



IRA A. FULTON SCHOOL OF ENGINEERING

Leading engineering discovery and innovative education for global impact on quality of life.

Development of a Synthetic Population Generator for Activity-Travel Demand Forecasting

Baltimore Metropolitan Council

**Kick-off Meeting and Workshop
November 9-10, 2010**



Acknowledgements

- *Software Development:* Karthik Konduri and Bhargava Sana
- *Graphic Support and Documentation:* Keith Christian
- *Methodology:* Xin Ye, University of Maryland; Hillel Bar-Gera, Ben-Gurion University, Israel
- *Sponsors:*
 - Arizona State University, School of Sustainable Engineering and the Built Environment, Ira A. Fulton School of Engineering
 - Exploratory Advanced Research Program (EARP), Federal Highway Administration, US Department of Transportation

Outline

- Motivation for population synthesis
- What is population synthesis?
 - Standard IPF procedure
- Motivation for enhanced population synthesis
- Design of a new population synthesizer
- New Iterative Proportional Updating (IPU) Algorithm
 - Explanation of procedure
 - Geometric Interpretation
- Test Application
 - Computing household weights
 - Generating a synthetic population
 - Algorithm performance
- Demonstration of PopGen Open Source Software Package

Microsimulation Models of Travel

- Increasing interest in microsimulation models for travel demand forecasting
- Microsimulation models simulate travel at the level of the individual decision-maker while recognizing inter-dependencies among activities, trips, persons, time, and space
- Microsimulation models of travel increasingly based on activity-based paradigm of travel behavior
 - Explicit recognition of derived nature of travel demand
 - Enhanced representation of time-space interactions and constraints

Microsimulation Models of Travel (continued)

- Activity-based microsimulation modeling approaches offer ability to address emerging policy questions of interest
- By simulating activities and travel at the level of the individual traveler, these models are able to address impacts of:
 - Greenhouse gas emissions reduction targets
 - Flexible working arrangements
 - Impact of information and communication technology (ICT)
 - Interactions between micro-scale land use changes and travel
 - Pricing-based policies
 - Non-motorized transportation mode enhancements

Why Population Synthesis?

- We need disaggregate household and person socio-demographic data for entire population of model region
- Such data for the entire population is generally not available
- This leads to the need to synthesize a regional population from known statistical distributions on the population
- We have:
 - Disaggregate data for a sample of the population (PUMS, travel surveys)
 - Marginal distributions for the entire region (census summary files, agency forecasts)

What is Population Synthesis?

Population synthesis involves generating a synthetic population by expanding the disaggregate sample data to mirror known aggregate distributions of household and person variables of interest.

Standard IPF-Based Procedure

- Standard IPF (iterative proportional fitting)-based procedure based on Beckman et al (1996)
- Procedure
 - Choose household-level control variables
 - Obtain the marginal distributions on these variables from census summary files (SF)
 - Generate a seed matrix of the joint distribution from a microdata sample data set (PUMS, travel survey)
 - Expand the seed matrix using an IPF-procedure to match the given marginal control totals while maintaining the joint distribution implied by the seed matrix

Standard IPF-Based Procedure (continued)

- Selection probabilities are estimated for households in the microdata sample
- Households are drawn using the selection probabilities to match the expanded cell frequencies
- The resulting synthetic population is checked for goodness-of-fit and households are redrawn if necessary
- The synthetic population is comprised of all individuals within the synthesized (drawn) households

Illustration of IPF Procedure

Sample Seed Data and Summary Marginal Distributions

		Income		Total	Household Size Marginals
		Low	High		
Household Size	Adjustment	--	--		
1	--	3.0	1.0	4.0	30.0
2	--	2.0	4.0	6.0	40.0
3 or more	--	2.0	1.0	3.0	30.0
Total		7.0	6.0		
Income Marginals		60.0	40.0		

Marginal Distributions

Seed Data

Marginal Distributions

Note: In the original image, a red box highlights the 'Income Marginals' row, and a yellow box labeled 'Seed Data' has an arrow pointing to the 'Income Marginals' row. A blue box highlights the 'Household Size Marginals' column, and another blue box labeled 'Marginal Distributions' has an arrow pointing to the 'Household Size Marginals' column.

Illustration of IPF Procedure (continued)

Iteration 1: Adjustment for Income

Household Size	Adjustment	Income		Total	Household Size Marginals
		Low	High		
		$60/7 = 8.57$	6.67		
1	--	$3 \times 8.57 = 25.7$	6.7	32.4	30.0
2	--	17.1	26.7	43.8	40.0
3 or more	--	17.1	6.7	23.8	30.0
Total		60.0	40.0		
Income Marginals		60.0	40.0		

Illustration of IPF Procedure (continued)

Iteration 1: Adjustment for Household Size

		Income		Total	Household Size Marginals
		Low	High		
Household Size	Adjustment	--	--		
1	$30.0/32.4 = 0.93$	$25.7 \times 0.93 = 23.8$	6.2	30.0	30.0
2	0.91	15.7	24.3	40.0	40.0
3 or more	1.26	21.6	8.4	30.0	30.0
Total		61.1	38.9		
Income Marginals		60.0	40.0		

Illustration of IPF Procedure (continued)

After 3 Iterations, convergence is achieved

Household Size	Adjustment	Income		Total	Household Size Marginals
		Low	High		
		--	--		
1	1.00	23.6	6.4	30.0	30.0
2	1.00	15.2	24.8	40.0	40.0
3 or more	1.00	21.3	8.7	30.0	30.0
Total		60.0	40.0		
Income Marginals		60.0	40.0		

Multiway frequency table matching known marginal distributions

Summary of IPF Procedure

- The standard IPF-based procedure explained in detail in Beckman et al (1996)
- The IPF-based procedure has been implemented widely in various population synthesizers
- Following the estimation of the cell frequencies in the joint distribution, households are drawn probabilistically

Motivation for Enhancement

- Key limitation of the standard IPF-based procedure
 - Controls only for household attributes and not person attributes
 - Synthetic populations fail to match distributions of person characteristics of interest
 - The method ignores differences in household composition among households within a cell
- ***Hence the need to re-assign weights to sample households based on household composition***

Recent Literature Addresses Issue

➤ Guo and Bhat (2007)

- “... deviation (in person attributes) could severely affect the accuracy of the subsequent microsimulation outcome ...”
- Household- and person- joint distributions are estimated using IPF procedure
- Household selection probabilities computed based on target distributions of household types
- A sample household is drawn so long as the household and person level frequency counts are within a certain threshold of the given distributions

Recent Literature (continued)

- Arentze and Timmermans (2007)
 - Person level marginal constraints are converted into household level constraints using relational matrices
 - Household constraints and the converted person level constraints are used to estimate household joint distributions using the standard IPF procedure

Recent Literature (continued)

➤ Pritchard and Miller (2009)

- IPF implemented with a sparse list-based data structure that can accommodate a large number of control variables
- A conditional Monte Carlo drawing procedure is adopted to simultaneously fit household and person marginal distributions
- Persons within households are drawn from a pool while maintaining person to household relationships
- Enhances the fit to person distributions while maintaining the match to household marginals

Recent Literature (continued)

➤ Srinivasan et al (2009)

- A “fitness value” is calculated for each sample household
- “Fitness value” captures the contribution of the sample household in matching both household and person distributions
- Synthetic population is generated by selecting sample households with the highest fitness values
- Drawing process continues until the expected number of households are drawn or all fitness values become negative

PopGen: A New Population Synthesizer

- Incorporates a new Iterative Proportional Updating (IPU) algorithm for estimating household weights
- The algorithm estimates sample household weights such that ***BOTH*** household and person distributions are matched
- Simple, practical, and computationally tractable algorithm with an intuitive interpretation
- Basic idea behind IPU algorithm in PopGen
 - ***Reallocate weights among sample households of a type to account for differences in household composition***

PopGen Methodology

Step 1: Estimate Household and Person Type Constraints

- household and person sample data
- household and person level marginal distributions

- Adjust priors to account for zero-cell problem
- Adjust marginals to account for the zero-marginal problem
- Run Iterative Proportional Fitting (IPF) procedure to estimate **household and person type constraints**

PopGen Methodology (continued)

Step 2: Estimate Household Weights

- household and person sample data
- household and person type constraints from Step 1

- Run the Iterative Proportional Updating (IPU) algorithm to estimate sample **household weights** that satisfy both household and person type constraints

PopGen Methodology (continued)

Step 3: Generate the Synthetic Population

- household and person sample data
- household weights from Step 2

- Apply rounding procedures to get the frequency of different household types in the synthetic population
- Estimate household selection probabilities using the computed weights
- Draw sample households based on selection probabilities for each household to match cell frequencies
- Repeat the process until a ***synthetic population*** with the best fit is obtained

PopGen Terminology

➤ Household Type

- Not to be confused with a household attribute 'household type'
- Refers to a combination of household-level variables of interest
- Represents a cell in the joint distribution of a set of household-level variables

➤ Person Type

- Similar to above – formed by a combination of multiple person-level variables of interest

PopGen Terminology (continued)

- A measure of fit (δ value)
 - Measures the absolute relative deviation between the IPU-adjusted cell frequency and the IPF-estimated household/person type constraints
 - Average δ value across all constraints is used as a goodness-of-fit measure
 - Average δ value is also used to monitor and set convergence criterion for the IPU algorithm

PopGen Terminology (continued)

- A measure of fit (δ value)

$$\delta_j = \frac{|d_{i,j} w_i - c_j|}{c_j}$$

$d_{i,j} w_i$ = adjusted cell frequency

c_j = the j^{th} IPF-estimated constraint

Illustration of IPU Algorithm

Frequency Matrix

Household ID	Initial 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_0		0.9143	0.9231	0.9011	0.8923	0.9327

Illustration of IPU Algorithm (continued)

Adjustment with respect to household type constraints

Household ID	Initial 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	35/3 = 11.67	65/5 = 13.00
Constraints		35.00	65.00	91.00	65.00	104.00		
δ_0		0.9143	0.9231	0.9011	0.8923	0.9327		
Weighted Sum 1		35.00	5.00	51.67	28.33	28.33		
Weighted Sum 2		35.00	65.00	111.67	88.33	88.33		

Illustration of IPU Algorithm (continued)

Adjustment with respect to person type constraints

Household ID	Initial 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	5	9	7	7					
Constraints		35	65	91	65	104	35/3 = 11.67	65/5 = 13.00	91/111.67 = 0.81	65/76.80 = 0.85	104/67.68 = 1.54
δ		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					
δ_1		0.0006	0.0015	0.1521	0.3222	0					

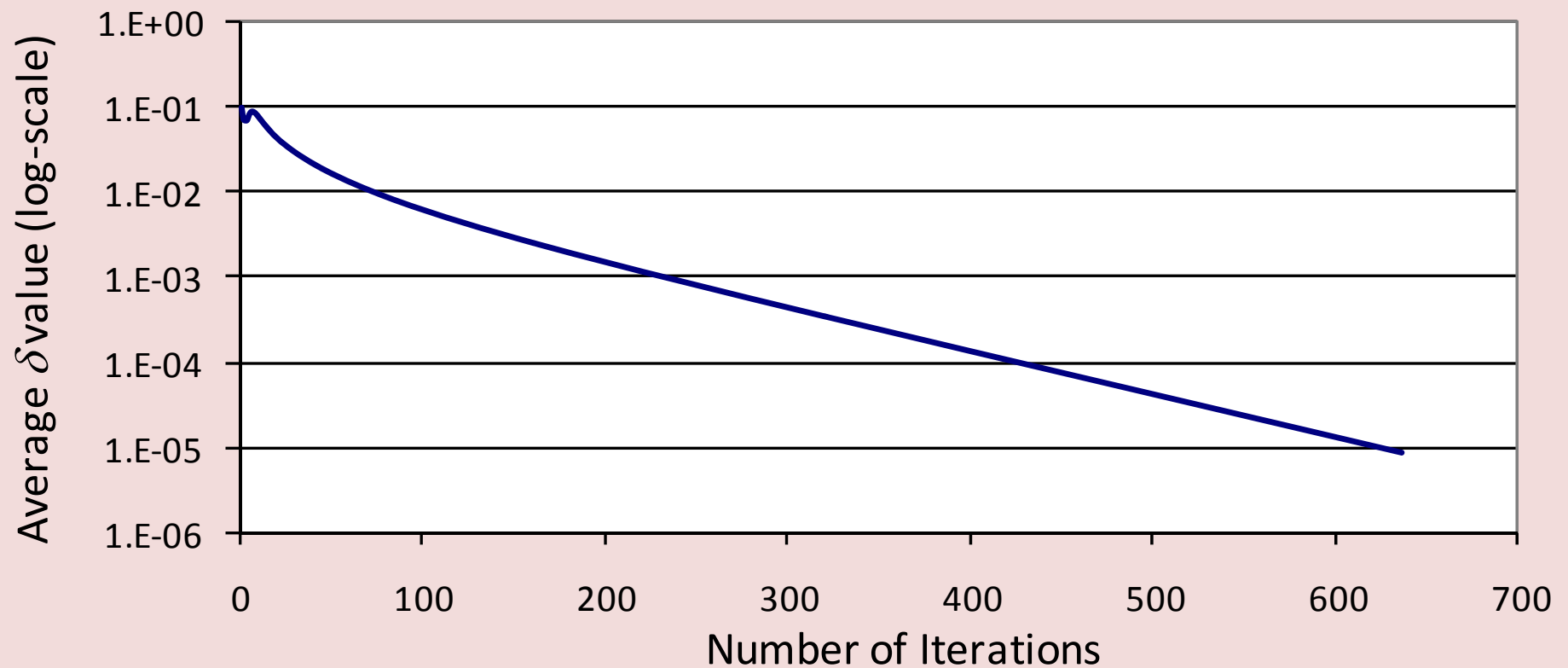
Illustration of IPU Algorithm (continued)

Final Results

Household ID	Initial Weights	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	IPU (New) 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
Constraints		35.00	65.00	91.00	65.00	104.00	
δ_0		0.9143	0.9231	0.9011	0.8923	0.9327	
δ_{IPU}		0.0000	0.0000	0.0000	0.0000	0.0000	

Illustration of IPU Algorithm (continued)

Improvement in Average δ Value



IPU: Geometric Interpretation

- Consider the following household structure and population constraints

Household ID	Household Type 1	Person Type 1	Weights
1	1	0	w_1
2	1	1	w_2
Constraints	4	3	

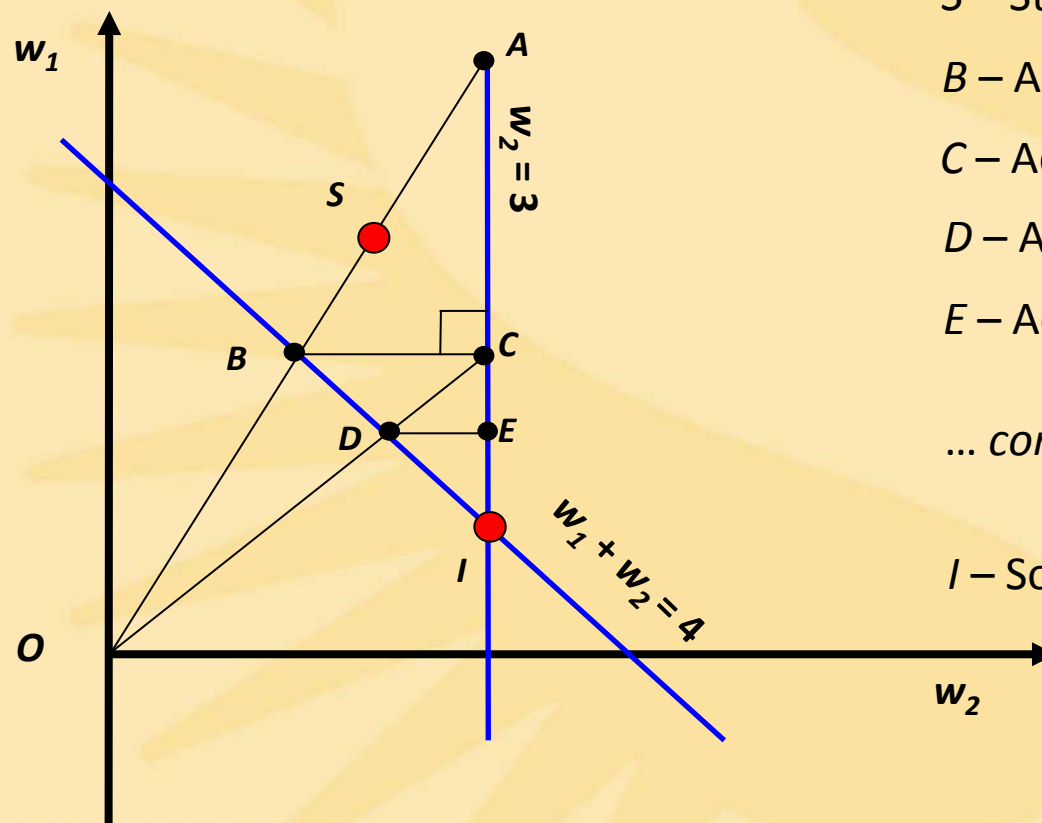
- Weights can be estimated by solving the following system of linear equations

$$w_1 + w_2 = 4$$

$$w_2 = 3$$

IPU: Geometric Interpretation (continued)

When solution is **within** the feasible region



S – Starting Point

B – Adjustment for Household Constraint

C – Adjustment for Person Constraint

D – Adjustment for Household Constraint

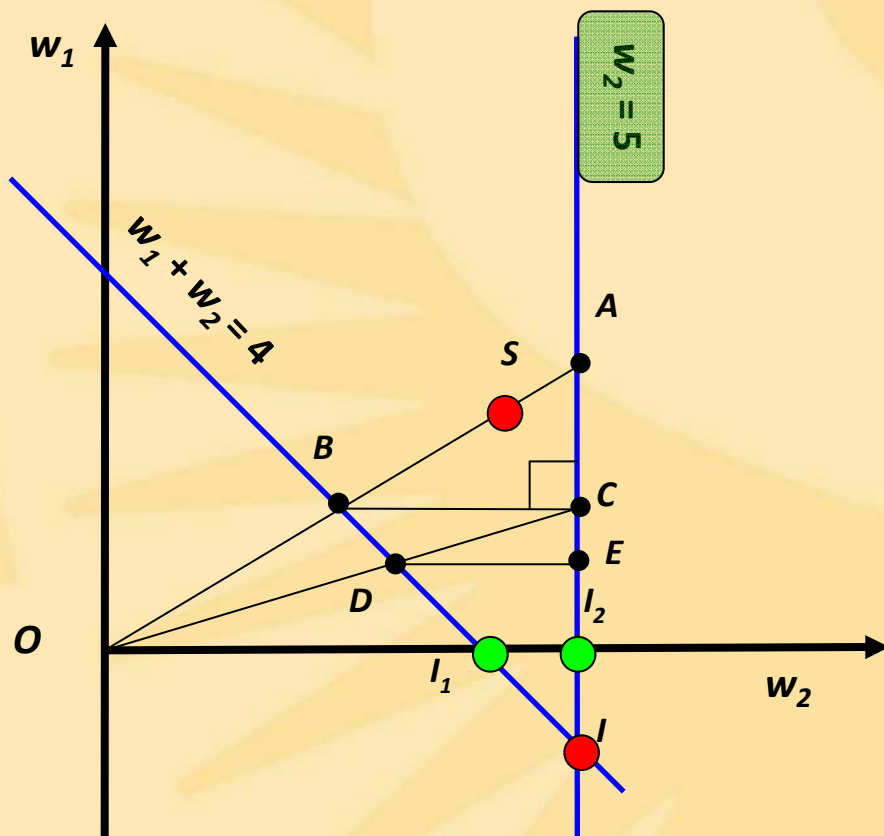
E – Adjustment for Person Constraint

... continue to convergence

I – Solution

IPU: Geometric Interpretation (continued)

When solution is **outside** the feasible region



S – Starting Point

B – Adjustment for household constraint

C – Adjustment for person constraint

D – Adjustment for household constraint

E – Adjustment for person constraint

... continue to convergence

I – Solution outside feasible region

I_1 – Corner solution where household constraint is satisfied

I_2 – Corner solution where person constraint is satisfied

A Test Application of PopGen

- Test area – Maricopa County, Arizona
- Population estimates from Census 2000
 - 3,071,219 individuals
 - 1,133,048 households and 44,689 group quarters
 - 2,090 blockgroups
- Sample household and person data obtained from 2000 PUMS
 - 254,205 individuals
 - 95,066 households
 - 5,489 groupquarters
- Marginal distributions of attributes obtained from 2000 Census Summary Files
- Synthetic population generated at level of blockgroup

Test Application: Control Variables

Household Attributes

- Household Type (5 categories)
 - 1) Family: Married Couple; 2) Family: Male Householder, No Wife; 3) Family: Female Householder, No Husband; 4) Non-family: Householder Alone; 5) Non-family: Householder Not Alone
- Household Size (7 categories)
 - 1) 1 Person; 2) 2 Persons; 3) 3 Persons; 4) 4 Persons; 5) 5 Persons; 6) 6 Persons; 7) 7 or more Persons
- Household Income (8 categories)
 - 1) \$0 - \$14,999; 2) \$15,000 - \$24,999; 3) \$25,000 - \$34,999; 4) \$35,000 - \$44,999; 5) \$45,000 - \$59,999; 6) \$60,000 - \$99,999; 7) \$100,000 - \$149,999; 8) Over \$150,000
- Presence of Own Children (2 categories)
 - 1) Yes; 2) No
- **560 household type constraints**

Test Application: Control Variables (continued)

Person Attributes

➤ Gender (2 categories)

1) *Male*; 2) *Female*

➤ Age (10 categories)

1) *Under 5 years*; 2) *5 to 14 years*; 3) *15 to 24 years*; 4) *25 to 34 years*; 5) *35 to 44 years*; 6) *45 to 54 years*; 7) *55 to 64 years*; 8) *65 to 74 years*; 9) *75 to 84 years*; 10) *85 and more*

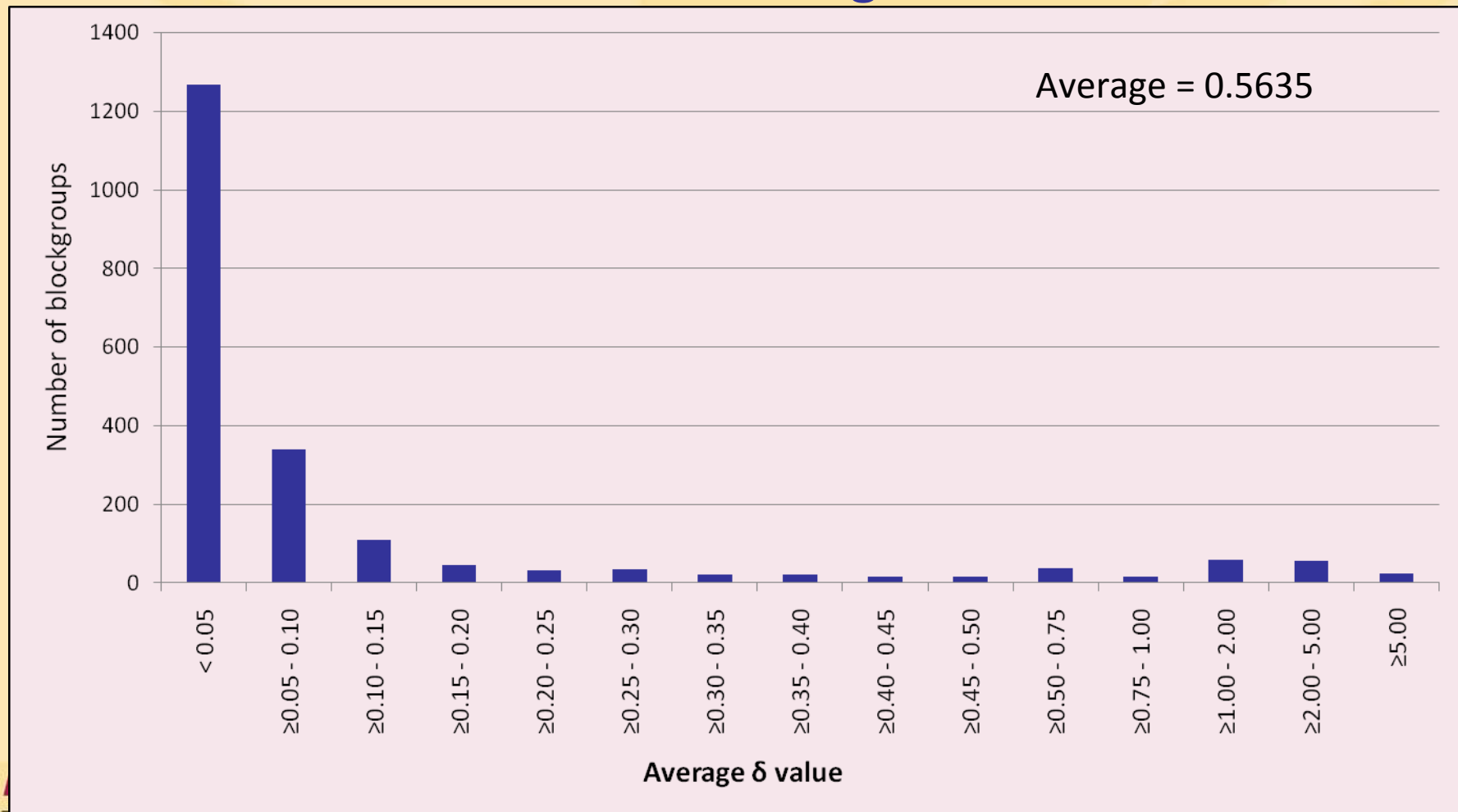
➤ Ethnicity (7 categories)

1) *White alone*; 2) *Black or African American alone*; 3) *American Indian and Alaska Native alone*; 4) *Asian alone*; 5) *Native Hawaiian and Other Pacific Islander alone*; 6) *Some other race alone*; 7) *Two or more races*

➤ **140 person type constraints**

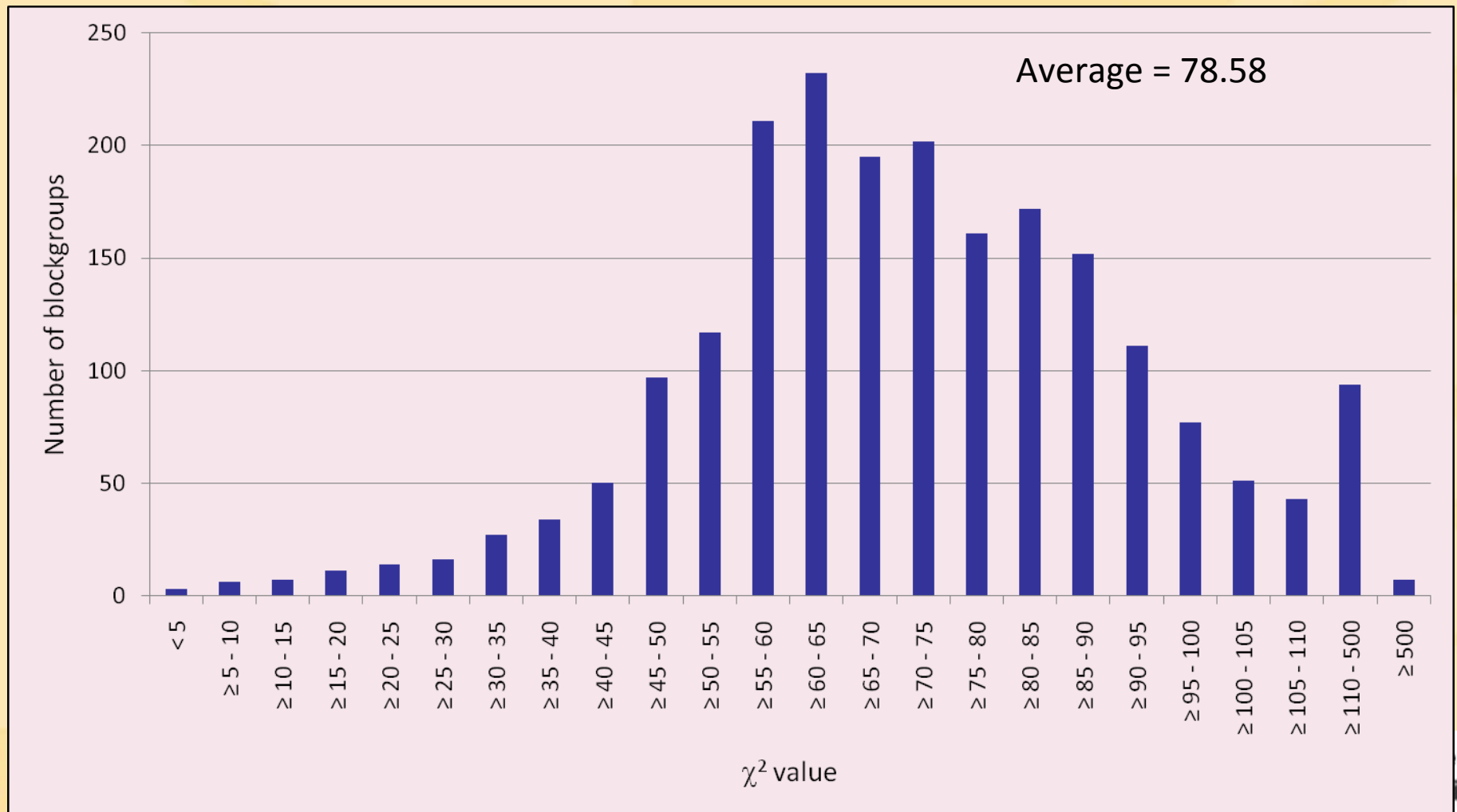
Test Application: Regional Results

Distribution of Average δ Value



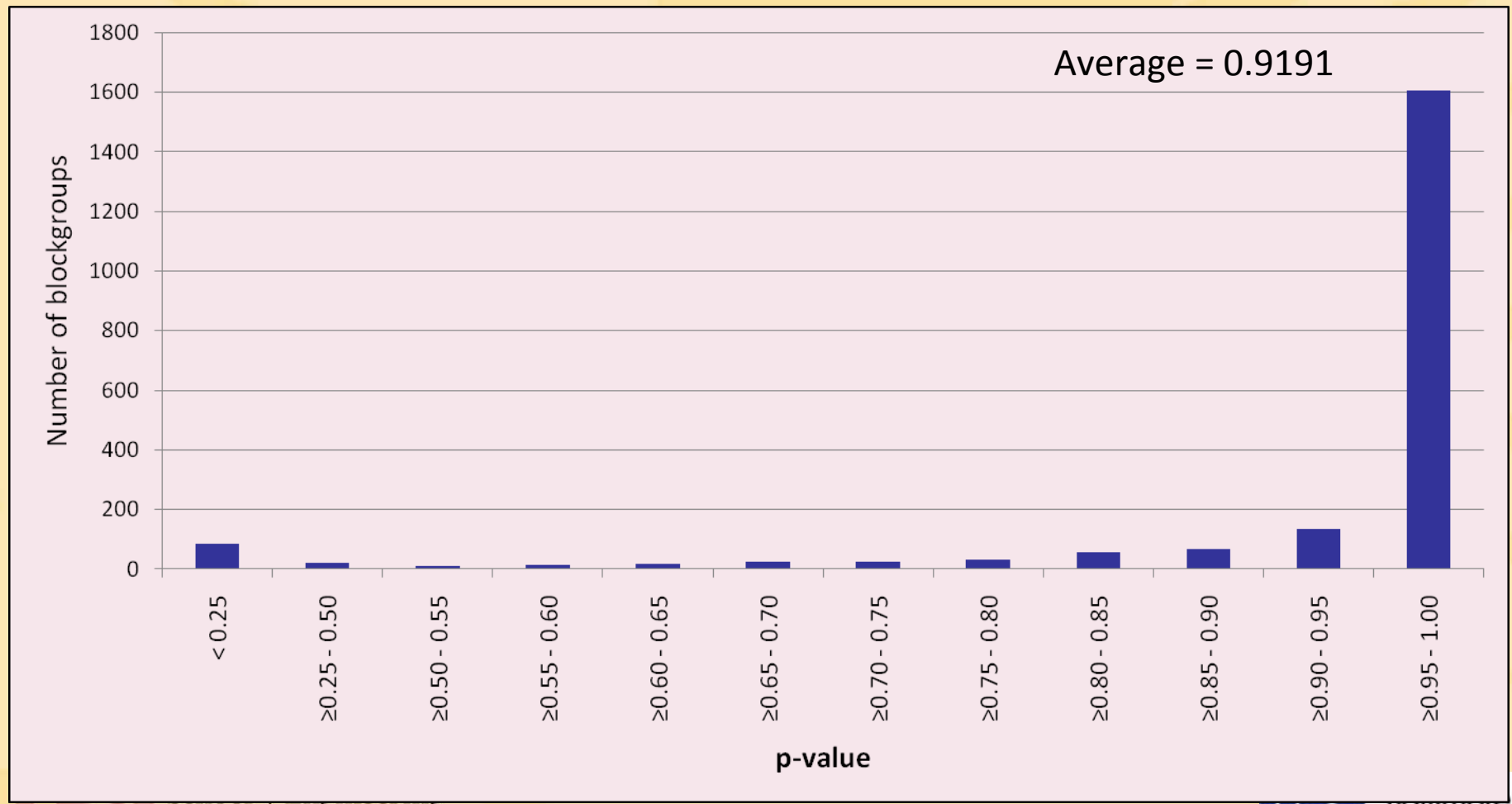
Test Application: Regional Results (continued)

Distribution of χ^2 -value



Test Application: Regional Results (continued)

Distribution of p-value

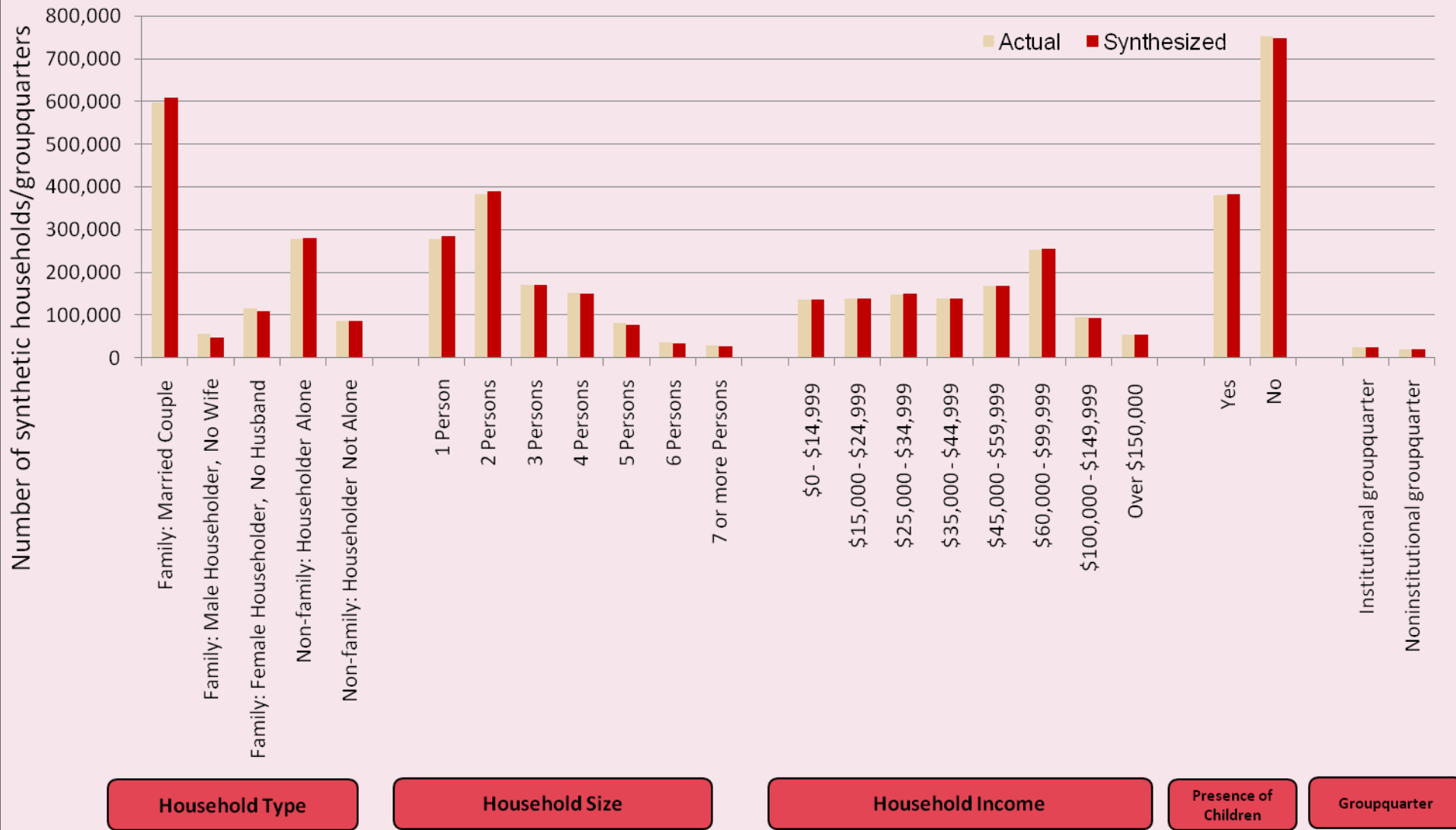


Test Application: Regional Results_(continued)

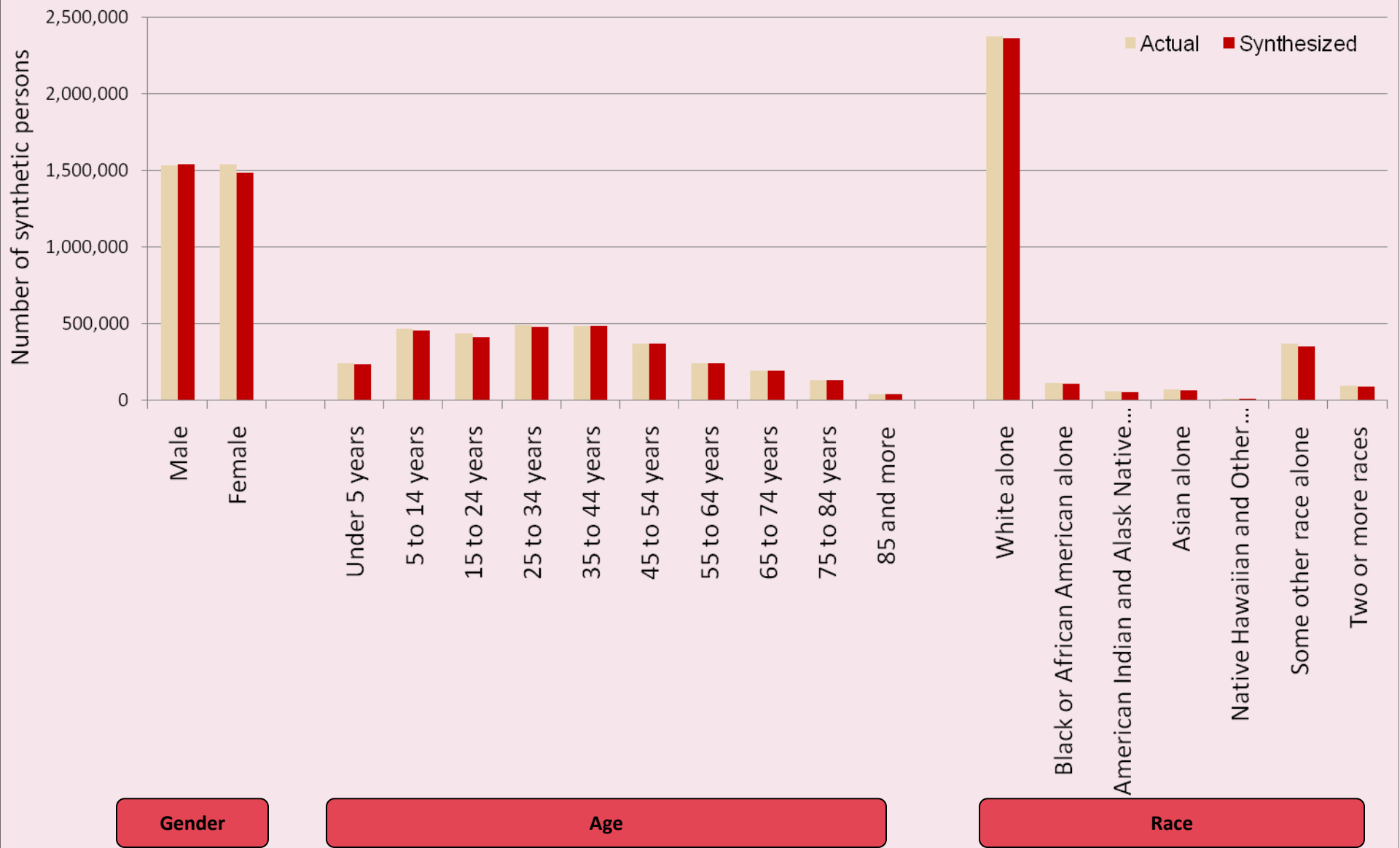
Comparison of estimated and actual frequencies

	<i>Estimated</i>	<i>Actual</i>
Households	1,133,048	1,133,048
Groupquarters	44,689	44,689
<i>Total</i>	<i>1,177,737</i>	<i>1,177,737</i>
Persons	3,020,695	3,072,149

Test Application: Regional Results (continued)



Test Application: Regional Results (continued)



Test Application: Regional Results (continued)

➤ Computational Performance

- Dell Precision T5400, quad core machine with Intel Xeon Processors and 4 GB of RAM
- Average **processing** time per blockgroup – 32 seconds
- Average **run** time per blockgroup using a parallel version of the code – 8 seconds
- Total **processing** time for 2090 blockgroups – approximately 18 hours and 35 minutes
- Total **run** time for 2090 blockgroups – approximately 4 hours and 40 minutes

Test Application: Sample Results

Results for two illustrative block groups

Blockgroup A

County – Maricopa

Tract ID – 111602

Blockgroup ID – 5

Near Perfect Solution Reached

Blockgroup B

County – Maricopa

Tract ID – 104203

Blockgroup ID – 2

Corner Solution Reached

Test Application: Sample Results (continued)

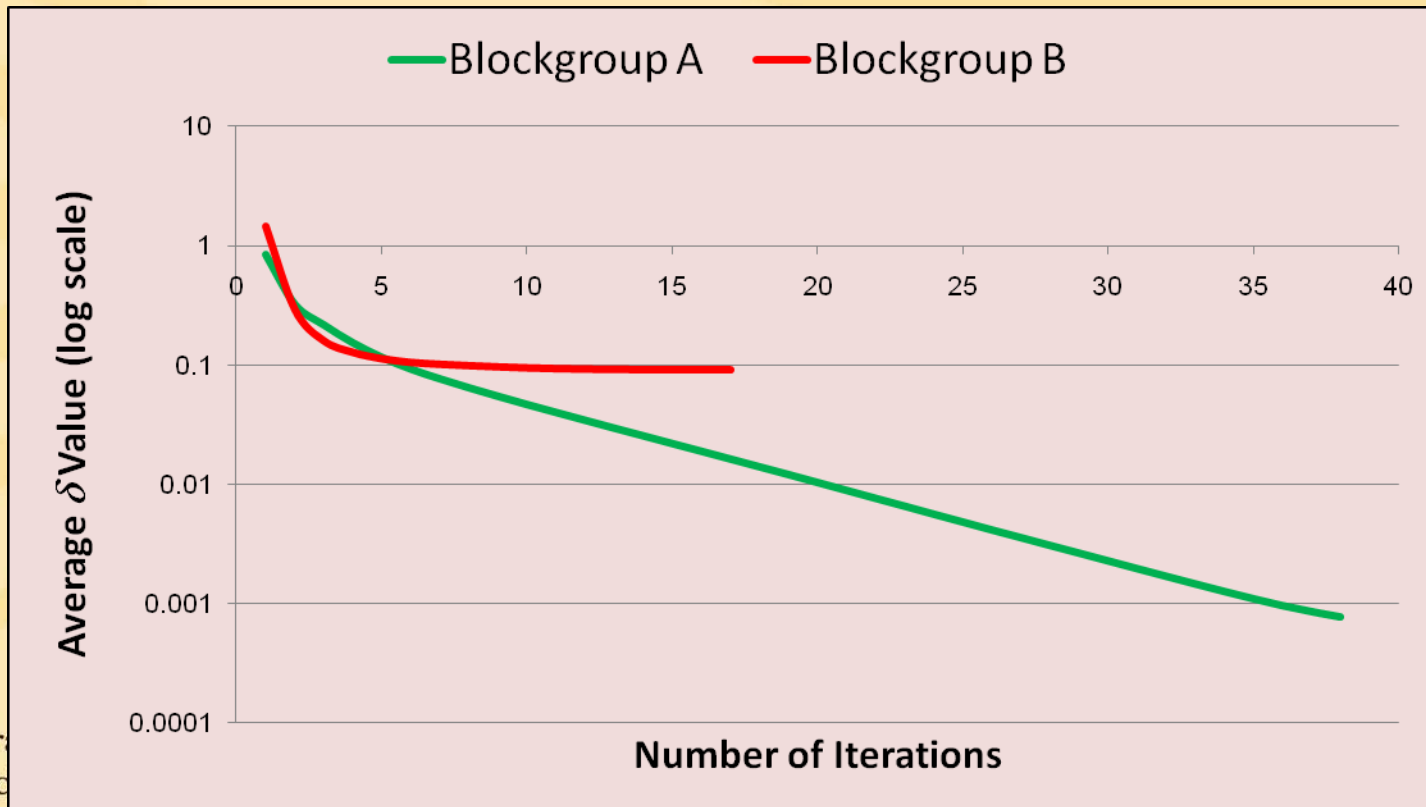
Reduction in Average Absolute Relative Difference (δ value)

Blockgroup A

δ 0.8385 \rightarrow 0.0008 in 38 iterations

Blockgroup B

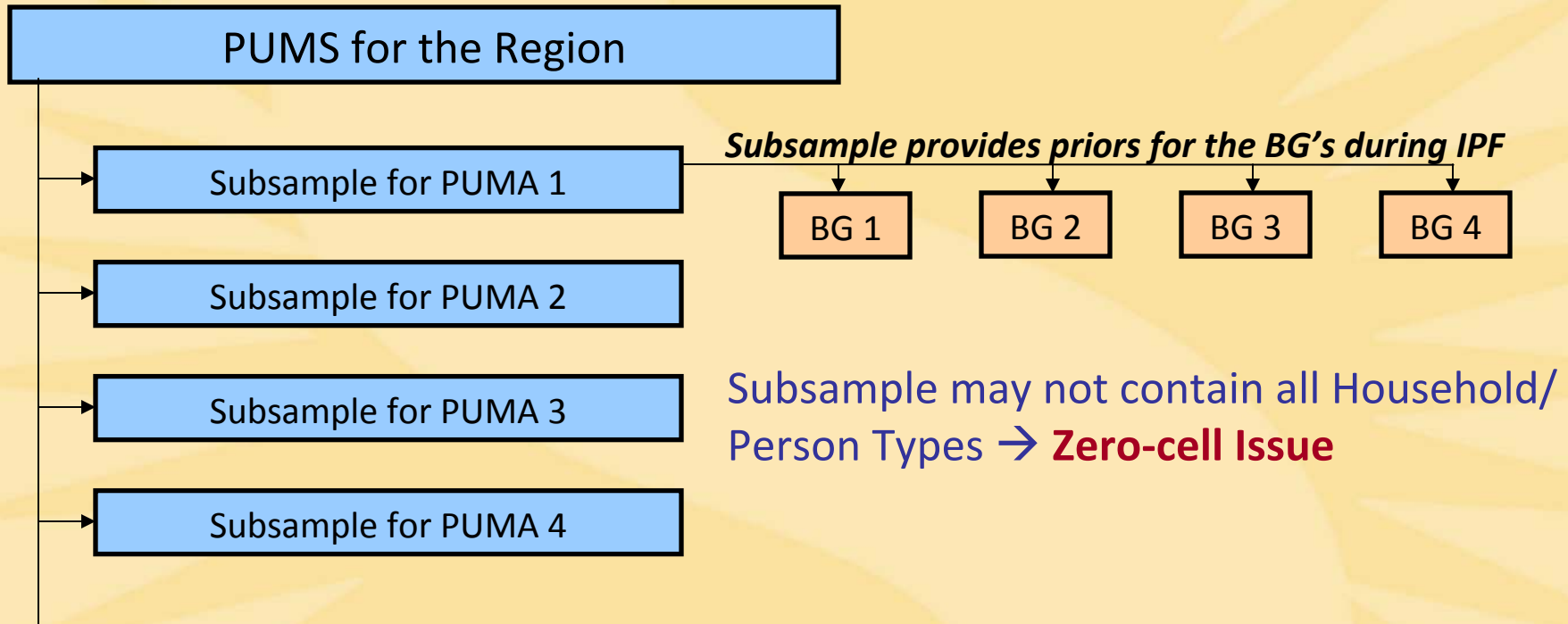
δ 1.4502 \rightarrow 0.0903 in 17 iterations



Small Geographies: Zero-Cell Correction

- Zero-cell Issue
 - The seed matrix from the sub-region (PUMA) to which the small geography belongs does not include infrequent household types
 - IPF for the geography may fail to converge
- Typical Approach
 - Add a small arbitrary number to the zero-cells (Beckman et al 1996)
 - This procedure introduces an arbitrary bias (Guo and Bhat, 2007)
- Solution Implemented in PopGen
 - ***Borrow prior information for the zero cells from the PUMS data for the entire region subject to an upper limit***

Small Geographies: Zero-Cell Correction (continued)



Small Geographies: Zero-Cell Correction (continued)

Priors from PUMA

		Household Income		
		High	Low	
Household Size Category	1	3	0	12
	2	2	4	
	3 or more	2	1	
Total				

Priors from Total PUMS

		Household Income		
		High	Low	
Household Size Category	1	7	2	33
	2	8	10	
	3 or more	3	3	
Total				

Probabilities from PUMA

		Household Income		
		High	Low	
Household Size Category	1	0.25	0.00	
	2	0.17	0.33	
	3 or more	0.17	0.08	

Threshold Probability = $1/12 = 0.083$

Probabilities from Total PUMS

		Household Income		
		High	Low	
Household Size Category	1	0.21	0.06	
	2	0.24	0.30	
	3 or more	0.09	0.09	

Small Geographies: Zero-Cell Correction (continued)

Zero-cell Adjustment

		Household Income	
		High	Low
Household Size Category	1	0.25	0.06
	2	0.17	0.33
	3 or more	0.17	0.08

Probability sum adds up to more than 1.00 (1.06)
 → adjust probabilities for other cells

Zero-cell Adjustment

		Household Income	
		High	Low
Household Size Category	1	0.25×0.94	0.06
	2	0.17×0.94	0.33×0.94
	3 or more	0.17×0.94	0.08×0.94

Adjustment factor = $(1.00 - 0.06)$
 = 0.94

Adjusted priors

		Household Income	
		High	Low
Household Size Category	1	0.23	0.06
	2	0.16	0.31
	3 or more	0.16	0.08

Small Geographies: Zero-Marginal Correction

➤ Issue

- The marginal values for certain categories of an attribute take a zero value
- IPF procedure will assign a zero to all household/person type cells that comprise the zero-marginal category
- As a result the IPU algorithm may fail to proceed

➤ Solution implemented in PopGen

- Add a small value (0.001) to the zero-marginal categories
- IPU algorithm now proceeds to compute weights
- Effect of this small value on results is negligible

Small Geographies: Zero-Marginal Correction (continued)

Iteration 1 of IPU algorithm without correction

Household ID	Initial 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	0.00	0.00	0.00
2	1	1	0	1	0	1	11.67	11.67	0.00	0.00	0.00
3	1	1	0	2	1	0	11.67	11.67	0.00	0.00	0.00
4	1	0	1	1	0	2	1.00	13.00	0.00	0.00	0.00
5	1	0	1	0	2	1	1.00	13.00	13.00	55.00	150.00
6	1	0	1	1	1	0	1.00	13.00	0.00	0.00	0.00
7	1	0	1	2	1	2	1.00	13.00	0.00	0.00	0.00
8	1	0	1	1	1	0	1.00	13.00	0.00	0.00	0.00
Weighted Sum		3	5	9	7	7					
Constraints		35	65	0	110	150					
δ		0.9143	0.9231	-	0.9364	0.9533					
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		0.00	13.00	0.00	26.00	13.00					
Weighted Sum 4		0.00	55.00	0.00	110.00	55.00					
Weighted Sum 5		0.00	150.00	0.00	300.00	150.00					
δ_i		1.0000	1.3077	-	1.7273	0.0000					

Adjustment

$35/3 = 11.67$	$65/5 = 13.00$	$0/111.67 = 0.00$	$110/26 = 4.23$	$150/55 = 2.73$
----------------	----------------	-------------------	-----------------	-----------------

Small Geographies: Zero-Marginal Correction (continued)

Iteration 2 of IPU algorithm without correction

Household ID	Initial 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					
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	5	9	7	7					
Constraints		35	65	0	110	150					
δ_1		1.0000	1.3077	-	1.7273	0.0000					
Weighted Sum 1											
Weighted Sum 2											
Weighted Sum 3											
Weighted Sum 4											
Weighted Sum 5											
δ_2											

35/0 = undefined

Adjustment

Small Geographies: Zero-Marginal Correction (continued)

Iteration 1 of IPU algorithm with correction

Household ID	Initial 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	0.00	0.00	0.00
2	1	1	0	1	0	1	11.67	11.67	0.00	0.00	0.00
3	1	1	0	2	1	0	11.67	11.67	0.00	0.00	0.00
4	1	0	1	1	0	2	1.00	13.00	0.00	0.00	0.00
5	1	0	1	0	2	1	1.00	13.00	13.00	55.00	150.00
6	1	0	1	1	1	0	1.00	13.00	0.00	0.00	0.00
7	1	0	1	2	1	2	1.00	13.00	0.00	0.00	0.00
8	1	0	1	1	1	0	1.00	13.00	0.00	0.00	0.00
Weighted Sum		3	5	9	7	7					
Constraints		35	65	0.001	110	150					
δ_0		0.9143	0.9231	0.9011	0.8923	0.9327					
Weighted Sum 1		35.0000	5.0000	51.6700	28.3300	28.3300					
Weighted Sum 2		35.0000	65.0000	111.6700	88.3300	88.3300					
Weighted Sum 3		0.0003	13.0005	0.0010	26.0006	13.0007					
Weighted Sum 4		0.0010	55.0004	0.0035	110.0000	55.0006					
Weighted Sum 5		0.0019	149.9978	0.0064	299.9944	150.0000					
δ_1		0.9999	1.3077	5.3619	1.7272	0.0000					

Small Geographies: Zero-Marginal Correction (continued)

Iteration 2 of IPU algorithm with correction

Household ID	Initial 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	21.83	21.83	0.00	0.00	0.00
2	1	1	0	1	0	1	5.16	5.16	0.00	0.00	0.00
3	1	1	0	2	1	0	8.01	8.01	0.00	0.00	0.00
4	1	0	1	1	0	2	0.00	0.00	0.00	0.00	0.00
5	1	0	1	0	2	1	150.00	65.00	65.00	55.00	150.00
6	1	0	1	1	1	0	0.00	0.00	0.00	0.00	0.00
7	1	0	1	2	1	2	0.00	0.00	0.00	0.00	0.00
8	1	0	1	1	1	0	0.00	0.00	0.00	0.00	0.00
Weighted Sum		3	5	9	7	7					
Constraints		35	65	0.001	110	150					
δ_1		0.9999	1.3077	5.3619	1.7272	0.0000					
Weighted Sum 1		35.0000	149.9978	43.0097	329.8319	176.9928					
Weighted Sum 2		35.0000	65.0000	43.0074	159.8380	91.9946					
Weighted Sum 3		0.0008	64.9989	0.0010	129.9984	64.9995					
Weighted Sum 4		0.0007	54.9997	0.0009	110.0000	55.0003					
Weighted Sum 5		0.0017	149.9985	0.0018	299.9983	150.0000					
δ_2		1.0000	1.3077	0.8139	1.7273	0.0000					

Person Total Inconsistency

➤ Issue

- The person total range derived from the household size distribution is not consistent with the given person total
- Results in a corner solution and the estimated weights do not match the person-type constraints
- Synthetic population generated doesn't match given person totals

➤ Solution implemented in PopGen

- Adjust the household marginal distributions

Person Total Inconsistency (continued)

For household size category 3

- p_tot_min calculated using a value of 3
- p_tot_max calculated using a value of 8

hhldsize1	hhldsize2	hhldsize3	p_tot_min	p_tot_max	given_ p_tot
1026	816	359	3735	5530	3503
443	539	212	2157	3217	3612
773	679	235	2836	4011	4321
323	412	204	1759	2779	1523

Person Total Inconsistency (continued)

Step 1: Calculate the person total difference

Average household size for category 3 – 3.7

hhldsize1	hhldsize2	hhldsize3	p_tot_eq	given_p_tot	p_diff
1026	816	359	3986.3	3503	-483.3
443	539	212	2305.4	3612	1306.6
773	679	235	3000.5	4321	1320.5
323	412	204	1901.8	1523	-378.8

Person Total Inconsistency (continued)

Step 2: Calculate the corresponding household difference

$$hhld_diff = p_diff / (phhldsize1 + phhldsize2 * 2 + phhldsize3 * 3.7)$$

phhldsize1	phhldsize2	phhldsize3	p_diff	hhld_diff
0.47	0.37	0.16	-483.30	-266.85
0.37	0.45	0.18	1306.60	676.71
0.46	0.40	0.14	1320.50	742.44
0.34	0.44	0.22	-378.80	-187.03

Person Total Inconsistency (continued)

Step 3: Revise the household marginals

$$rhhldsize1 = hhldsize1 + phhldsize1 * hhld_diff$$

rhhldsize1	rhhldsize2	rhhldsize3	rp_tot
901.61	717.07	315.47	3503.00
694.07	844.48	332.15	3612.00
1113.19	977.82	338.42	4321.00
258.66	329.94	163.37	1523.00

Test Application: Synthetic Population

- Synthetic population generation process can be divided into three steps
 - Estimating whole frequencies
 - Calculating selection probabilities
 - Drawing households

Test Application: Estimating Frequencies

- IPF-estimated household type constraints provide target frequencies
 - Rounding procedures are employed to convert decimal values to whole frequencies
- Rounding procedures implemented in PopGen
 - Arithmetic Rounding (default)
 - Bucket Rounding
 - Stochastic Rounding

Test Application: Estimating Frequencies (continued)

➤ Arithmetic Rounding Procedure

- Round the decimal frequencies
- Account for the difference between the rounded frequency sum and the actual frequency sum

Test Application: Estimating Frequencies (continued)

Illustration of Arithmetic Rounding Procedure

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

Test Application: Estimating Frequencies (continued)

➤ Bucket Rounding Procedure

- The procedure ensures that the rounded frequency sum and the actual frequency sum are the same
- Keeps track of the accumulated rounding error
- Accumulated rounding error is used to bias the rounding of the next frequency value

Test Application: Estimating Frequencies (continued)

Illustration of Bucket Rounding Procedure

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

Test Application: Estimating Frequencies (continued)

➤ Stochastic Rounding Procedure

- Frequencies are randomly rounded up or rounded down
- Account for the difference between the rounded frequency sum and the actual frequency sum

1. Consider a household type frequency of 22.41
2. It can be rounded up with a probability of 0.41 and rounded down with a probability of 0.59
3. We randomly draw a number between 0 and 1 to decide which way the frequency gets rounded
 - Say if the random number was 0.20, then $0.00 \leq 0.20 \leq 0.41$, so the frequency gets rounded up to 23.00
 - Alternatively if the random number was 0.78, then $0.41 < 0.78 \leq 1.00$, so the frequency gets rounded down to 22.00

Test Application: Estimating Frequencies (continued)

Illustration of Stochastic Rounding Procedure

Household Type	Frequency	Rounded Frequency	Difference	Ranking to Receive a Household	Adjustment	Adjusted Frequency
1	64.85	64	-0.85	1	1	65
2	12.34	12	-0.34	8		12
3	10.36	10	-0.36	7		10
4	0.43	0	-0.43	4		0
5	0.49	0	-0.49	2	1	1
6	0.47	1	0.53	13		1
7	0.44	1	0.56	14		1
8	0.39	0	-0.39	5		0
9	0.48	0	-0.48	3		0
10	0.10	0	-0.10	12		0
11	0.12	0	-0.12	11		0
12	0.20	1	0.80	16		1
13	0.27	0	-0.27	10		0
14	0.28	0	-0.28	9		0
15	0.38	1	0.62	15		1
16	0.37	0	-0.37	6		0
Total	91.97	90	-1.97		2	92

Test Application: Selection Probabilities

- Synthetic households are drawn probabilistically based on IPU-estimated weights
- Selection probabilities are estimated for each household type that needs to be synthesized
- No additional adjustments to match person constraints are needed
- The individuals from the synthetic households comprise the synthetic population

Test Application: Selection Probabilities (continued)

Household ID	Household Type 1	Household Type 2	Person Type 1	Person Type 2	Person Type 3	Final Weights	Household Type 1		Household Type 2	
							Cumulative Sum	Probability	Cumulative Sum	Probability
1	1	0	1	1	1	1.36	1.36	0.0389	-	-
2	1	0	1	0	1	25.66	27.02	0.7720	-	-
3	1	0	2	1	0	7.98	35.00	1.0000	-	-
4	0	1	1	0	2	27.79	-	-	27.79	0.4276
5	0	1	0	2	1	18.45	-	-	46.24	0.7115
6	0	1	1	1	0	8.64	-	-	54.88	0.8444
7	0	1	2	1	2	1.47	-	-	56.35	0.8671
8	0	1	1	1	0	8.64	-	-	64.99	1.0000

Test Application: Drawing Households

- Rounded frequencies and the selection probabilities from earlier steps are used to generate a synthetic population
- For each household type, we use the corresponding selection probabilities to draw households
- The persons in the drawn households comprise the synthetic population for the target year
- As the drawing procedure is probabilistic, the fit of the synthetic population is checked
- The drawing procedure is repeated until a synthetic population with the best fit is obtained

Test Application: Drawing Households (continued)

Household ID	Household Type 1		Household Type 2	
	Cumulative Sum	Probability	Cumulative Sum	Probability
1	1.36	0.0389	-	-
2	27.02	0.7720	-	-
3	35.00	1.0000	-	-
4	-	-	27.79	0.4276
5	-	-	46.24	0.7115
6	-	-	54.88	0.8444
7	-	-	56.35	0.8671
8	-	-	64.99	1.0000
Frequency	35		65	

1. Consider Household Type 1
2. Generate a random number between 0 and 1, e.g. 0.23
3. **0.0389 < 0.23 < 0.7720**
4. Household ID – 2 is added to the synthetic population
5. The process is repeated until 35 households of Household Type 1 are included
6. The process is repeated for Household Type 2

Test Application: Synthetic Population

➤ χ^2 goodness-of-fit statistic

- A goodness-of-fit measure to check match against person-level distributions
- The corresponding p-value represents the level of confidence at which the synthetic population matches the given constraints
- A synthetic population is drawn repeatedly until a desired p-value is achieved or a maximum number of draws is reached
- Maximum number of draws is user specified and dependent on geographic context

$$\chi^2 = \sum_j \left[\frac{(n_j - c_j)^2}{c_j} \right]$$

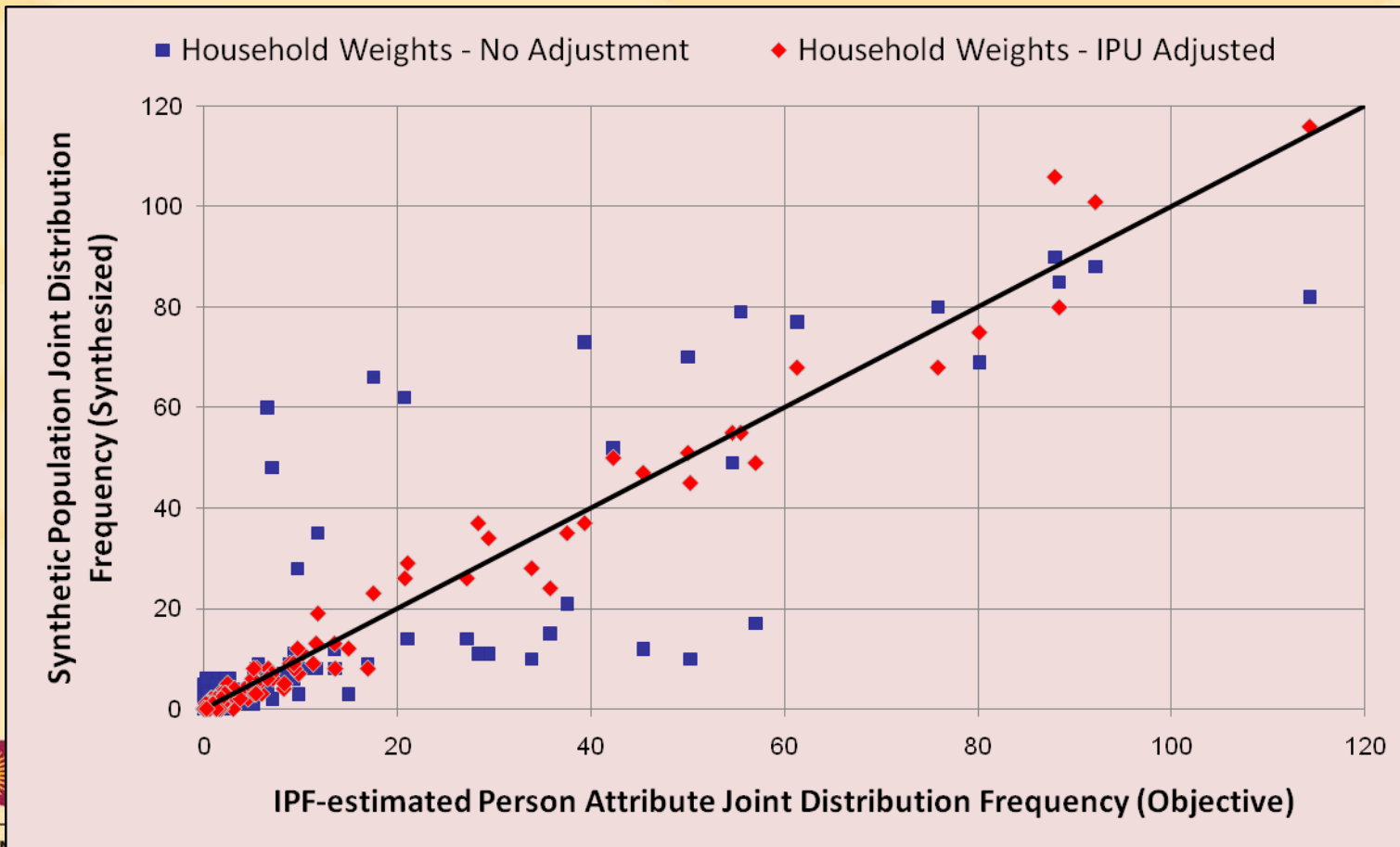
n_j = frequency of synthetic persons of the j^{th} person-type

c_j = the j^{th} IPF-estimated person-type constraint

Test Application: Performance

Blockgroup A

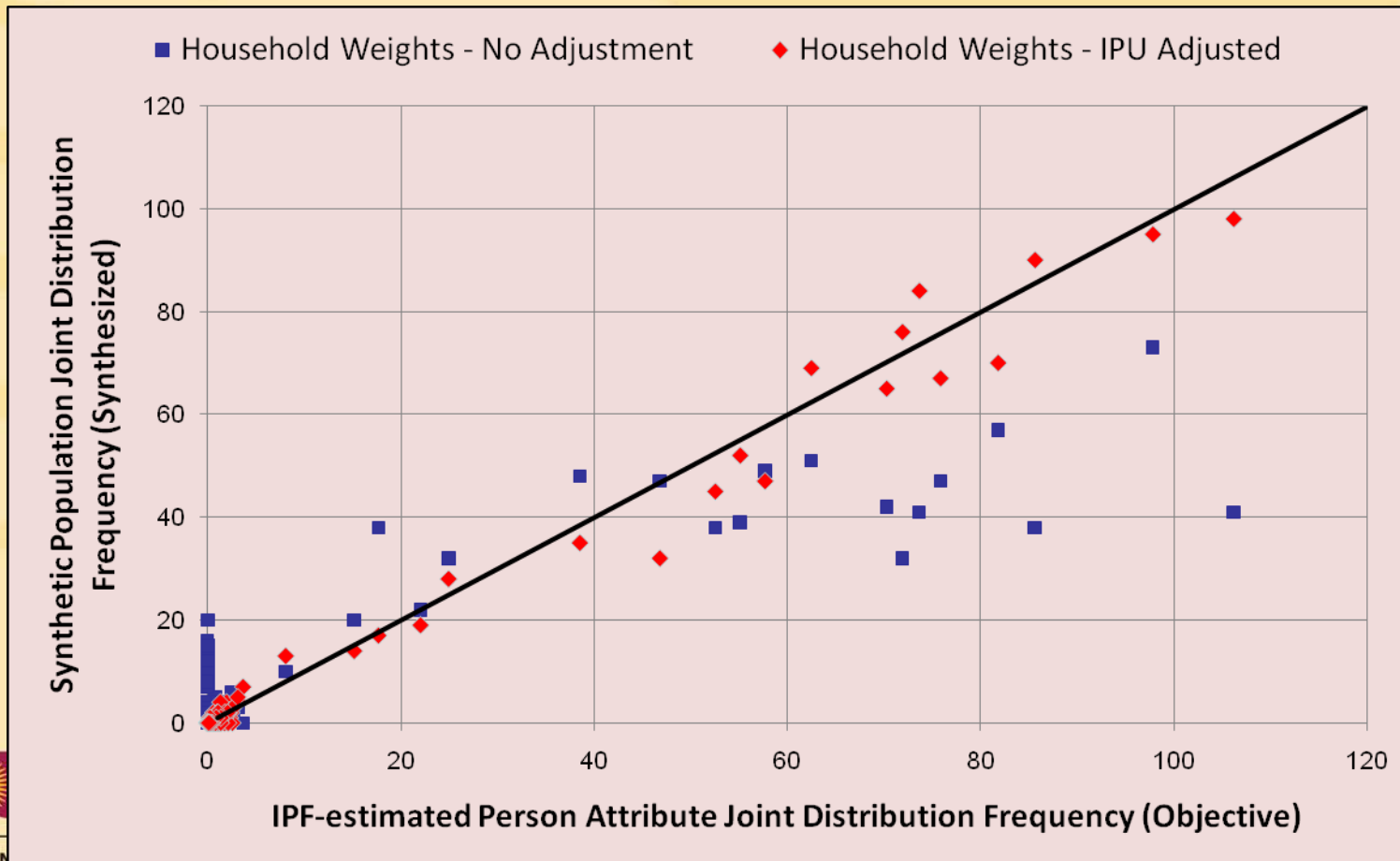
$\chi^2 = 93.8, df = 120, p\text{-value} = 0.9632$



Test Application: Performance (continued)

Blockgroup B

$\chi^2 = 61.9, df = 108, p\text{-value} = 0.9999$



Summary and Conclusions

- State of practice moving towards disaggregate microsimulation modeling of travel demand
- Need synthetic population to implement microsimulation models
- Standard IPF-based procedures for synthetic population generation generally do not control for both household- and person-control variables
- PopGen incorporates new IPU algorithm based on concept of redistributing household weights to reflect differences in household composition
- Test application shows procedure is practical, computationally feasible, and provides a synthetic population that is more representative of the true population

Development of PopGen Software

- Features of the package
 - A stand-alone application
 - Graphic User Interface to enhance user-friendliness
 - Data downloading, processing and editing capabilities
 - Modify marginal distributions to match person totals more closely
 - Synthesis using classic and IPU approach
 - Interface for visualizing and exporting the results

Development of PopGen Software (continued)

- PopGen 1.0 was released on July 15, 2009
- PopGen 1.1 was released on November 15, 2009

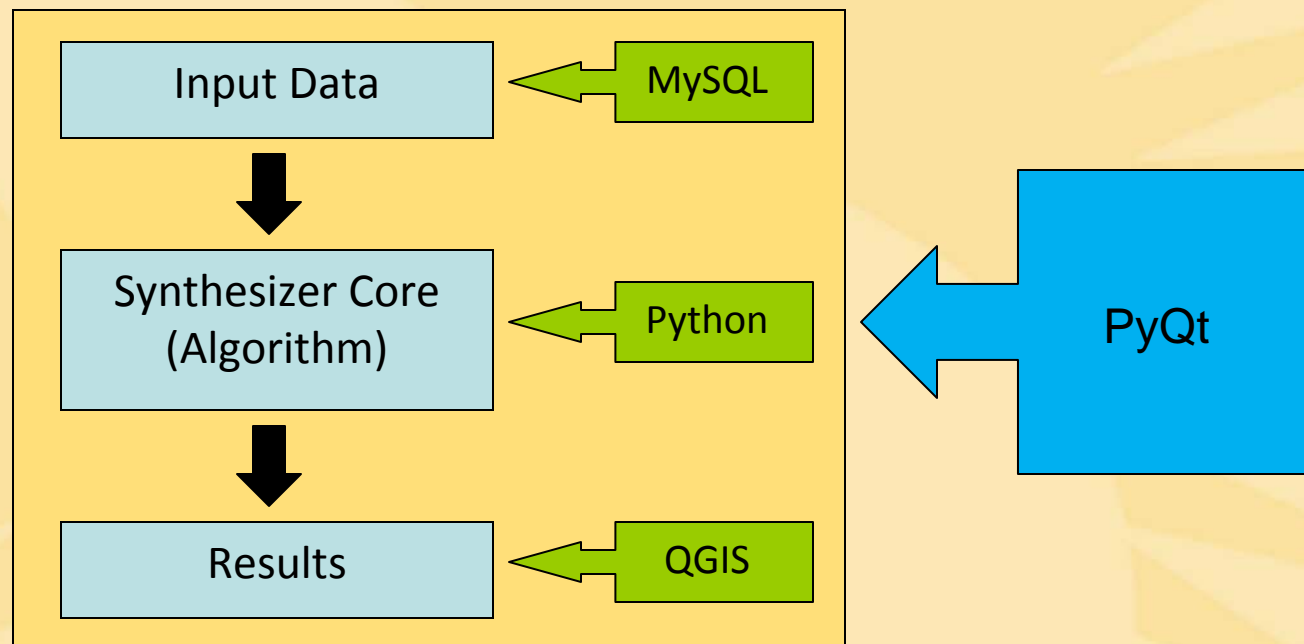
Main Website:

<http://urbanmodel.asu.edu/popgen.html>

Wiki site:

<http://simtravel.wikispaces.asu.edu/Population>
[+Synthesis](#)

PopGen: Open Source Framework



Population Evolution Model

- Evolve the base year synthetic population to obtain population for a future year
- A composite set of models to capture all the evolutionary processes
 - Migration of households in and out of a region
 - Person-level life cycle events
 - Household-level changes over time

Population Evolution Model

Household Migration Models

- Emigration Model: Rate-based probability model which deletes households in excess in a future year
- Immigration Model: Rate-based probability model which adds households of types that are deficient in the future year

Population Evolution Model

Person Evolution Models

- Aging Model: Deterministic model that increments the age of a person by one year
- Fertility Model: Rate-based fertility model for women predicting the birth of a child based on age, race etc.
- Mortality Model: Mortality rate-based model predicting the death of a person based on age, gender etc.
- Educational Attainment Model: Logit-based model predicting if a person will be in school in year t controlling for age, schooling completed, status in year $t-1$ etc.

Population Evolution Model

Person Evolution Models (continued)

- Occupation Choice Model: Multinomial-logit model to predict occupation choice based on age, experience, wage etc. Includes a non-work choice.
- Wage Model: This model would set the wages to clear the job market in each occupation choice. It would be used by the occupation choice model.
- Mobility Options Model: Rate-based model for predicting driver-license and or transit pass holder status.

Population Evolution Model

Household Evolution Models

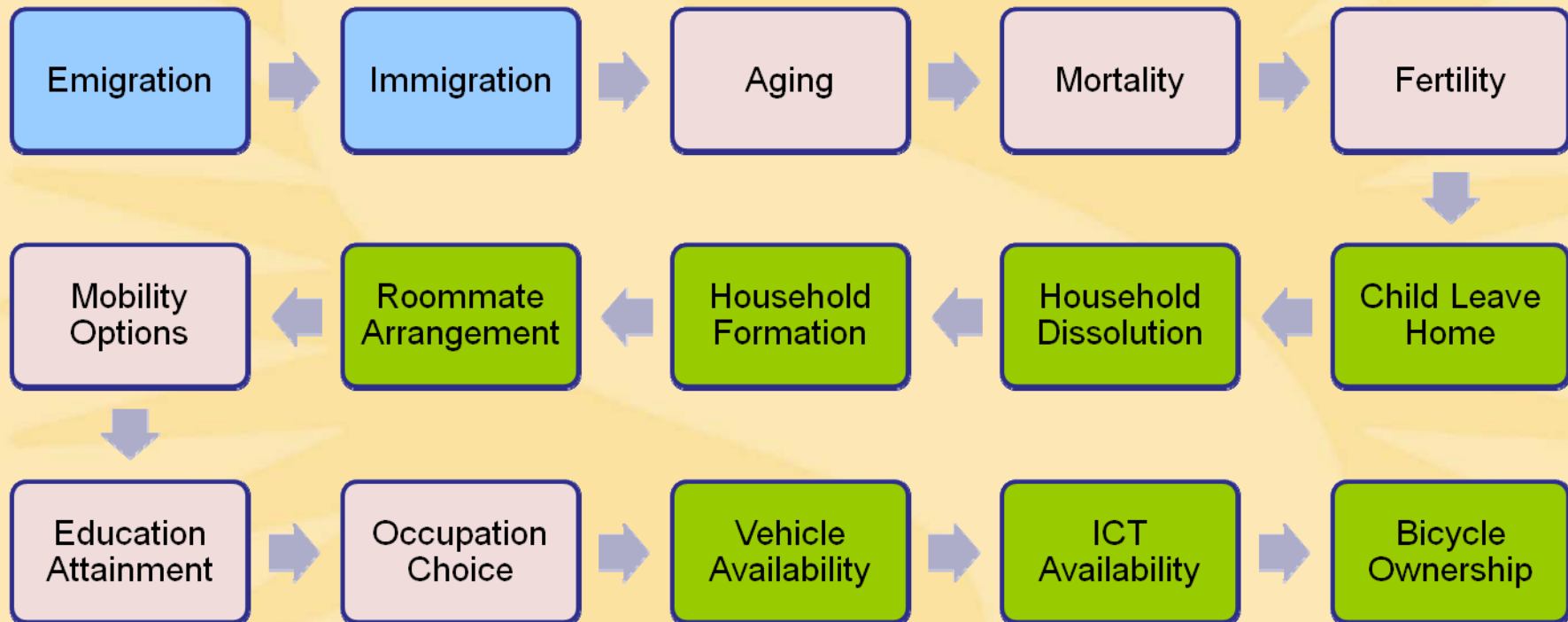
- Child Leaving Home Model: Model based on rates specific to age and gender. A new household is created and assigned to the person.
- HH Separation Model: Predicts the probability that a household with two or more adults chooses to separate. Children, if present, are allocated to a primary residence.
- Household Formation Model: This model predicts the event of persons of different genders combining to form a household. Need to develop “scoring” system that will “match” individuals.

Population Evolution Model

Household Evolution Models (continued)

- Roommate Model: Probability that two persons of the same gender will cohabitate is predicted. Need a “scoring” system that will “match” individuals together.
- Auto Availability Model: This ordered-probit/ multinomial logit model would predict number of vehicles by type for a household.
- Bicycle Ownership Model: Number of bicycles present in the household is predicted.
- ICT Availability Model: Predicts cell phones, computers, and internet connectivity for a household. Use market penetration statistics to determine ICT availability.

Population Evolution Model



Population Evolution Model

➤ Challenges

- Data availability
- Model estimation
- Reconciling household interactions and dependencies
- Modeling simultaneous choices, e.g., Education and Occupation choices
- Endogeneity of choices, e.g., auto ownership and residential/workplace location choices (typically in land use model)
- Overall sequencing of events

Questions

?