff	Max Diff
Region	Person-level - race	-6.1	10.6	-1.9	-1.8	-0.9	-0.8
	Household-level - presence of children in household	-4.7	2.5	0.0	0.0	0.0	0.0
County 1	Person-level - age X gender	-14.5	-3.8	-14.4	-3.8	-14.3	-3.9
	Household-level - household type X household size X household income	0.0	0.0	-1.1	3.1	-0.6	4.0
County 3	Person-level - age X gender	-4.7	-1.1	-4.6	-1.1	-2.9	-0.7
	Household-level - household type X household size X household income	0.0	0.0	-1.1	3.1	-0.6	4.1
County 5	Person-level - age X gender	-3.2	-0.8	-3.2	-0.8	-1.7	-0.4
	Household-level - household type X household size X household income	0.0	0.0	-1.1	3.1	-0.6	3.8
County 9	Person-level - age X gender	-1.4	-0.3	-1.3	-0.3	-0.8	-0.2
	Household-level - household type X household size X household income	0.0	0.0	-1.1	3.1	-0.6	3.8

Notes:

1) Scenario 1 - Synthetic Population without Multilevel Controls using IPU

2) Scenario 2 - Synthetic Population with Multilevel Controls using IPU

3) Scenario 3 - Synthetic Population with Multilevel Controls using Entropy

CHAPTER 4: DEMONSTRATION OF THE ENAHNCED IPU-BASED APPROACH TO POPULATION SYNTHESIS

The specific objectives of the analysis presented in this chapter are twofold namely 1) to demonstrate the applicability of the enhanced approaches for generating a synthetic population while controlling for marginal distributions at TAZ- and county-level spatial resolutions and 2) to demonstrate the value of additional controls at higher levels of spatial resolution in generating a synthetic population. To this end the enhanced IPU approach is applied to generate a synthetic population at the Traffic Analysis Zone (TAZ) level for the 2012 model year. The remaining chapter is organized as follows. The input data and the population synthesis setup are described in the next section. In the following section, results are presented starting with the convergence properties of the IPU and entropy procedures for estimating county-level sample household weights. In the next section, performance results of the estimated sample household weights are presented. Some concluding thoughts are presented in the last section.

Description of the Population Synthesis Scenarios

Results from a TAZ-level population synthesis using the enhanced population synthesis methodologies for the 2012 model year are presented below. Unlike the previous two chapters where the focus was on the sample household weight estimation step of synthetic population generation, in this analysis, both steps of the population synthesis process (including sample household weight estimation and drawing households) were carried out. The performance results compare the generated synthetic population against given marginal distribution values. The marginal distributions at both TAZ- and county-levels provided by BMC were used in the analysis. Sample data was prepared using 2008-2012 5-year American Community Survey (ACS) Public Use Microdata Sample (PUMS) data. The model region consists of ten counties across Maryland and District of Columbia namely District of Columbia, Anne Arundel, Baltimore, Carroll, Frederick, Harford, Howard, Montgomery, Prince George's, and Baltimore City.

The marginal distributions at the person-level include people who live in both households and groupquarters, therefore, both households and groupquarter housing units were synthesized simultaneously. The marginal distributions that were available at the household-level include TAZ-level marginal distributions of household size, household income category, and number of workers in the household, and county-level marginal distributions of householder age. To perform the different scenarios (described below), county-level marginal distributions of household size, household income category, and number of workers were derived from the TAZ-level totals by aggregating across all TAZs that belong to a given county unit. Marginal distributions available at the groupquarter level include TAZ-level distribution of type of groupquarter, and county-level marginal distribution of total number of groupquarter units. The marginal distributions available at the person-level include TAZ-level marginal distribution of employment status and county-level marginal distribution of age of individuals. Total numbers of synthetic population units to be synthesized include 2,076,236 households (from the number of workers distribution), 145,718 groupquarters (from the type of groupquarter unit distribution) and 5,416,563 (based on the employment status distribution).

Synthetic population was generated jointly for TAZs in the state of Maryland and District of Columbia. This entailed preparing a single sample dataset by stacking 2008-2012 ACS PUMS records from Maryland and District of Columbia. The household sample file consisted of

123,027 records, groupquarter sample file consisted of 8,912 and person sample file consisted of 310,252 records. While PUMS records only belonging to the PUMA to which an individual geographic unit (i.e. TAZ) belongs were used in the Iterative Proportional Fitting (IPF) step of the synthetic population generation, sample records from both Maryland and District of Columbia were used when estimating sample household weights and when drawing households.

The enhanced synthetic population methodology was implemented in PopGen 2.0 software. More details regarding the software are included in Chapter 5. In this analysis, the IPU procedure was used for generating a synthetic population. The focus of the analysis was not on illustrating differences between the IPU and Entropy procedures. As noted in the previous chapter, there are small differences in performance between the IPU and Entropy procedures and the choice of a procedure for an analysis depends on the objective for a particular use case and available computational resources. However, the trends and observations identified across scenarios using IPU procedure described below will also hold for the Entropy procedure.

In this analysis, seven different scenarios were performed as described below:

- All marginal distributions at the TAZ- and county-level were controlled (Scenario 1): Household-level control variables include household size (5 category), household income (5 category), and number of workers (4 category) at the TAZ-level, and age of householder (6 category) at the county-level. Person-level control variables include employment status (2 category) at the TAZ-level, and age of person (18 category) at the county-level. Groupquarter-level control variables include type of groupquarter (2 category) at the TAZ level, and number of groupquarter units (1 category) at the county-level.
- No worker count control at the TAZ-level (Scenario 2): All control variables are the same as Scenario 1. The only exception being household-level worker count variable is not controlled at the TAZ-level.
- No household size control at the TAZ-level (Scenario 3): All control variables are the same as Scenario 1. The only exception being household-level person count variable is not controlled at the TAZ-level.
- No household income control at the TAZ-level (Scenario 4): All control variables are the same as Scenario 1. The only exception being household-level income variable is not controlled at the TAZ-level.
- Worker count control at the region-level (Scenario 5): All control variables are the same as Scenario 1. The only exception being household-level worker count variable is controlled at county-level instead of at the TAZ-level.
- Household size control at the region-level (Scenario 6): All control variables are the same as Scenario 1. The only exception being household-level person count variable is controlled at county-level instead of at the TAZ-level.
- Household income control at the region-level (Scenario 7): All control variables are the same as Scenario 1. The only exception being household-level income variable is controlled at county-level instead of at the TAZ-level.

Scenario 1 served as the baseline providing a synthetic population that utilizes all marginal distribution information that is available. Scenarios 2 to 7 comprise variations of scenario 1 and represent synthetic populations generated using a subset of the information available. Results from Scenario 1 provide evidence in support of the applicability of the enhanced population synthesis procedure. Population synthesis results from all Scenarios shed light on the value of including additional control variables at various spatial resolutions when generating a synthetic population.

Results

In this section, aggregate to disaggregate comparisons of the synthetic population generated in the different scenarios are presented.

Aggregate Comparison of the Synthetic Population Totals

Table 16 presents an aggregate comparison of the synthetic population generated in the different scenarios against the given totals. As mentioned in the previous section, the given total values for the number of household units were generated using the worker count variable – this information is important because by using a marginal distribution for a different variable, the total would be slightly different because of known inconsistencies between the marginal distributions of household-level control variables. As expected, the count of households in the synthetic population matches the given totals perfectly. This observation is reasonable because the synthetic population is drawn such that the total values perfectly match the given total of households for each individual geographic unit. Small deviations (39 less units of households) in the household totals can be observed for Scenario 2 and Scenario 5. The difference in scenario 2 is reasonable because worker count is not included as a control variable. As a result, number of households will match the total implied by the household income category – the last variable controlled during the IPF step of the population synthesis. Even though worker count is included as a control variable in Scenario 5, differences are still observed because worker count is used as a control at the region level. Region level control variables only influence the sample household weight estimates but they do not influence the number of households that are synthesized. Number of households synthesized is influenced by the control variables at the individual geography level (i.e. household size, and household income) and the last variable controlled in the IPF step at the individual geography level (i.e. household income). Since household income is the last household level variable that is controlled at the individual geography level in Scenario 5, the synthesized household totals match the household totals derived from the marginal distribution of the income variable (i.e. 2,076,197).

The number of synthesized groupquarters perfectly matches the given totals for all scenarios. This is not surprising because only one control variable was used to control for synthesizing groupquarter units at the individual geography level. In synthetic population generation process whenever there is a single control variable for a housing-level synthetic population entity (e.g. households, groupquarters), the marginal distribution including the total for the entity are perfectly matched. The person totals for all scenarios are closely matched with deviations of up to two percent. The differences are reasonable because household-level marginal distributions and person-level marginal distributions appear to be inconsistent. This observation can be further reinforced by observing the performance of the synthetic population whenever household size variables are controlled at the individual geography level, the difference in person total is close to -1.6%. Individual geography level constraints serve as hard constraints that must be adhered to and if they are at odds with person-level marginal distributions then there will be higher deviations in the person totals. In such instances, the synthetic population generation procedure selects a corner solution where the household-level marginal distributions are matched perfectly and person-level marginal distributions are matched only closely. It is interesting to note that in Scenario 3 when household size variable is not used as a control variable, the difference in person total is the highest. This also points to the opposite effect of the household size variable wherein excluding the variable is lowering the performance of the synthetic population. The observation also alludes to the importance of including household-level control variables that provide information about household composition when

synthesizing a population. In the absence of household composition variables, the number of persons synthesized will be inaccurate.

It can also be seen from Table 16 that in Scenario 6 when household size variable is included as a control variable but only at the region-level, the difference in the person total is the least. This observation shows that by including a household composition variable even if it is only at the region level, the fit in person total increases compared to when it is not considered at all. The observation also points to the behavior of the synthetic population as a function of the count of variables at the individual geographic level. It can be seen that when the number of control variables at the individual geographic level are high (household size, household income, and worker count in Scenario 1), the match in the person totals is low whereas when the number of control variables are fewer (household income and worker count in Scenario 6), the match in person totals is better. It must be noted that the only difference between Scenario 1 and Scenario 6 is that household size variable is used as a control at the region level as opposed to at the individual geography level in Scenario 6. Similar observations of a better match in person totals can be made when moving one of the control variables from the individual geography level to the region level in Scenarios 5 and 7. Even though there is an improvement in the match of person total by moving a control variable from the individual geography level to the region level, more detailed analysis of performance must be performed at a disaggregate level to confirm whether the better match in person total also translates to other measures of synthetic population performance.

Comparison of the Marginal Distributions for the Entire Model Region

Tables 17a and 17b show the comparison of marginal distributions for the various household-level controlled and uncontrolled variables for the entire model region. Tables 18a and 18b show the comparison of marginal distributions for person-level and groupquarter-level controlled and uncontrolled variables. For any given scenario, the controlled variables are highlighted in color and uncontrolled variables are not highlighted. Variables that are controlled at the individual geographic unit level are highlighted in yellow and variables that are controlled at the region level are highlighted in green. It must also be noted that the marginal distribution for householder age and age of person at the region level doesn't include values for District of Columbia (County FIPS Code 001). As a result, comparison of the absolute values of marginal distribution for these variables is not reasonable. Instead percentage distribution values of the synthetic population are compared against given values of the percentage distribution across the 9 counties (for which region-level distributions are provided) to test for reasonableness of the generated synthetic population.

At the region-level, percentage distribution of householder age compares well with given percentage distributions across all scenarios (Tables 17a and 17b). Across all scenarios there is an overestimation of about 1.6% to 2.2% percent for the first category of householder age (i.e. age of householder from 14 to 24 years) and an overestimation from 0.05% to 0.3% for second category of householder age (i.e. age of householder from 25 to 34 years). For the fourth, fifth, and sixth categories of householder age there is the opposite trend with a small percentage of underestimation ranging from -0.1% to -1.1%. For the third category of householder age, both overestimation (Scenarios 1, and 6) and underestimation (Scenarios 2, 3, 4, 5, and 7) in the percentage for the category are observed across scenarios. There isn't a significant difference in percentage distribution for the householder age variable even when additional control variables are included at the region levels (i.e. Scenarios 5, 6, and 7) compared to scenarios when there is a

single control variable included at the region level (i.e. Scenarios 2, 3, 4). They follow similar trends with small changes in the deviation percentage across categories.

At the region-level, percentage distribution of person age compares well with the given percentage distribution across all scenarios (Tables 18a and 18b). The deviation in the percentage value across any category ranges from -0.7% to 1.4%. Similar to the householder age variable, there isn't a significant difference in the percentage distribution for the age of person variable when additional household-level control variables are included at the region level. This is reasonable because additional control variables (worker count in Scenario 5, household size in Scenario 6, and household income in Scenario 7) do not contain any extra information about person age to improve the fit. Similar to the comparison of the person totals, Scenario 6 appears to offer the best match with the person age percentage distribution with deviation in percentage values ranging from -0.3% and 0.2%.

Unlike the householder age and person age variables at the region-level, absolute values of the synthetic population marginal distributions for all other variables can be compared directly against their given values. It can be seen from the tables that whenever a variable is controlled at the individual geographic unit level (highlighted in yellow) for a given scenario, the synthetic population marginal distribution for the variable almost perfectly closely matches the given marginal distribution. The small deviations in matching the marginal distributions are in part due to the order in which the control variables are used in the IPF step of the population synthesis – the later the variable is controlled in the IPF process, the better the match. The deviations are also in part due to the rounding procedure in the household drawing step of the population synthesis wherein non-integer constraints are converted into integer values to obtain the counts of different household types to draw. On the other hand, whenever a variable is not controlled at the individual geography level, the synthetic population doesn't match the given marginal distribution for the variable. The deviation values are large with percentages ranging from -48.7% (for category 5 of the household size variable in Scenario 3) to 28.6% (for category 3 of the household size variable in Scenario 3) when the variable is not controlled at all (Scenarios 2, 3, and 4). However, the deviation values drop considerably when the variable is controlled at the region-level with values ranging from -2.6% (for category 3 of worker count variable in Scenario 5) to 5.6% (for category 1 of worker count variable in Scenario 5). These observations point to the importance of including control variables at the individual geography level for synthesizing an accurate synthetic population. Further, the results also indicate that if marginal distributions for a variable are not readily available at the individual geography level but are available at higher levels of spatial resolution then a synthetic population that controls the variable at higher level of spatial resolution will provide a more accurate synthetic population then one that doesn't consider the variable at all.

Among the three household-level variables, excluding household size variable (i.e. Scenario 3) as a control at any spatial resolution has the highest impact on the quality of the synthetic population with percent deviations varying from -48.7% to 28.6% compared to excluding worker count and household income variables (i.e. Scenarios 2, and 4) with percent deviations varying from -18.0% to 20.9% and -16.3% to 13.6% respectively. Among scenarios where one of the three household-level variables is excluded at the individual geography level (i.e. Scenarios 5, 6 and 7), Scenario 7 performs the best in matching given marginal distributions followed by Scenario 6 and Scenario 5 in the same order. Performance is best when size variables (i.e. variables providing information about number of person units in the household) namely number of persons and number of workers are controlled at the individual geography.

When at least one of the size variables is not included, the performance of the synthetic population drops.

From Tables 18a and 18b, it can also be seen that the synthetic population matches the groupquarter marginal distribution perfectly. On the other hand synthetic population matches the person-level marginal distribution of employment type only closely. The deviations in the distributions follow observations of person total match described in the previous subsection. Scenario 6 appears to perform best. The smallest absolute percentage deviation is for Scenario 6 and the highest absolute percentage deviation is for Scenario 1. Across all scenarios the number of the people that are unemployed is underestimated. However, in Scenarios 1, 4, 6, and 7 count of people who are employed is overestimated and in Scenarios 2, 3, and 5, count of people who are employed is underestimated. It looks like whenever persons and workers are both controlled (irrespective of whether they are controlled at the individual geography level or at the region level), the number of people who are employed is overestimated.

Comparison of the Marginal Distributions at the Level of Individual Geographic Unit

In an effort to further confirm whether observations of synthetic population performance hold across scenarios, a more disaggregate analysis was performed by comparing the synthetic population against the given marginal distribution values for the different control variables at the level of the individual geographic unit. Only marginal distributions for the household size, household income, and worker count were compared at the individual geography level. Therefore, for each scenario, measures of match between the synthetic population and the marginal distributions for these three variables were compared. For each scenario, Tables 19a and 19b present summary statistics for absolute percentage difference for different categories of household size, household income and worker count variables across different individual geographic units. Similar to observations presented in the previous subsection, performance of the synthetic population is best in scenarios where the variables are controlled at the individual geography level followed by scenarios in which the variables are controlled at the region level and finally by scenarios in which the variables are not controlled at all. Again, among scenarios where one of the household-level variable was not controlled, household size appears to have the highest implication on the performance of the synthetic population as can be seen from the high values of average absolute percent deviation when household size was not controlled (Scenario 3). It is interesting to note that in the counterpart scenarios (i.e. Scenarios 5, 6, and 7) wherein one of the household-level variables is controlled at the region level, Scenario 6 appears to perform the best followed by Scenario 5 and Scenario 7. It appears like not including income at the individual geography level results in the poorest performance followed by scenarios where one of the size variables is included. However the order of performance of the synthetic population is in contrast to the observations from Table 17 when the overall marginal distributions for the entire model region were compared – in that comparison, Scenario 7 seemed to perform best followed by Scenario 6 and Scenario 5. The ordering of the synthetic population performance based on the comparison of marginal distributions at the individual geographic level is more accurate. This observation also points to the importance of carrying out detailed performance analysis of the synthetic population by evaluating various aggregate and disaggregate measures to identify the best configuration of inputs and parameters for generating a synthetic population.

Conclusions

In this chapter, the IPU based enhanced methodology was applied to generate a TAZ-level synthetic population for the BMC model region for the 2012 model year. To this end, the specific objectives of the analysis presented in this chapter are twofold namely 1) to demonstrate the applicability of the multilevel population synthesis procedures and 2) to demonstrate the value of additional controls at higher levels of spatial resolution when generating a synthetic population. Following are the key takeaways from the analysis that was conducted.

- The enhanced population synthesis procedure can be applied to generate a synthetic population for the BMC use case i.e. to generate a TAZ level synthetic population while controlling for marginal distributions at county- and TAZ-levels. Synthetic population performance is consistent with expectations.
- It is desirable to control for variables at the lowest spatial resolution when possible. When marginal distributions for control variables of interest are not available at the lowest spatial resolution, the variables can be controlled at higher spatial resolution. Controlling for variables even if only at higher spatial resolution results in a synthetic population that is more representative of the underlying population than not controlling for the variable at all.
- Among the different household-level attributes, size variables i.e. variables that provide information about the number and type of people in the household play an important role not only in improving the fit of the household-level attributes but also in improving the fit of person-level attributes. An important consideration when including multiple size variables is to ensure consistency.

Table 16: Aggregate Comparison of the Synthetic Population Totals

Scenario	Household Total				Groupquarter Total				Person Total			
	Given	Synthesized	Diff	% Diff	Given	Synthesized	Diff	% Diff	Given	Synthesized	Diff	% Diff
Scenario 1	2076236	2076236	0	0.00	145718	145718	0	0.00	5416563	5329428	-87135	-1.63
Scenario 2	2076236	2076197	-39	0.00	145718	145718	0	0.00	5416563	5330086	-86477	-1.62
Scenario 3	2076236	2076236	0	0.00	145718	145718	0	0.00	5416563	5319189	-97374	-1.83
Scenario 4	2076236	2076236	0	0.00	145718	145718	0	0.00	5416563	5331153	-85410	-1.60
Scenario 5	2076236	2076197	-39	0.00	145718	145718	0	0.00	5416563	5330839	-85724	-1.61
Scenario 6	2076236	2076236	0	0.00	145718	145718	0	0.00	5416563	5367195	-49368	-0.92
Scenario 7	2076236	2076236	0	0.00	145718	145718	0	0.00	5416563	5336594	-79969	-1.50

Table 17a: Comparison of Household-level Marginal Distributions for Entire Model Region for Scenarios 1 through 4

Variable Name	Category	Given	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
			Synthesized	% Diff	Synthesized	% Diff	Synthesized	% Diff	Synthesized	% Diff
Householder Age	1	3.4%	5.1%	1.6%	5.5%	2.1%	5.2%	1.7%	5.2%	1.8%
	2	15.8%	15.9%	0.2%	16.1%	0.3%	15.9%	0.1%	15.8%	0.0%
	3	18.0%	18.0%	0.1%	17.5%	-0.4%	17.8%	-0.1%	17.7%	-0.3%
	4	21.6%	20.9%	-0.7%	20.7%	-0.9%	20.9%	-0.7%	20.9%	-0.7%
	5	19.6%	19.1%	-0.5%	19.1%	-0.5%	19.1%	-0.4%	19.3%	-0.3%
	6	21.7%	20.9%	-0.7%	21.0%	-0.7%	21.1%	-0.6%	21.1%	-0.5%
	Total	**100.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**
Household Size	1	603,074	601,063	-0.3%	603,085	0.0%	443,417	-26.5%	601,062	-0.3%
	2	610,091	611,070	0.2%	610,082	0.0%	744,250	22.0%	611,054	0.2%
	3	359,188	359,687	0.1%	359,192	0.0%	461,738	28.6%	359,711	0.1%
	4	281,083	281,415	0.1%	281,067	0.0%	312,553	11.2%	281,409	0.1%
	5	222,776	223,001	0.1%	222,771	0.0%	114,278	-48.7%	223,000	0.1%
	Total	**2,076,212**	**2,076,236**	**0.0%**	**2,076,197**	**0.0%**	**2,076,236**	**0.0%**	**2,076,236**	**0.0%**
Household Income Category	1	204,447	204,047	-0.2%	204,302	-0.1%	204,451	0.0%	171,024	-16.3%
	2	205,957	205,711	-0.1%	206,042	0.0%	205,936	0.0%	233,672	13.5%
	3	310,696	310,603	0.0%	310,740	0.0%	310,710	0.0%	284,128	-8.6%
	4	675,915	675,941	0.0%	675,961	0.0%	675,953	0.0%	616,002	-8.9%
	5	679,182	679,934	0.1%	679,152	0.0%	679,186	0.0%	771,410	13.6%
	Total	**2,076,197**	**2,076,236**	**0.0%**	**2,076,197**	**0.0%**	**2,076,236**	**0.0%**	**2,076,236**	**0.0%**
Household Worker Count	0	432,102	431,868	-0.1%	388,469	-10.1%	432,140	0.0%	431,989	0.0%
	1	755,855	756,010	0.0%	913,897	20.9%	755,925	0.0%	755,995	0.0%
	2	718,006	718,095	0.0%	588,931	-18.0%	718,191	0.0%	717,944	0.0%
	3	170,273	170,263	0.0%	184,900	8.6%	169,980	-0.2%	170,308	0.0%
	Total	**2,076,236**	**2,076,236**	**0.0%**	**2,076,197**	**0.0%**	**2,076,236**	**0.0%**	**2,076,236**	**0.0%**

Table 17b: Comparison of Household-level Marginal Distributions for Entire Model Region for Scenarios 5 through 7

Variable Name	Category	Given	Scenario 5		Scenario 6		Scenario 7	
			Synthesized	% Diff	Synthesized	% Diff	Synthesized	% Diff
Householder Age	1	3.4%	5.6%	2.2%	5.2%	1.7%	5.3%	1.9%
	2	15.8%	15.7%	0.0%	15.9%	0.1%	15.8%	0.1%
	3	18.0%	17.3%	-0.6%	17.9%	0.0%	17.7%	-0.3%
	4	21.6%	20.5%	-1.1%	21.0%	-0.6%	20.9%	-0.7%
	5	19.6%	19.2%	-0.3%	19.1%	-0.5%	19.2%	-0.3%
	6	21.7%	21.6%	-0.1%	20.8%	-0.8%	21.1%	-0.6%
	Total	**100.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**
Household Size	1	603,074	603,085	0.0%	592,464	-1.8%	601,062	-0.3%
	2	610,091	610,082	0.0%	606,421	-0.6%	611,054	0.2%
	3	359,188	359,192	0.0%	362,484	0.9%	359,711	0.1%
	4	281,083	281,067	0.0%	285,848	1.7%	281,409	0.1%
	5	222,776	222,771	0.0%	229,019	2.8%	223,000	0.1%
	Total	**2,076,212**	**2,076,197**	**0.0%**	**2,076,236**	**0.0%**	**2,076,236**	**0.0%**
Household Income Category	1	204,447	204,302	-0.1%	204,451	0.0%	205,036	0.3%
	2	205,957	206,042	0.0%	205,936	0.0%	207,482	0.7%
	3	310,696	310,740	0.0%	310,710	0.0%	311,733	0.3%
	4	675,915	675,961	0.0%	675,953	0.0%	675,684	0.0%
	5	679,182	679,152	0.0%	679,186	0.0%	676,301	-0.4%
	Total	**2,076,197**	**2,076,197**	**0.0%**	**2,076,236**	**0.0%**	**2,076,236**	**0.0%**
Household Worker Count	0	432,102	456,460	5.6%	432,140	0.0%	431,989	0.0%
	1	755,855	746,288	-1.3%	755,925	0.0%	755,995	0.0%
	2	718,006	699,355	-2.6%	718,191	0.0%	717,944	0.0%
	3	170,273	174,094	2.2%	169,980	-0.2%	170,308	0.0%
	Total	**2,076,236**	**2,076,197**	**0.0%**	**2,076,236**	**0.0%**	**2,076,236**	**0.0%**

Table 18a: Comparison of Person- and Groupquarter-level Marginal Distributions for Entire Model Region for Scenarios 1 through 4

Variable Name	Category	Given	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
			Synthesized	% Diff	Synthesized	% Diff	Synthesized	% Diff	Synthesized	% Diff
Groupquarter Type	1	52,638	52,638	0.0%	52,638	0.0%	52,638	0.0%	52,638	0.0%
	2	93,080	93,080	0.0%	93,080	0.0%	93,080	0.0%	93,080	0.0%
	Total	**145,718**	**145,718**	**0.0%**	**145,718**	**0.0%**	**145,718**	**0.0%**	**145,718**	**0.0%**
Number of groupquarters	**1**	145,718	145,718	0.0%	145,718	0.0%	145,718	0.0%	145,718	0.0%
	Total	**145,718**	**145,718**	**0.0%**	**145,718**	**0.0%**	**145,718**	**0.0%**	**145,718**	**0.0%**
Employment status	1	2,745,282	2,763,166	0.7%	2,742,923	-0.1%	2,741,164	-0.2%	2,762,681	0.6%
	2	2,671,281	2,566,262	-3.9%	2,587,163	-3.1%	2,578,025	-3.5%	2,568,472	-3.8%
	Total	**5,416,563**	**5,329,428**	**-1.6%**	**5,330,086**	**-1.6%**	**5,319,189**	**-1.8%**	**5,331,153**	**-1.6%**
Person Age	1	6.2%	5.9%	-0.3%	5.6%	-0.6%	5.7%	-0.5%	5.6%	-0.6%
	2	6.3%	6.1%	-0.2%	5.6%	-0.6%	5.7%	-0.5%	5.7%	-0.6%
	3	6.5%	6.0%	-0.5%	5.8%	-0.7%	5.9%	-0.5%	5.8%	-0.6%
	4	6.8%	7.1%	0.3%	7.1%	0.3%	7.3%	0.5%	7.6%	0.8%
	5	6.9%	7.3%	0.4%	8.3%	1.4%	7.4%	0.5%	7.9%	1.0%
	6	7.2%	7.3%	0.2%	7.9%	0.7%	7.4%	0.2%	7.2%	0.1%
	7	6.8%	7.0%	0.2%	6.8%	0.1%	7.0%	0.3%	6.9%	0.2%
	8	6.5%	6.5%	0.0%	6.4%	-0.2%	6.7%	0.2%	6.4%	-0.2%
	9	6.9%	6.8%	-0.1%	6.6%	-0.2%	7.0%	0.1%	6.7%	-0.2%
	10	7.5%	7.3%	-0.2%	7.3%	-0.2%	7.4%	-0.1%	7.3%	-0.2%
	11	7.6%	7.4%	-0.2%	7.3%	-0.2%	7.5%	-0.1%	7.3%	-0.2%
	12	6.7%	6.8%	0.0%	6.8%	0.1%	6.8%	0.1%	6.8%	0.1%
	13	5.6%	5.7%	0.1%	5.8%	0.2%	5.8%	0.2%	5.8%	0.2%
	14	4.2%	4.1%	-0.1%	4.1%	-0.1%	3.9%	-0.3%	4.1%	-0.1%
	15	2.9%	2.9%	0.0%	2.9%	0.0%	2.8%	-0.1%	2.9%	0.0%
	16	2.1%	2.2%	0.1%	2.2%	0.0%	2.1%	0.0%	2.2%	0.1%
	17	1.6%	1.7%	0.1%	1.7%	0.1%	1.7%	0.1%	1.8%	0.1%
	18	1.8%	1.8%	0.1%	1.8%	0.1%	1.8%	0.0%	1.9%	0.1%
	Total	**100.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**

Table 18b: Comparison of Person- and Groupquarter-level Marginal Distributions for Entire Model Region for Scenarios 5 through 7

Variable Name	Category	Given	Scenario 5		Scenario 6		Scenario 7	
			Synthesized	% Diff	Synthesized	% Diff	Synthesized	% Diff
Groupquarter Type	1	52,638	52,638	0.0%	52,638	0.0%	52,638	0.0%
	2	93,080	93,080	0.0%	93,080	0.0%	93,080	0.0%
	Total	**145,718**	**145,718**	**0.0%**	**145,718**	**0.0%**	**145,718**	**0.0%**
Number of groupquarters	**1**	145,718	145,718	0.0%	145,718	0.0%	145,718	0.0%
	Total	**145,718**	**145,718**	**0.0%**	**145,718**	**0.0%**	**145,718**	**0.0%**
Employment status	1	2,745,282	2,727,209	-0.7%	2,762,025	0.6%	2,749,545	0.2%
	2	2,671,281	2,603,630	-2.5%	2,605,170	-2.5%	2,587,049	-3.2%
	Total	**5,416,563**	**5,330,839**	**-1.6%**	**5,367,195**	**-0.9%**	**5,336,594**	**-1.5%**
Person Age	1	6.2%	5.7%	-0.6%	6.0%	-0.3%	5.7%	-0.6%
	2	6.3%	5.7%	-0.6%	6.3%	0.1%	5.7%	-0.6%
	3	6.5%	5.8%	-0.6%	6.5%	0.1%	5.9%	-0.6%
	4	6.8%	7.2%	0.4%	6.8%	0.0%	7.5%	0.7%
	5	6.9%	8.4%	1.4%	6.9%	0.0%	7.9%	1.0%
	6	7.2%	7.3%	0.1%	7.2%	0.1%	7.3%	0.1%
	7	6.8%	6.8%	0.1%	6.9%	0.2%	6.9%	0.1%
	8	6.5%	6.4%	-0.1%	6.6%	0.1%	6.4%	-0.2%
	9	6.9%	6.7%	-0.1%	6.9%	0.0%	6.7%	-0.2%
	10	7.5%	7.3%	-0.2%	7.4%	-0.1%	7.3%	-0.2%
	11	7.6%	7.3%	-0.2%	7.4%	-0.2%	7.4%	-0.2%
	12	6.7%	7.0%	0.3%	6.7%	0.0%	6.8%	0.1%
	13	5.6%	6.0%	0.4%	5.8%	0.1%	5.8%	0.2%
	14	4.2%	4.0%	-0.2%	4.0%	-0.2%	4.1%	-0.1%
	15	2.9%	2.8%	-0.1%	2.8%	-0.1%	2.9%	0.0%
	16	2.1%	2.1%	0.0%	2.2%	0.0%	2.2%	0.1%
	17	1.6%	1.7%	0.1%	1.7%	0.1%	1.7%	0.1%
	18	1.8%	1.8%	0.0%	1.8%	0.1%	1.9%	0.1%
	Total	**100.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**	**100.0%**	**0.0%**

Table 19a: Comparison of Marginal Distributions at the Level of Individual Geographic Unit for Scenarios 1 through 4

Variable Name	Category	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
		Min	Max	Average	Min	Max	Average	Min	Max	Average	Min	Max	Average
Household Income	1	0.00	1.00	0.03	0.00	1.00	0.03	0.00	0.33	0.00	0.00	42.00	0.66
	2	0.00	3.00	0.03	0.00	2.00	0.03	0.00	1.00	0.00	0.00	47.00	0.71
	3	0.00	1.00	0.02	0.00	2.00	0.02	0.00	1.00	0.00	0.00	27.00	0.55
	4	0.00	2.00	0.02	0.00	2.00	0.01	0.00	1.00	0.00	0.00	24.57	0.52
	5	0.00	1.00	0.01	0.00	1.00	0.01	0.00	0.33	0.00	0.00	211.00	1.32
Household Size	1	0.00	1.00	0.01	0.00	1.00	0.00	0.00	3.70	0.33	0.00	1.00	0.01
	2	0.00	5.50	0.02	0.00	1.00	0.01	0.00	70.00	0.32	0.00	5.50	0.02
	3	0.00	5.67	0.01	0.00	0.50	0.00	0.00	12.00	0.38	0.00	5.67	0.01
	4	0.00	6.00	0.01	0.00	0.50	0.00	0.00	15.00	0.24	0.00	6.00	0.01
	5	0.00	2.50	0.01	0.00	1.00	0.00	0.00	1.00	0.47	0.00	2.50	0.01
Household Worker Count	0	0.00	2.00	0.03	0.00	27.00	0.49	0.00	1.00	0.01	0.00	1.00	0.01
	1	0.00	1.00	0.02	0.00	72.00	0.37	0.00	2.00	0.01	0.00	0.67	0.01
	2	0.00	1.00	0.02	0.00	47.00	0.25	0.00	1.00	0.01	0.00	1.00	0.01
	3	0.00	2.00	0.04	0.00	20.00	0.37	0.00	1.00	0.01	0.00	1.00	0.01

Table 19b: Comparison of Marginal Distributions at the Level of Individual Geographic Unit for Scenarios 5 through 7

Variable Name	Category	Scenario 5			Scenario 6			Scenario 7		
		Min	Max	Average	Min	Max	Average	Min	Max	Average
Household Income	1	0.00	1.00	0.03	0.00	0.33	0.00	0.00	51.00	0.69
	2	0.00	2.00	0.03	0.00	1.00	0.00	0.00	29.00	0.50
	3	0.00	2.00	0.02	0.00	1.00	0.00	0.00	37.00	0.56
	4	0.00	2.00	0.01	0.00	1.00	0.00	0.00	25.00	0.49
	5	0.00	1.00	0.01	0.00	0.33	0.00	0.00	327.00	0.94
Household Size	1	0.00	1.00	0.00	0.00	3.70	0.12	0.00	1.00	0.01
	2	0.00	1.00	0.01	0.00	18.75	0.16	0.00	5.50	0.02
	3	0.00	0.50	0.00	0.00	3.33	0.13	0.00	5.67	0.01
	4	0.00	0.50	0.00	0.00	3.50	0.13	0.00	6.00	0.01
	5	0.00	1.00	0.00	0.00	2.13	0.15	0.00	2.50	0.01
Household Worker Count	0	0.00	30.00	0.35	0.00	1.00	0.01	0.00	1.00	0.01
	1	0.00	99.00	0.26	0.00	2.00	0.01	0.00	0.67	0.01
	2	0.00	51.00	0.18	0.00	1.00	0.01	0.00	1.00	0.01
	3	0.00	19.00	0.25	0.00	1.00	0.01	0.00	1.00	0.01

CHAPTER 5: POPGEN 2.0 – SOFTWARE IMPLEMENTATION OF THE ENHANCED POPULATION SYNTHESIS APPROACHES

The enhanced population synthesis methodology was implemented as a stand-alone software package dubbed PopGen 2.0. PopGen 2.0 builds on legacy versions PopGen 1.0 and 1.1. However, PopGen 2.0 is a complete reimplementation of the software and it features a number of desirable features that enhance the user experience, and improve computational performance. Additionally, the software codebase now employs software best practices that will ensure longevity of the software beyond the life of the project. BMC can access future versions of the PopGen software by visiting the PopGen GitHub repository. Detailed instructions for installing PopGen are provided in Appendix A and steps for using the software are described in Appendix B.

REFERENCES

- Ye, X., Konduri, K.C., Sana, B., and Pendyala, R.M. (2009). A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations. Proceedings of the 88th Annual Meeting of the Transportation Research Board, Washington, DC.
- Bar-Gera, H., Konduri, K.C., Sana, B., Ye, X., and Pendyala, R.M. (2009). Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. Proceedings of the 88th Annual Meeting of the Transportation Research Board, Washington, DC.
- Beckman, R.J., Baggerly, K.A., Mckay, M.D. (1996). Creating synthetic baseline populations. Transportation Research Part A, 30(6), 415-429.
- Pendyala, R.M. and K.P. Christian (2011) PopGen 1.1 User's Guide. Lulu.com Publishers, USA. ~113 pages

APPENDIX A: INSTALLING POPGEN 2.0

In this appendix, instructions for installing PopGen 2.0 software are described. The general approach to installing PopGen 2.0 is very similar to the approach for installing PopGen 1.1 (Pendyala and Christian 2011). The installation instructions provided in this appendix are specific to Windows operating system. However, PopGen 2.0 is not platform specific and can be run on computers with other operating systems including Mac OSX and Linux. Also, one can setup 32-bit and 64-bit installations of PopGen 2.0. In this document instructions for both 32-bit and 64-bit installations of PopGen 2.0 are provided. For installation instructions for other operating systems, please visit the PopGen GitHub repository.

Note: From this point forward PopGen 2.0 will be referred to as just PopGen – any reference to the word PopGen in the remaining Appendix should be associated with the version 2.0.

There are two steps for installing PopGen as described in the following sections. In the first section, instructions for installing Python programming language and dependencies are presented. In the second section, instructions for installing PopGen are provided.

Install Python Programming Language and Dependencies

Instructions for Installing Python Programming Language

Users must first install Python programming language. Python Version 2.7 is recommended because PopGen software has been developed and tested against this version. Installers for both 32-bit and 64-bit Windows operating systems can be downloaded from the Python website. Depending on the project setup (i.e. number of geographies, size of the synthetic population), PopGen may require large amounts of RAM which may not always be accommodated by 32-bit installations. Therefore, it is recommended that 64-bit version of the Python programming language and libraries are installed to avoid any runtime issues due to insufficient memory. After downloading the installer, double-click the installer and follow the instructions in the wizard to complete the Python programming language installation.

Install Python Libraries

After installing Python programming language, users must install the following Python libraries. PopGen utilizes these libraries for specifying a project, for implementing the different population synthesis algorithms, and for analyzing the results.

- PyYAML (Version 3.11): PyYAML is a python library that is capable of parsing YAML files.
- Numpy (Version 1.9.2): Numpy is a scientific python library that facilitates array and matrix operations.
- Scipy (Version 0.15.1): Scipy is a scientific python library that builds on top of Numpy and features functionality to develop software for applications in mathematics, science, and engineering.
- Pandas (Version 0.16.1): Pandas is a scientific python library that provides high-performance and user-friendly data structures and data analysis tools.

In addition to identifying the names of the libraries, versions of the libraries are also identified. PopGen has been tested against these versions of the libraries. Therefore, it is recommended that same versions of the libraries be installed (to the extent possible). While it is possible to run

PopGen with other versions of the libraries, there may be undesirable behaviors because of unknown incompatibilities between versions of Python and/or libraries. Users must install Numpy, Scipy and Pandas in the same order because each builds on the other.

If a 32-bit (64-bit) version of Python programming language is installed then 32-bit (64-bit) version of the Python libraries must also be installed. Installers for 32-bit versions of the Python libraries can easily be accessed from the official websites of these libraries. Links to the official websites of the libraries are provided below:

- PyYAML (Version 3.11): http://pyyaml.org/
- Numpy (Version 1.9.2): http://www.numpy.org/
- Scipy (Version 0.15.1): http://www.scipy.org/
- Pandas (Version 0.16.1): http://pandas.pydata.org/

32-bit installers can also be accessed from the official Python Package Index. To install any library, download the installation file, then double-click the installation file and follow the instructions in the wizard.

Unlike the 32-bit version of libraries, installing 64-bit version of the libraries is an involved process. 64-bit versions of the installation files for the above libraries are not readily available. Some alternative options to install the 64-bit versions of the Python libraries are presented below. The options are ordered in increasing order of difficulty.

Option 1: Install a Python Distribution (Recommended for all users)

An easy way to overcome the installation issues with 64-bit versions of the Python libraries is to use a Python distribution such as Anaconda. Anaconda is a pre-packaged distribution that includes 64-bit versions of both the Python programming language and a host of popular libraries useful for data analytics and scientific computing. Anaconda distribution includes the four libraries noted above for running PopGen. Users can install Anaconda distribution by downloading the installation file from the Anaconda website. The only downside to using Anaconda (and more generally any Python distribution) is that it comes pre-packaged with a number of other libraries that may not all be useful – Anaconda distribution comes pre-packaged with more than 200 libraries in addition to the four that are needed for running PopGen.

Option 2: Install Individual Packages from Binaries

- 64-bit PyYAML can be installed by downloading the 64-bit installation file from the PyYAML page on Python Package Index or by visiting the download page on the official website.
- 64-bit versions of the other three packages including Numpy, Scipy, and Pandas must be installed from their Python Wheels. Wheels are a new way to distribute Python libraries wherein users can install the libraries without compiling from source code. Wheels for the three libraries can be accessed for some libraries (namely Pandas) from Python Package Index and for other packages (namely Scipy, and Numpy) from unofficial resources on the web. For example, Christoph Gohlke maintains 32-bit and 64-bit wheels for a host scientific Python libraries on his website at http://www.lfd.uci.edu/~gohlke/pythonlibs/ including Numpy, Scipy, and Pandas. After downloading the wheel file from either the official or unofficial resources, pip (a tool for installing Python libraries) must be used to install the library. To install any library, issue the following command from the shell:

```
c:\default\windows\shell>pip install <full_location_of_wheel_file.whl>
```

Note: Installation file for 32-bit and 64-bit PyYAML (Version 3.11) and Wheel files for Numpy (Version 1.9.2), Scipy (Version 0.15.1) and Pandas (Version 0.16.1) have been archived on June 11, 2015. If users have any difficulty finding specific versions of the files to install these libraries, they can access them from Google drive folder maintained by the authors of PopGen.

Option 3: Install from Source Code

Download the source code for the libraries from the official website (see above) and issue the following command for each:

```
c:\library\source\code\root\folder>python setup,py install
```

It must be noted that some of the libraries may require other software to be installed before the library can be compiled from source code.

Install PopGen

There are two ways to install PopGen as described below:

Option 1: Install from Source Code

Source code for the latest version of PopGen can be downloaded from https://github.com/foss-transportationmodeling/popgen/releases. PopGen codebase is distributed as a zip file – download and extract the zip file (advanced users can also choose to check out the latest development version of the code from the GitHub master repository). The code must be compiled before it can be used for setting up population synthesis projects. Steps to compile the code are presented below:

- Step 1: Open the command line shell and browse to the root folder of the PopGen codebase. The root folder refers to the folder that contains the docs, popgen, test, and tutorial folders.

```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Corporation.  All rights reserved.
```

- Step 2: From the root folder issue the following command:

```
C:\pogen\source\code\root\folder>python setup,py install
```

Option 2: Install using Python Setup Tools

In addition to making the source code available on GitHub in a zip file, PopGen is also distributed on the Python Package Index. In order to install PopGen from the Python Package Index, issue the following command from the shell:

```
C:\default\windows\shell>pip install popgen
```

Note: This option is currently not available because a stable release of PopGen 2.0 is yet to be released. PopGen is now in an open beta release. Soon after testing is complete on the beta

release, a stable release of PopGen will be made available along with a distribution of the package on Python Package Index.

APPENDIX B: INSTRUCTIONS FOR RUNNING POPGEN 2.0

In this appendix, instructions for setting and running a project in PopGen 2.0 are provided. In the first section, instructions are provided for defining the configuration file. In the second section, instructions for preparing the input files are discussed. In the last section, steps for launching a synthetic population run are presented.

Note: From this point forward PopGen 2.0 will be referred to as just PopGen – any reference to the word PopGen in the remaining Appendix should be associated with the version 2.0.

Defining the Configuration File

PopGen uses a YAML-based configuration file setup to specify a project. YAML files can be created/edited using any source code editor (e.g. Notepad++ or Atom). An example of the configuration file can be accessed by visiting the root folder of the PopGen installation – this is located under <Python27_Folder>\Lib\site-packages\popgen-2.0b1-py2.7.egg. The file can also be accessed by visiting the root folder of the PopGen source code that you downloaded when installing PopGen. Once you have opened the root folder, the file is located under **tutorials\1_basic_popgen_setup\configuration.yaml**.

The configuration file comprises of a series of key-value pairs that specify information about the inputs, population synthesis methods, and outputs to be generated. There are three types of keys in the configuration file namely **required**, **project specific, and expressive**. **Required** keys must be specified for every project and they contain important information to run PopGen. **Project specific** keys are unique to a project and specified only if the project warrants their inclusion. Lastly, **expressive keys** are used to structure the configuration file. The names of the required keys and expressive keys are always fixed and the names of the project specific keys depend on the values provided for required keys that appear earlier in the configuration file.

Values for both required and project specific keys are compulsory. Currently values can be of 5 types, namely, Boolean, numbers, strings, list of strings, or list of numbers. Expressive keys don't accept any values directly – required keys and project specific keys are generally embedded under an expressive key. It must be noted that there is no special formatting when specifying values for keys. Numbers and strings can be expressed without any formatting (e.g. double quotations for string). List of string values (list of numbers) is defined by providing multiple strings (numbers) separated by a comma all of which is enclosed in square brackets (e.g. of list of string [variable1, variabl2]).

When setting up a new project, users can just edit the configuration file located in the tutorial instead of creating a configuration file from scratch – all the required keys and expressive keys should remain the same and only the values and project specific keys need to be revised to create a new configuration file.

The YAML-based configuration file can be divided into four sections. Each of the sections in the configuration file is described in detail below. For each section, an extract of the configuration file is presented followed by a discussion of the various configuration elements. Line numbers are included in the figures and tables below just for referencing the configuration elements and need not be included when setting up a new project. Required keys are presented in red color,

project specific keys are shown in green color, and expressive keys are presented in purple color in the extracts of the configuration file. Values for keys where applicable are presented in black color.

Project Attributes

Figure B.1 shows an extract of the project attributes from an example configuration file. In this portion of the configuration file, some general information regarding the PopGen project is specified. Particular information that needs to be specified by the user for various required and project specific keys are shown in Table B.1. Table B.1 also lists the description of expressive keys where a value is not expected.

```
1  project:
2   name: example
3   location: ../tutorials/1_basic_popgen_setup/
```

Figure B.1: Extract of an example configuration file showing the Project Attributes section

Table B.1: Description of keys and expected values in the Project Attributes section of the configuration file

Line No.	Key Name	Description	Value Type
1	project	Indicates the start of a PopGen project	-
2	name	Name/description of the project	Alphanumeric
3	location	Provides a location where all the inputs files are located. This also serves as the folder where outputs are stored. A valid folder path must be specified. If a folder doesn't exist, the program will create one	String

Input Files Configuration

Figure B.2 shows an extract of the input files configuration. In this portion of the configuration file, information regarding the input files is specified. Particular information that needs to be specified by the user for various required and project specific keys related to Input Files are shown in Table B.2. Table B.2 also lists the description of expressive keys where a value is not expected.

```
4   inputs:
5    entities: [household, person]
6    housing_entities: [household]
7    person_entities: [person]
8    column_names:
9     hid: hid
10    pid: pid
11    geo: geo
12    region: region
13    sample_geo: sample_geo
14   location:
15    geo_corr_mapping:
16     geo_to_sample: geo_sample_mapping.csv
17     region_to_sample: region_sample_mapping.csv
18     region_to_geo: region_geo_mapping.csv
19    sample:
20     household: household_sample.csv
```

```
21      person: person_sample.csv
22    marginals:
23     region:
24      household: region_household_marginals.csv
25      person: region_person_marginals.csv
26     geo:
27      household: household_marginals.csv
28      person: person_marginals.csv
```

Figure B.2: Extract of an example configuration file showing the Input Files section

Table B.2: Description of keys and expected values in the Input Files section of the configuration file

Line No.	Key Name	Description	Value Type
4	inputs	Indicates the start of the configuration element where input files are defined	-
5	entities	Specifies the types of housing and person population units that will be synthesized	List of strings
6	housing_entities	Specifies which of the "entities" correspond to housing population units. Users must ensure that the values provided here are a subset of the values provided in entities	List of strings
7	housing_entities	Specifies which of the "entities" correspond to person population units. Users must ensure that the values provided here are a subset of the values provided in entities	List of strings
8	column_names	Indicates the start of the configuration element where information regarding column names for key variables of interest is provided	-
9	hid	Name of the column in the sample files for housing entities that uniquely identifies the sample housing unit	String
10	pid	Name of the column in the sample files for person entities that represents the ID of a person within a sample housing unit	String
11	geo	Name of the column in the marginal files and the geographic correspondence files that contains IDs for the lower level spatial units	String
12	region	Name of the column in the marginal files and the geographic correspondence files that contains IDs for the higher level spatial units	String
13	sample_geo	Name of the column in the sample files and the geographic correspondence files that contains values for the spatial unit at which sample data is available.	String
14	location	Indicates the start of the configuration element inside "inputs" where file names will be specified	-
15	geo_corr_mapping	Indicates the start of the configuration element inside "location" where file names	-

Line No.	Key Name	Description	Value Type
		for geographic correspondence files will be specified	
16	geo_to_sample	Name of the file containing geographic unit to sample correspondence	String
17	region_to_sample	Name of the file containing region to sample correspondence	String
18	region_to_geo	Name of the file containing region to geographic unit correspondence	String
19	sample	Indicates the start of the configuration element inside “location” where file names for sample files will be provided	-
20	household (project specific key under “sample”)	Name of the file containing the sample data for household – a housing entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	String
21	person (project specific key under “sample”)	Name of the file containing the sample data for person – a person entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	String
22	marginals	Indicates the start of the configuration element inside “location” where file names of marginal files will be provided	-
23	region	Indicates the start of the configuration element inside “marginals” where file names of marginal files for higher level spatial units will be provided	-
24	household (project specific key under “marginal” for “region”)	Name of the file containing the marginal data for higher level spatial units for household – a housing entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	String
25	person (project specific key under “marginal” for “region”)	Name of the file containing the marginal data for higher level spatial units for person – a person entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	String
26	geo	Indicates the start of the configuration element inside “marginals” where file name for marginal files for lower level spatial units will be provided	-
27	household (project specific key under “marginal” for “geo”)	Name of the file containing the marginal data for lower level spatial units for household – a housing entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	String
28	person (project specific key under	Name of the file containing the marginal data for lower level spatial units for	String

Line No.	Key Name	Description	Value Type
	"marginal" for "geo")	person – a person entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	

Scenario Specification: Control Variables and Parameters

Figure B.3 shows an extract of the scenario specification where information regarding the controls and parameters are provided. Particular information that needs to be specified by the user for various required and project specific keys are shown in Table B.3. Table B.3 also lists the description of expressive keys where a value is not expected.

```
29  scenario:
30   - description: all_controls_entropy
31     control_variables:
32      region:
33       household: [rhhldtype]
34       person: []
35      geo:
36       household: [hhldtype]
37       person: [ptype]
38     parameters:
39      ipf:
40       tolerance: 0.0001
41       iterations: 250
42       zero_marginal_correction: 0.00001
43       rounding_procedure: bucket
44       archive_performance_frequency: 1
45      reweighting:
46       procedure: entropy
47       tolerance: 0.0001
48       inner_iterations: 1
49       outer_iterations: 1000
50       archive_performance_frequency: 1
51      draws:
52       pvalue_tolerance: 0.9999
53       iterations: 25
54       seed: 0
55     geos_to_synthesize:
56      region:
57       ids: [1]
```

Figure B.3: Extract of an example configuration file showing the Control Variables and Parameters section for a given scenario

Table B.3: Description of keys and expected values in the Control Variables and Parameters section of the configuration file

Line No.	Key Name	Description	Value Type
29	scenario	Indicates the start of the configuration element inside a "project" where information regarding scenarios are provided	-
30	description	Any user specified label for identifying the	String

Line No.	Key Name	Description	Value Type
		scenario being prepared	
31	control_variables	Indicates the start of the configuration element inside a scenario where the variables to be controlled are provided	-
32	region	Indicates the start of the configuration element inside “control_variables” where variables to be controlled at the higher level spatial resolution are provided	-
33	household (project specific key under “control_variables” for “region”)	List of variables to be controlled at higher level spatial unit of analysis for household – a housing entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	List of strings
34	person (project specific key under “control_variables” for “region”)	List of variables to be controlled at higher level spatial unit of analysis for person – a person entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	List of strings
35	geo	Indicates the start of the configuration element inside “control_variables” where variables to be controlled at the lower level spatial resolution are provided	-
36	household (project specific key under “control_variables” for “geo”)	List of variables to be controlled at lower level spatial unit of analysis for household – a housing entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	List of Strings
37	person (project specific key under “control_variables” for “geo”)	List of variables to be controlled at lower level spatial unit of analysis for person – a person entity. Note that this project specific key was derived from the values of the entities provided in the example configuration file	List of Strings
38	parameters	Indicates the start of the configuration element inside a scenario where parameters for the run are specified	-
39	ipf	Indicates the start of the configuration element inside “parameters” where IPF related parameters are provided	-
40	tolerance (required key under “ipf”)	Threshold value to check for convergence in the IPF procedure	Number
41	Iterations (required key under “ipf”)	Maximum number of iterations to be performed in the IPF procedure if convergence is not achieved	Number
42	zero_marginal_correction (required key under “ipf”)	Value to be assigned to zero marginals when performing the IPF procedure	Number
43	rounding_procedure (required key under “ipf”)	Type of rounding procedure to be applied to convert the IPF estimated cell values to integers. Currently only bucket rounding	String (valid

Line No.	Key Name	Description	Value Type
		procedure is implemented	value – bucket)
44	archive_performance_ frequency (required key under “ipf”)	This specifies the frequency at which the performance measures for the IPF procedure should be stored	Number
45	reweighting	Indicates the start of the configuration element inside “parameters” where parameters for sample weight estimation procedure are provided	-
46	procedure (required key under “reweighting”)	Type of procedure to be applied to estimate the sample weights. Currently IPU and entropy-based procedures are supported	String (valid values – ipu, entropy)
47	tolerance (required key under “reweighting”)	Threshold value to check for convergence in the reweighting procedure	Number
48	inner_iterations (required key under “reweighting”)	Maximum number of outer iterations to be performed in the reweighting procedure	Number
49	outer_iterations (required key under “reweighting”)	Maximum number of inner iterations to be performed for each outer iteration in the reweighting procedure	Number
50	archive_performance_ frequency (required key under “reweighting”)	This specifies the frequency at which the performance measures from the weighting procedure must be archived. In the weighting step, constant archiving of the performance measures will affect runtimes. Therefore, caution must be exercised when specifying the frequency value – low values for this key mean additional runtimes	Number
51	draws	Indicates the start of the configuration element inside “parameters” where parameters for synthetic population drawing procedure are provided	-
52	pvalue_tolerance (required key under “draws”)	Threshold value to satisfied by the drawing step to stop the iterative procedure for selecting synthetic population	Number
53	iterations (required key under “draws”)	Maximum number of iterations to be performed in the drawing step if tolerance is not satisfied	Number
54	seed (required key under “draws”)	Value of the seed to use in the random sampling involved in the drawing step	Number
55	geos_to_synthesize	Indicates the start of the configuration element inside a scenario for limiting the synthesis to selected geographies	-
56	region	Indicates the start of the configuration element inside “geos_to_synthesize” where select higher level geographies can be specified	-
57	ids (required key under	Provides the list of higher level geographies for which synthetic population should be	List of Numbers

Line No.	Key Name	Description	Value Type
	"geos_to_synthesize" for "region")	generated	

Scenario Specification: Outputs

Figure B.4 shows an extract of the scenario configuration where the specification for generating outputs is provided. Particular information that needs to be specified by the user for various required and project specific keys for generating outputs are shown in Table B.2. Table B.2 also lists the description of expressive keys where a value is not expected.

```
58     outputs:
59      performance: [ipf, reweighting, drawing]
60      weights:
61       export: True
62       collate_across_geos: False
63      multiway:
64       - variables: [ptype]
65         filename: ptype.csv
66         filetype: csv
67         entity: person
68       - variables: [hhldtype]
69         filename: hhldtype.csv
70         filetype: csv
71         entity: household
72       - variables: [rhhldtype]
73         filename: rhhldtype.csv
74         filetype: csv
75         entity: household
76      summary:
77       region:
78        filename: summary_region.csv
79        filetype: csv
80       geo:
81        filename: summary_geo.csv
82        filetype: csv
83      synthetic_population:
84       housing:
85        filename: housing_synthetic.csv
86        filetype: csv
87       person:
88        filename: person_synthetic.csv
89        filetype: csv
```

Figure B.4: Extract of an example configuration file showing the Outputs section for a given scenario

Table B.4: Description of keys and expected values in the Outputs section of the configuration file

Line No.	Key Name	Description	Value Type
58	outputs	Indicates the start of the configuration element inside a scenario where information for generating outputs is provided	-
59	performance	Specify the types of performance measures to be generated	List of Strings - (valid

Line No.	Key Name	Description	Value Type
			values - ipf, reweighting, drawing)
60	weights	Indicates the start of the configuration element inside "outputs" where information for writing out sample weights is provided	-
61	export	Specify whether estimated sample weights need to be written out	Boolean (valid values - True, False)
62	collate_across_geos	Indicates whether the sample weights for each household should be aggregated across geographies	Boolean (valid values - True, False)
63	multiway	Indicates the start of the configuration element inside "outputs" where information for generating multiway cross tabulations is provided	-
64, 68, 72	variables (required key under "multiway" for generating a multiway table)	For a given cross tabulation, this provides the list of variables to consider for producing a cross tabulation from the synthetic population	List of Strings
65, 69, 73	filename (required key under "multiway" for generating a multiway table)	For a given cross tabulation, this provides the filename for storing the cross tabulation	String
66, 70, 74	filetype (required key under "multiway" for generating a multiway table)	For a given cross tabulation, this provides the file type for storing the cross tabulation. Only comma-separated value format is supported at this time	String (valid value - csv)
67, 71, 75	entity (required key under "multiway" for generating a multiway table)	For a given cross tabulation, this provides the entity type from which to generate a cross tabulation	String (valid values - housing, person)
76	summary	Indicates the start of the configuration element inside "outputs" where information for generating summary tabulations is provided	-
77	region	Indicates the start of the configuration element inside "summary" where information for generating summary tabulations at the higher level spatial resolution is provided	-
78	filename (required key under "summary" for "region")	This provides the filename for storing the summary of the synthetic population at the higher level of spatial resolution	String
79	filetype (required key under "summary" for "region")	This provides the file type for storing the summary at the higher level spatial resolution. Only comma-separated value format	String (valid value -

Line No.	Key Name	Description	Value Type
		is supported at this time	csv)
80	geo	Indicates the start of the configuration element inside “summary” where information for generating summary tabulations at the lower level spatial resolution is provided	-
81	filename (required key under “summary” for “geo”)	This provides the filename for storing the summary of the synthetic population at the lower level of spatial resolution	String
82	filetype (required key under “summary” for “geo”)	This provides the file type for storing the summary at the lower level spatial resolution. Only comma-separated value format is supported at this time	String (valid value - csv)
83	synthetic_population	Indicates the start of the configuration element inside “outputs” where information for exporting synthetic population files is provided	-
84	housing	Indicates the start of the configuration element inside “synthetic_population” where information for exporting synthetic population records for the housing entities is provided	-
85	filename (required key under “synthetic_population” for “housing”)	This provides the filename for storing the synthetic population for the housing entities	String
86	filetype (required key under “synthetic_population” for “housing”)	This provides the file type for storing the synthetic population for the housing entities	String (valid values - csv)
87	person	Indicates the start of the configuration element inside “synthetic_population” where information for exporting synthetic population records for the person entities is provided	-
88	filename (required key under “synthetic_population” for “person”)	This provides the filename for storing the synthetic population for the person entities	String
89	filetype (required key under “synthetic_population” for “person”)	This provides the file type for storing the synthetic population for the person entities	String (valid values - csv)

Preparing the Input Files

The structure and layout of the different input files are described in this section. For each input file, there are two types of columns namely required columns, and optional columns. Required columns contain information that is needed to run PopGen. Therefore, all required columns must definitely be included in the input files for PopGen to proceed. It must be noted that required columns need not have the same name as shown in the example table layouts below. In the table layouts for the different input files, only required columns are presented because optional columns are not always necessary to run PopGen. It is important that the order of the required

columns as shown in the layouts below be maintained for the different input files to avoid any undesirable behaviors. A codebook is provided at the end of this subsection that provides a description of all the required columns in each of the input files.

In each input file, the first row should contain the variable names. From second row and onwards, column values must be included. Users must note that the only exception to this layout of the column names and values is for the marginal input files – see the discussion on Marginal Files for more information. Use only alphanumeric column names – do not use any special characters other than underscore symbol (_) when defining the column names. All input files must be specified in a comma-separated values (CSV) format.

Three types of input files are needed to setup a project in PopGen including Geographic Correspondence Files, Sample Files, and Marginal Files. For each of these types of inputs, specific files and their layouts are described below:

Geographic Correspondence Files

There are three types of spatial units/resolutions that are used in PopGen, namely, **geographic unit**, **region**, and **sample geographic unit**. **Geographic unit** is used to reference the lower level spatial resolution at which population synthesis will be performed. Next, **region** is used to reference the higher level spatial resolution at which control marginal distributions are provided. It must be noted that these marginal distributions are in addition to the marginal distributions typically provided for each geographic unit during population synthesis. Lastly, **sample geographic unit** is used to reference the spatial resolution corresponding to the sample units. Three types of geographic correspondence files are utilized by PopGen to gather information regarding these three spatial units/resolutions as described below:

- Region to geographic unit correspondence – This file provides the mapping between higher level spatial unit and lower level spatial unit at which controls will be provided. The structure of the file is shown in Table B.5 below.

  Table B.5: Layout of the region to geographic unit correspondence file

region	geo

- Region to sample correspondence – This file provides the mapping between region and sample geographic unit. The layout of the file is shown in Table B.6 below.

  Table B.6: Layout of the region to sample correspondence file

region	sample_geo

- Geographic unit to sample correspondence – This file provides the mapping between the geographic unit of analysis and the sample geographic unit. The structure of the file is shown in Table B.7.

  Table B.7: Layout of the geographic unit to sample correspondence file

geo	sample_geo

Note: In the region to sample correspondence (geographic unit to sample correspondence), if multiple entries are provided for each region (geo) then sample units from those sample geographies are used as seed during the IPF procedure

Sample Files

As the name suggests, these files provide sample data for the housing and person entities to PopGen. The number of sample files will depend on the number of housing and person entities specified by the user. For each of the entities a separate sample file must be provided. The structure of the input files for housing and person entities is shown in Table B.8 and B.9 respectively.

Table B.8: Layout of the sample file for each housing entity

hid	sample_geo	attribute_1	attribute_2	…	attribute_k	…		

Table B.9: Layout of the sample file for each person entity

hid	pid	sample_geo	attribute_1	attribute_2	…	attribute_k	…		

Marginal Files

These files provide the marginal distributions that must be controlled during the population synthesis procedure for each of the housing and person entities. Similar to the sample files, the number of marginal files will depend on the number of housing and person entities being synthesized. Additionally, for each entity two files must be provided – first file provides the marginal distribution for the entity at the geographic unit level (lower level spatial resolution) and the second file provides the marginal distribution for the entity at the region level (higher level spatial resolution). The structure of the input files is similar for both housing and person entities and is as shown in Table B.10 respectively.

Table B.10: Layout of the marginal file for both housing and person entities

variable_names	attribute_1	attribute_1	…	attribute_1	…	attribute_k	…
variable_categories	value_1	value_2	…	value_p	…	value_1	…
geo							

It must be noted that the layout of the marginal files is different from the other input files. Differences include the following:

- First three rows provide the column definitions in the marginal files whereas only the first row provides the column definitions for the geographic correspondence and sample files
- Data values for the different columns start from the fourth row

Definitions of the Required Columns

Table B.11 below provides the definitions of each of the required columns. As mentioned earlier, the required columns are necessary to run PopGen. However, they need not have the same names as shown in the layout tables above.

Table B.11: Definitions of the required columns in the input files

Column Name	Column Definition
region	The column provides IDs of the higher level spatial units at which marginal distributions are controlled
geo	The column provides IDs of the lower level spatial units at which marginal distributions are controlled
sample_geo	The column provides IDs of the spatial units at which sample data is available
hid	This column provides a unique ID to each sample unit for each housing entity that is synthesized
pid	This a value indexed at 1 that provides a unique ID to every person in the sample housing unit
attribute_k	For a given entity, this represents the k^{th} variable name
variable_names	This is just a row description that informs PopGen that the first row in the marginal files corresponds to the names of the control variables
variable_categories	This is just a row description that informs PopGen that the second row in the marginal files correspond to the category for each control variable
value_p	This is a numeric value for the p^{th} category of a given control variable

Launching a PopGen Run

In this section, steps for launching a population synthesis run in PopGen are described. First, prepare the configuration file and all the input files for the specific application as described in the previous two sections. After that PopGen can be launched by preparing a batch file as described below:

Using PopGen Run Script

When PopGen is installed using the approach described in Appendix A, it installs a run script titled popgen_run. This script can be accessed from the command line and be used to launch a PopGen run. The script requires only one input namely location of the configuration file. It is executed from the command line as shown below (highlighted in blue):

```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Corporation.  All rights reserved.
```

Using Python Scripting

When PopGen is installed, it is also setup as a Python library that can be called from any Python script. Therefore, in addition to the run script approach described above, PopGen can also be called from a Python script. The steps to launch PopGen using Python scripting is described below:

Note: The steps outlined here are specific to running the example in the tutorial folder. However, one can easily modify these instructions to fit any general use case. The commands to be issued in each step are highlighted in blue.

Step 1: Launch the Windows command line and issue the change directory command to switch to the root folder where you extracted the PopGen source code

```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Corporation.  All rights reserved.
```

Step 2: Launch the Python shell by issuing the following command from the windows command line

```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Corporation.  All rights reserved.
```

This will launch a Python shell as shown below with a blinking cursor next to >>>.

```
Microsoft Windows [Version 6.1.7601]

Copyright (c) 2009 Microsoft Corporation.  All rights reserved.


C:\popgen\source\code\root\folder>python

Python 2.7.10 (default, May 23 2015, 09:44:00) [MSC v.1500 64 bit (AMD64)]
on win32
```

Step 2: Import PopGen Project class by issuing the command from the Python shell

```
C:\Users\kkonduri>python

Python 2.7.10 (default, May 23 2015, 09:44:00) [MSC v.1500 64 bit (AMD64)]
on win32

Type "help", "copyright", "credits" or "license" for more information.
```

Step 3: Create a new Project class object by issuing the below command

```
>>>from popgen import Project

>>>p = Project(“./tutorial/1_basic_popgen_setup/configuration.yaml”)
```

Step 4: Load the project by calling the load_project function

```
>>>from popgen import Project

>>>p = Project(“./tutorial/1_basic_popgen_setup/configuration.yaml”)
```

Step 5: Run the scenarios in the configuration file by calling the run_scenarios function

```
>>>from popgen import Project

>>>p = Project(“./tutorial/1_basic_popgen_setup/configuration.yaml”)

>>>p.load_project()
```

Above runs the synthesis procedure and synthetic population results should be available for further use in the ./tutorial/1_basic_popgen_setup/<Data Time all_controls_ipu>.