Exploring the Implications of Alternative Household- and Person-level Constraints at Multiple Spatial Resolutions in Synthetic Population Generation

Karthik C. Konduri
University of Connecticut

Daehyun You, Venu M. Garikapati, and Ram M. Pendyala
Georgia Institute of Technology

Innovations in Travel Modeling Conference, Denver, Colorado. May 1-4, 2016

Outline

Introduction

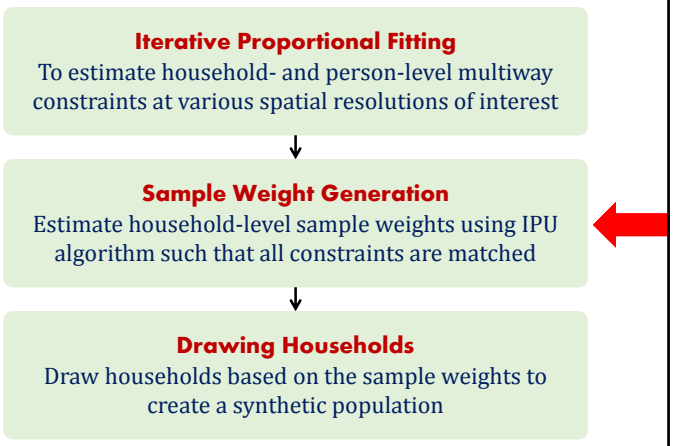
- Disaggregate socio-economic and demographic data for the entire population is a critical input to apply microsimulation models
  - This information is not readily available
- Instead, disaggregate information for a sample of households ("sample data"), and aggregate distributions ("marginal distributions") for the entire population are available
  - Available data are combined to synthesize a population
- Beckman et al (1996) was the earliest synthetic population generator (SPG)
  - Number of SPGs have been developed since
  - Ma and Srinivasan (2015) provide a comprehensive review of generators to date

Multi-level Marginal Distributions

- Agencies often maintain marginal distributions at multiple spatial resolutions
  - For example, control distributions for select variables at the traffic analysis zone (TAZ) level, and additional control distributions for other variables at the county level
- It would be desirable to generate a synthetic population that accounts for all of the information available
- Most of the SPGs in practice today are not able to accommodate multi-level distributions of household and person-level attributes as controls
- An enhanced SPG (PopGen 2.0) that can accommodate multi-level marginal distributions as controls

Overview of PopGen 1.0

- PopGen 1.0 (Ye et al 2009) is able to match given distributions of household- and person-level attributes
- In PopGen 2.0, sample weight generation step is enhanced to accommodate multi-level marginal distributions



Iterative Proportional Updating (IPU): Overview

- The IPU is a heuristic iterative procedure that solves the below optimization problem for estimating household sample weights so that the given constraints are matched

Minimize $\sum_j \left[\frac{\sum_i d_{i,j} w_i - c_j}{c_j}\right]^2$

Subject to $w_i \geq 0$

Where:
i denotes a sample household
j denotes the constraint or population characteristic of interest
$d_{i,j}$ represents the contribution of the sample household i to the population characteristic j
w_i is the weight attributed to sample household i
c_j is the value of the population characteristic j

Iterative Proportional Updating (IPU): Steps

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3
1	1	1	0	1	1	1
2	1	1	0	1	0	1
3	1	1	0	2	1	0
4	1	0	1	1	0	2
5	1	0	1	0	2	1
6	1	0	1	1	1	0
7	1	0	1	2	1	2
8	1	0	1	1	1	0
Weighted Sum		3.00	5.00	9.00	7.00	7.00
Constraints		35.00	65.00	91.00	65.00	104.00
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327

Iterative Proportional Updating (IPU): Steps (2)

Iteration 1: Adjustment with respect to Household Type 1 constraint

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weights 1
1	1	1	0	1	1	1	11.67
2	1	1	0	1	0	1	11.67
3	1	1	0	2	1	0	11.67
4	1	0	1	1	0	2	1.00
5	1	0	1	0	2	1	1.00
6	1	0	1	1	1	0	1.00
7	1	0	1	2	1	2	1.00
8	1	0	1	1	1	0	1.00
Weighted Sum		3.00	5.00	9.00	7.00	7.00	
Constraints		35.00	65.00	91.00	65.00	104.00	
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327	
Weighted Sum 1		***35.00***	5.00	51.67	28.33	28.33	

Iterative Proportional Updating (IPU): Steps (3)

Iteration 1: Adjustment with respect to Household Type 2 constraint

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weights 1	Weights 2
1	1	1	0	1	1	1	11.67	11.67
2	1	1	0	1	0	1	11.67	11.67
3	1	1	0	2	1	0	11.67	11.67
4	1	0	1	1	0	2	1.00	13.00
5	1	0	1	0	2	1	1.00	13.00
6	1	0	1	1	1	0	1.00	13.00
7	1	0	1	2	1	2	1.00	13.00
8	1	0	1	1	1	0	1.00	13.00
Weighted Sum		3.00	5.00	9.00	7.00	7.00		
Constraints		35.00	65.00	91.00	65.00	104.00		
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327		
Weighted Sum 1		***35.00***	5.00	51.67	28.33	28.33		
Weighted Sum 2		35.00	***65.00***	111.67	88.33	88.33		

Iterative Proportional Updating (IPU): Steps (4)

Iteration 1: Adjustment with respect to Person Type 1 constraint

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weights 1	Weights 2	Weights 3
1	1	1	0	1	1	1	11.67	11.67	9.51
2	1	1	0	1	0	1	11.67	11.67	9.51
3	1	1	0	2	1	0	11.67	11.67	9.51
4	1	0	1	1	0	2	1.00	13.00	10.59
5	1	0	1	0	2	1	1.00	13.00	13.00
6	1	0	1	1	1	0	1.00	13.00	10.59
7	1	0	1	2	1	2	1.00	13.00	10.59
8	1	0	1	1	1	0	1.00	13.00	10.59
Weighted Sum		3.00	5.00	9.00	7.00	7.00			
Constraints		35.00	65.00	91.00	65.00	104.00			
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327			
Weighted Sum 1		***35.00***	5.00	51.67	28.33	28.33			
Weighted Sum 2		35.00	***65.00***	111.67	88.33	88.33			
Weighted Sum 3		28.52	55.38	***91.00***	76.80	74.39			

Iterative Proportional Updating (IPU): Steps (5)

Iteration 1: Adjustment with respect to Person Type 2 constraint

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weights 1	Weights 2	Weights 3	Weights 4
1	1	1	0	1	1	1	11.67	11.67	9.51	8.05
2	1	1	0	1	0	1	11.67	11.67	9.51	9.51
3	1	1	0	2	1	0	11.67	11.67	9.51	8.05
4	1	0	1	1	0	2	1.00	13.00	10.59	10.59
5	1	0	1	0	2	1	1.00	13.00	13.00	11.00
6	1	0	1	1	1	0	1.00	13.00	10.59	8.97
7	1	0	1	2	1	2	1.00	13.00	10.59	8.97
8	1	0	1	1	1	0	1.00	13.00	10.59	8.97
Weighted Sum		3.00	5.00	9.00	7.00	7.00				
Constraints		35.00	65.00	91.00	65.00	104.00				
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327				
Weighted Sum 1		***35.00***	5.00	51.67	28.33	28.33				
Weighted Sum 2		35.00	***65.00***	111.67	88.33	88.33				
Weighted Sum 3		28.52	55.38	***91.00***	76.80	74.39				
Weighted Sum 4		25.60	48.50	80.11	***65.00***	67.68				

Iterative Proportional Updating (IPU): Steps (6)

Iteration 1: Adjustment with respect to Person Type 3 constraint

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Weights 1	Weights 2	Weights 3	Weights 4	Weights 5
1	1	1	0	1	1	1	11.67	11.67	9.51	8.05	12.37
2	1	1	0	1	0	1	11.67	11.67	9.51	9.51	14.61
3	1	1	0	2	1	0	11.67	11.67	9.51	8.05	8.05
4	1	0	1	1	0	2	1.00	13.00	10.59	10.59	16.28
5	1	0	1	0	2	1	1.00	13.00	13.00	11.00	16.91
6	1	0	1	1	1	0	1.00	13.00	10.59	8.97	8.97
7	1	0	1	2	1	2	1.00	13.00	10.59	8.97	13.78
8	1	0	1	1	1	0	1.00	13.00	10.59	8.97	8.97
Weighted Sum		3.00	5.00	9.00	7.00	7.00					
Constraints		35.00	65.00	91.00	65.00	104.00					
δ_b		0.9143	0.9231	0.9011	0.8923	0.9327					
Weighted Sum 1		***35.00***	5.00	51.67	28.33	28.33					
Weighted Sum 2		35.00	***65.00***	111.67	88.33	88.33					
Weighted Sum 3		28.52	55.38	***91.00***	76.80	74.39					
Weighted Sum 4		25.60	48.50	80.11	***65.00***	67.68					
Weighted Sum 5		35.02	64.90	104.84	85.94	***104.00***					

Iterative Proportional Updating (IPU): Steps (7)

Weights at the end of 638 iterations

Household ID	Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Final Weights
1	1	1	0	1	1	1	1.36
2	1	1	0	1	0	1	25.66
3	1	1	0	2	1	0	7.98
4	1	0	1	1	0	2	27.79
5	1	0	1	0	2	1	18.45
6	1	0	1	1	1	0	8.64
7	1	0	1	2	1	2	1.47
8	1	0	1	1	1	0	8.64
Weighted Sum		35.00	65.00	91.00	65.00	104.00	
Constraints		35.00	65.00	91.00	65.00	104.00	
δ_b		0.0000	0.0000	0.0000	0.0000	0.0000	

PopGen 1.0: Multi-level Marginal Distributions

- Weights are estimated for each geographic unit independently such that household- and person-level constraints are satisfied
  - As a result, PopGen 1.0 is not capable of accounting multi-level controls
- In PopGen 2.0, the IPU procedure is extended so that weights are generated for all geographic units simultaneously such that multi-level marginal distributions are matched

Illustration of Extended IPU Methodology

- Consider a region with two geographies (Geo 1 and Geo 2)
- The first level of controls is at the region level (say a county) and second level of controls is at the geographic unit level (say a TAZ)

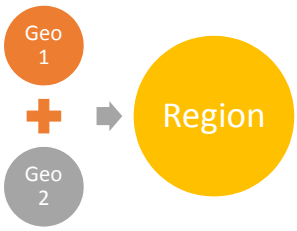
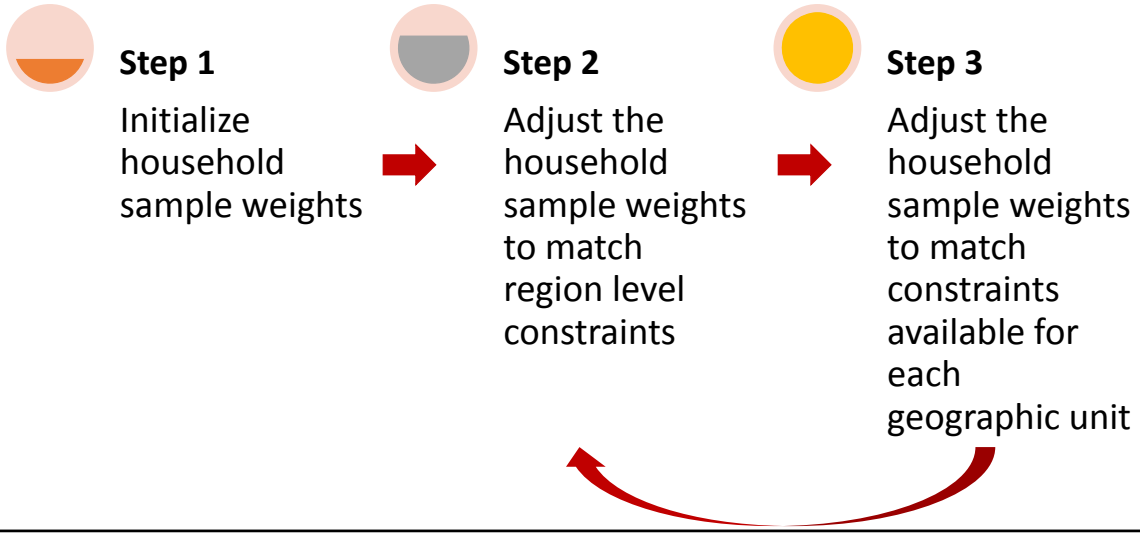


Illustration of Extended IPU Methodology (2)



Iteration 1, Step 1: Initial Household Sample Weight

			Region HH Type			HH Type		Person Type		
For first geographic unit (Geo 1)	hid	weight	1	2	3	1	2	1	2	3
	1	1	0	0	1	1	0	1	1	1
	2	1	1	0	0	1	0	1	0	1
	3	1	0	1	0	1	0	2	1	0
	4	1	1	0	0	0	1	1	0	2
	5	1	0	1	0	0	1	0	2	1
	6	1	0	0	1	0	1	1	1	0
	7	1	0	1	0	0	1	2	1	2
	8	1	0	0	1	0	1	1	2	0
		Geo 1			Weighted Sum	3	5	9	8	7
					Constraint	46	51	92	88	84
					δ	0.94	0.90	0.90	0.91	0.92
For second geographic unit (Geo 2)	hid	weight	1	2	3	1	2	1	2	3
	1	1	0	0	1	1	0	1	1	1
	2	1	1	0	0	1	0	1	0	1
	3	1	0	1	0	1	0	2	1	0
	4	1	1	0	0	0	1	1	0	2
	5	1	0	1	0	0	1	0	2	1
	6	1	0	0	1	0	1	1	1	0
	7	1	0	1	0	0	1	2	1	2
	8	1	0	0	1	0	1	1	2	0
		Geo 2			Weighted Sum	3	5	9	8	7
					Constraint	33	99	138	122	104
					δ	0.91	0.95	0.94	0.93	0.93
Match in constraints for Region		Weighted Sum	4.0	6.0	6.0					
		Constraint	86	61	82					
		δ	0.953	0.902	0.927					

Iteration 1, Step 2: Adjust weights to match region level constraints

			Region HH Type			HH 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
		Geo 1			Weighted Sum	45.33	69.17	124.67	95.33	108.67
					Constraint	46	51	92	88	84
					δ	0.01	0.36	0.36	0.08	0.29
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
		Geo 2			Weighted Sum	45.33	69.17	124.67	95.33	108.67
					Constraint	33	99	138	122	104
					δ	0.37	0.30	0.10	0.22	0.05
Match in constraints for Region		Weighted Sum	86.0	61.0	82.0					
		Constraint	86	61	82					
		δ	0.000	0.000	0.000					

Iteration 1, Step 3: Adjust weights to match geography level constraints

			Region HH Type			HH 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
	Geo 1				Weighted Sum	44.120	52.643	106.869	86.249	84.000
					Constraint	46	51	92	88	84
					δ	0.041	0.032	0.162	0.020	0.000
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
	Geo 2				Weighted Sum	27.844	87.550	122.800	110.679	104.000
					Constraint	33	99	138	122	104
					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	61	82					
		δ	0.216	0.019	0.035					

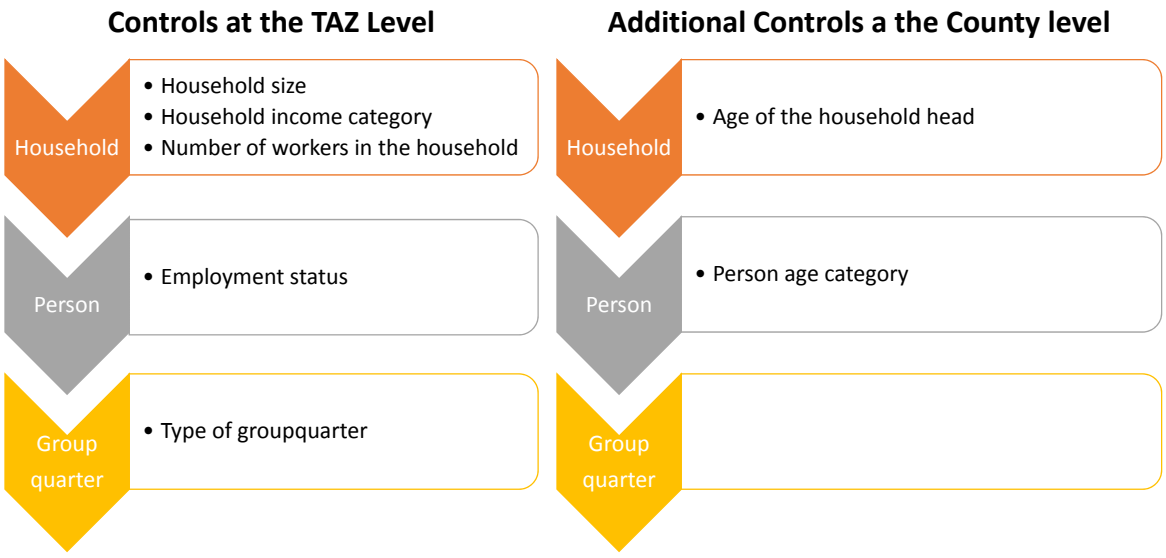
Adjusted weights at the end of 1000 Iterations

			Region HH Type			HH 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
	Geo 1				Weighted Sum	46.23	51.43	92.60	88.00	84.00
					Constraint	46	51	92	88	84
					δ	0.005	0.009	0.007	0.000	0.000
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
	Geo 2				Weighted Sum	33.17	99.77	139.00	122.00	104.00
					Constraint	33	99	138	122	104
					δ	0.005	0.008	0.007	0.000	0.000
Match in constraints for Region		Weighted Sum	86.0	61.7	82.9					
		Constraint	86.0	61.0	82.0					
		δ	0.000	0.011	0.011					

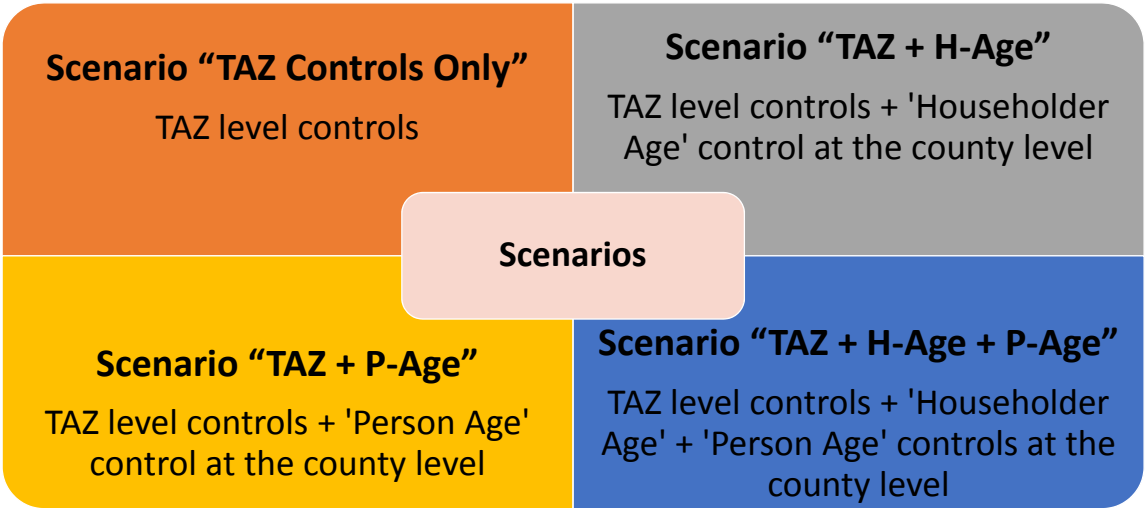
Case Study

- Model area: Baltimore Metropolitan Council (BMC) region
  - Households: 1,801,162
  - Persons: 4,793,980
  - Groupquarters: 104,522
- Marginal distributions are available at TAZ and County levels
- Generate a synthetic population utilizing the extended IPU procedure

Case Study: Control Variables



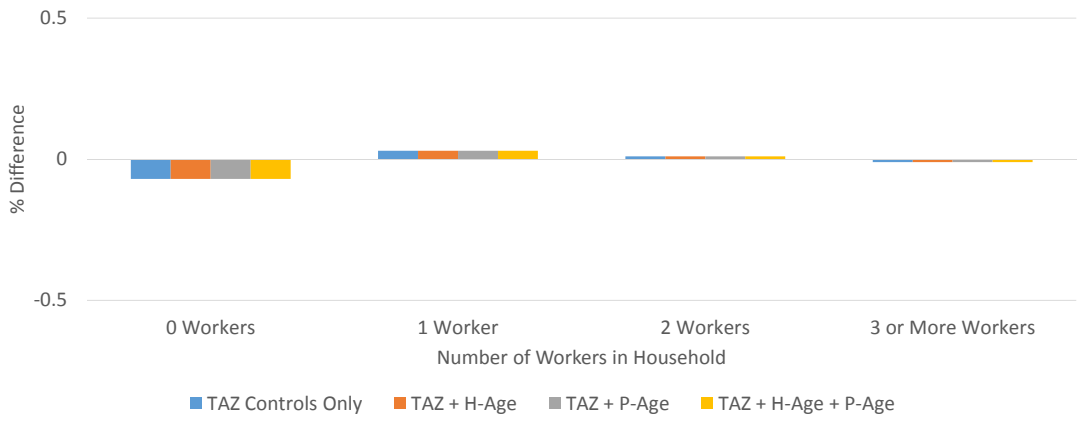
Case Study: Scenarios



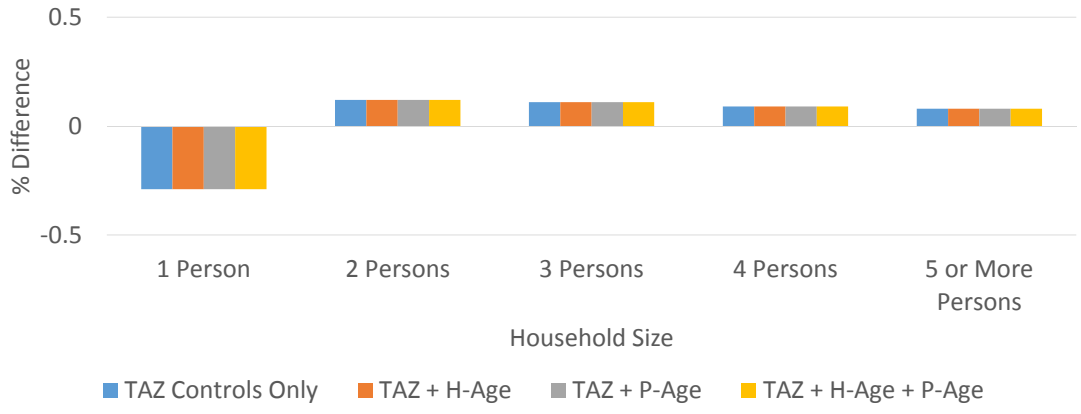
Results: Aggregate Comparisons

Scenario	Household Total			Person Total			Groupquarter Total		
	Given	Synthesized	% Diff	Given	Synthesized	% Diff	Given	Synthesized	% Diff
TAZ Controls Only	1801191	1801191	0.0%	4793980	4790075	-0.1%	104522	104522	0.0%
TAZ + H-Age	1801191	1801191	0.0%	4793980	4795380	0.0%	104522	104522	0.0%
TAZ + P-Age	1801191	1801191	0.0%	4793980	4722872	-1.5%	104522	104522	0.0%
TAZ + H-Age + P-Age	1801191	1801191	0.0%	4793980	4721624	-1.5%	104522	104522	0.0%

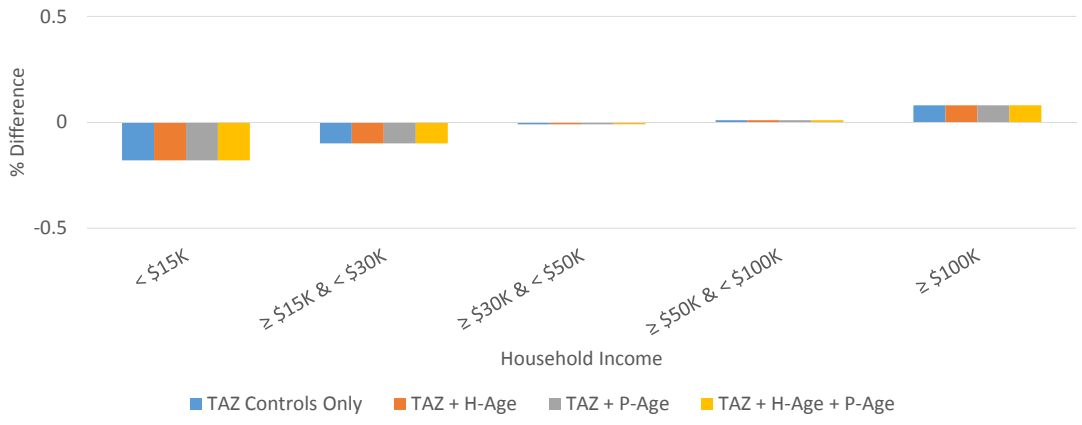
Results: Comparison of Worker Count Distribution



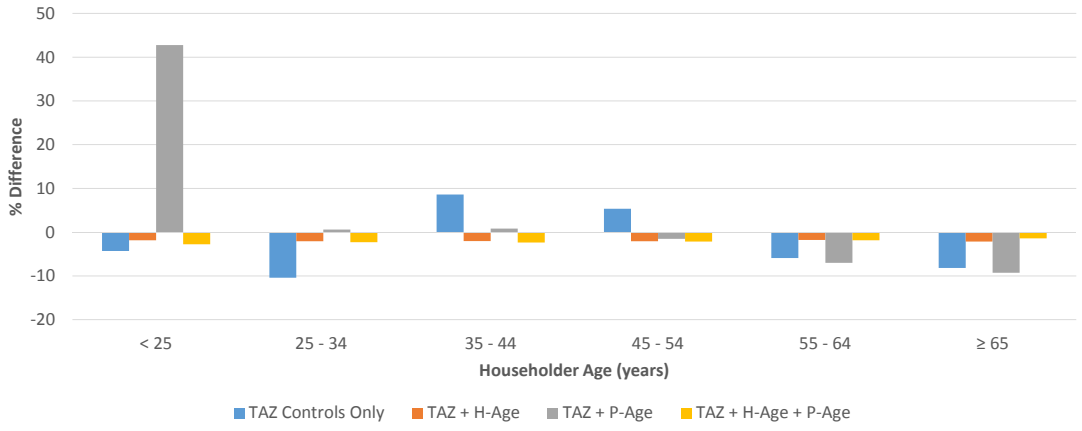
Results: Comparison of Household Size Distribution



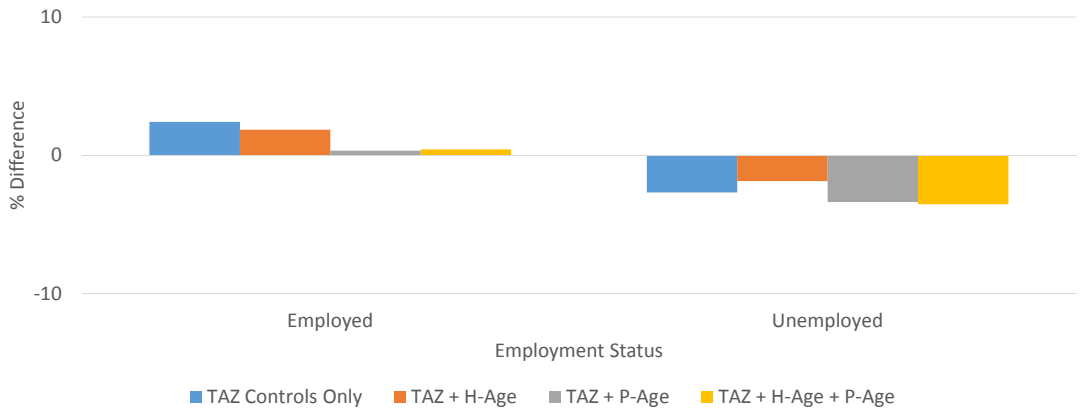
Results: Comparison of Household Income Distribution



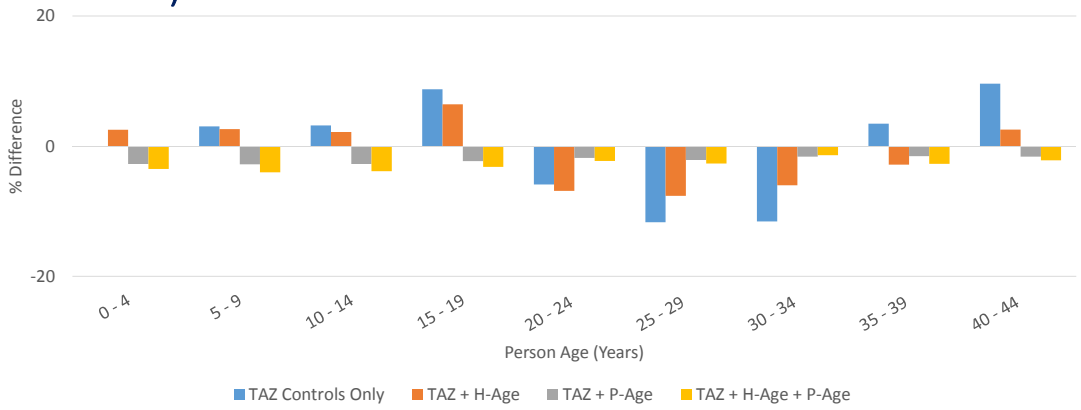
Results: Comparison of Householder Age Distribution
(county-level control)



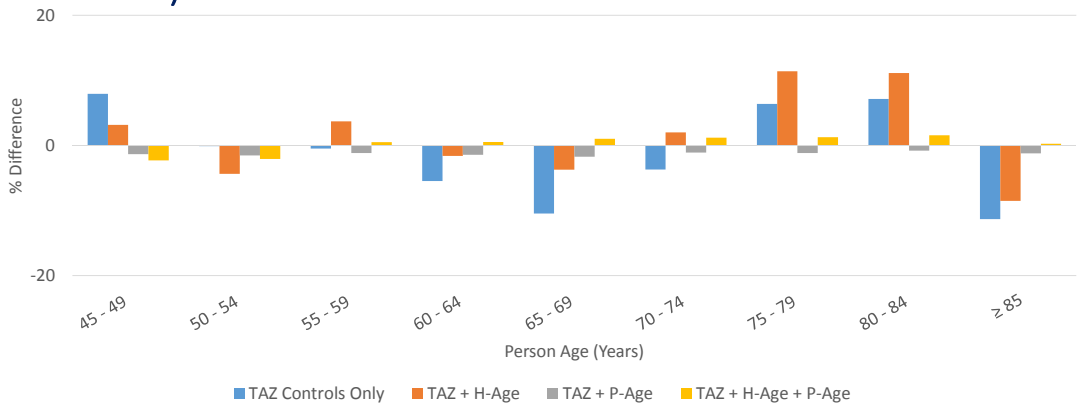
Results: Comparison of Employment Status Distribution



Results: Comparison of Person Age Distribution (county-level control)



Results: Comparison of Person Age Distribution (county-level control)



County Level Comparison of Fit Due to Additional Control: Householder Age

Scenario	County ID	Householder Age		Scenario	County ID	Householder Age	
		% Deviation Across Categories				% Deviation Across Categories	
		Min	Max			Min	Max
TAZ Controls Only	3	-10.08	9.50	TAZ + H-Age	3	-3.06	-2.01
	5	-10.03	9.11		5	-2.31	-1.14
	13	-12.68	58.81		13	0.13	4.32
TAZ + P-Age	3	-14.56	68.53	TAZ + H-Age + P-Age	3	-3.52	-1.37
	5	-8.67	43.22		5	-3.37	-1.32
	13	-9.19	161.83		13	1.33	4.37

County Level Comparison of Fit Due to Additional Control: Person Age

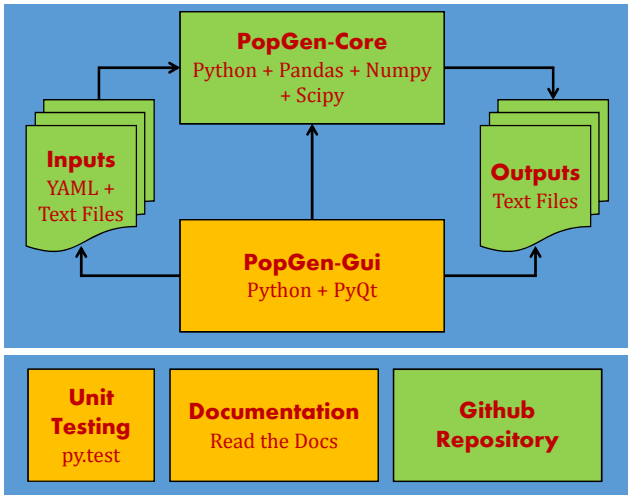
Scenario	County ID	Person Age		Scenario	County ID	Person Age	
		% Deviation Across Categories				% Deviation Across Categories	
		Min	Max			Min	Max
TAZ Controls Only	3	-9.89	17.05	TAZ + H-Age	3	-4.94	24.21
	5	-32.11	13.10		5	-28.18	12.31
	13	-14.38	30.04		13	-8.09	19.75
TAZ + P-Age	3	-1.77	1.66	TAZ + H-Age + P-Age	3	-3.38	5.83
	5	-1.74	0.14		5	-3.34	2.47
	13	-3.08	1.72		13	-5.25	6.97

Conclusions

- PopGen 2.0 is able to accommodate controls at multiple spatial resolutions simultaneously
- Highly efficient algorithm facilitates generation of representative synthetic population
- There is a considerable improvement in fit when additional controls are utilized
- More variables will not necessarily improve the quality of synthesis as much as more unique variables will

PopGen 2.0

- The extended IPU procedure has been incorporated into a completely new implementation named PopGen 2.0
- Beta release available https://github.com/foss-transportationmodeling/popgen/releases
- More information: http://www.simtravel.org/popgen.html



Acknowledgements

- Sponsors:
  - US Department of Transportation's Federal Highway Administration under its Exploratory Advanced Research Program
  - Southern California Association of Governments
  - Maricopa Association of Governments
  - Baltimore Metropolitan Council
- People
  - Bhargava Sana
  - Xin Ye
  - Hillel Bar-Gera
  - Keith Christian

Questions?