Comparing Methods of Synthetic Population Generation

Shannon Gallagher and Lee Richardson
sgallagh, leerich @stat.cmu.edu
Carnegie Mellon University
Department of Statistics
Models of Infectious Disease Agent Study (MIDAS)

Joint work with Rebecca C. Steorts
Duke University

May 10, 2015

Outline

1. Motivation
2. Task and Data
3. Methods: SRS, IPF, MMM
4. Implementation in Butler and Sierra Leone
5. Summary/Future Work

What is a Synthetic Population?

- Micro-data with a row for every person in the population
- Desired Characteristics: School, Workplace, Has Car, etc...

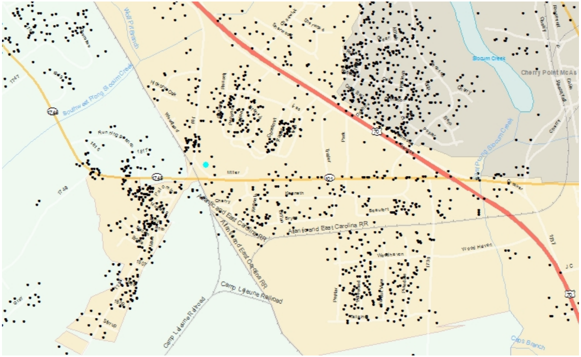


Figure : Source: http://www.gisagents.org/2012/05/synthetic-population-data-for-us.html

We were asked to generate synthetic populations

- For use in Agent Based Models (ABMs)
    - Models of Infectious Disease Agent Study (MIDAS)
    - Building on work done by Research Triangle Institute (RTI)
    - Model spread of disease through individual interactions
    - Require synthetic populations as input
- US → Western Africa
    - Motivated by the Ebola Outbreak

Finding data was a challenge

Lot's of considerations:

- Trustworthy
- Recency
- Geographic granularity
- Household vs. Individual data
- Variables to include

We utilized microdata repositories

Microdata is individual level data

Country	Year	Occupation	Household	Age	Gender
USA	2010	Statistician	1234	72	M
USA	2010	Data Scientist	1234	54	F
USA	2010	Epidemiologist	1234	56	M
USA	2010	Student (Stats)	1234	23	F
USA	2010	Student (CS)	1239	21	F
USA	2010	Artist	1239	24	M

Main data sources:

- Microdata: IPUMS-I, 5% Representative Sample
- Population Counts: Geohive
- Other: Summary Tables, Demographic Averages

We utilized microdata repositories

Microdata is individual level data

Country	Year	Occupation	Household	Age	Gender
USA	2010	Statistician	1234	72	M
USA	2010	Data Scientist	1234	54	F
USA	2010	Epidemiologist	1234	56	M
USA	2010	Student (Stats)	1234	23	F
USA	2010	Student (CS)	1239	21	F
USA	2010	Artist	1239	24	M

Main data sources:

- Microdata: IPUMS-I, 5% Representative Sample
- Population Counts: Geohive
- Other: Summary Tables, Demographic Averages

We utilized microdata repositories

Microdata is individual level data

Country	Year	Occupation	Household	Age	Gender
USA	2010	Statistician	1234	72	M
USA	2010	Data Scientist	1234	54	F
USA	2010	Epidemiologist	1234	56	M
USA	2010	Student (Stats)	1234	23	F
USA	2010	Student (CS)	1239	21	F
USA	2010	Artist	1239	24	M

Main data sources:

- Microdata: IPUMS-I, 5% Representative Sample
- Population Counts: Geohive
- Other: Summary Tables, Demographic Averages

Three Different Methods

Goal:

- Populations which are as accurate as possible
- Match best method with data availability

Methods:

- Simple Random Sampling (SRS)
- Iterative Proportional Fitting (IPF)
- Method of Moment Matching (MMM)

SRS allowed us to ‘get off the ground’

- Baseline comparison
- Let us meet deadlines
- Focus on syncing unharmonized sources

IPF is excellent for detailed data

- Well documented:
  - Deming and Stephan (1940)
  - Beckman, Baggerly, McKay (1996)
  - RTI- Wheaton et al. (2010)
- Idea:
  - Have Microdata *AND* marginal counts of 2+ variables
    Want contingency table
  - Fill in table based on IPF Algorithm
  - Sample PUMS based on 'distance' from cell

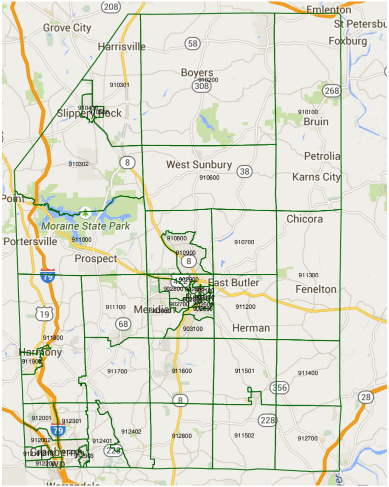
Marginal totals			i	
	Age/sex	Male	Female	T
j	Under-50	4	4	8
	Over-50	$\frac{8}{3}$	$\frac{4}{3}$	4
	T	$6\frac{2}{3}$	$5\frac{1}{3}$	12

MMM is our new method- adapted for Western Africa

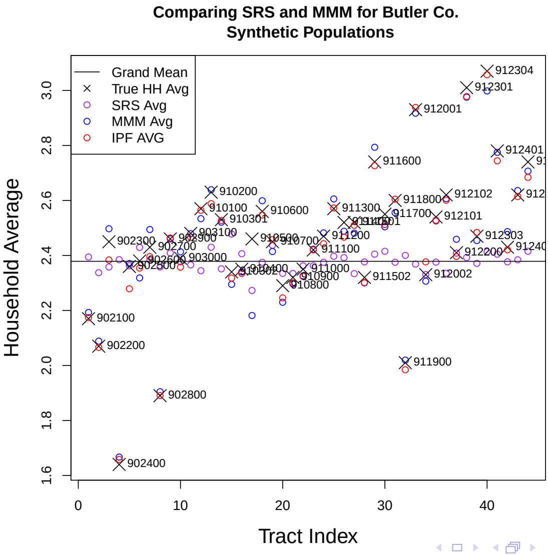
- Marginal counts often not available– required for IPF
- But first moment usually is
    - eg. average household size for a given region
- Formulation of quadratic program to sample from microdata
- Minimize ℓ^2 norm– include as much data as possible

For US example, we use Butler Co., PA

- Moderate population size ($\sim 180,000$ people)
- 44 Tracts, (income, gender, household size, etc.)

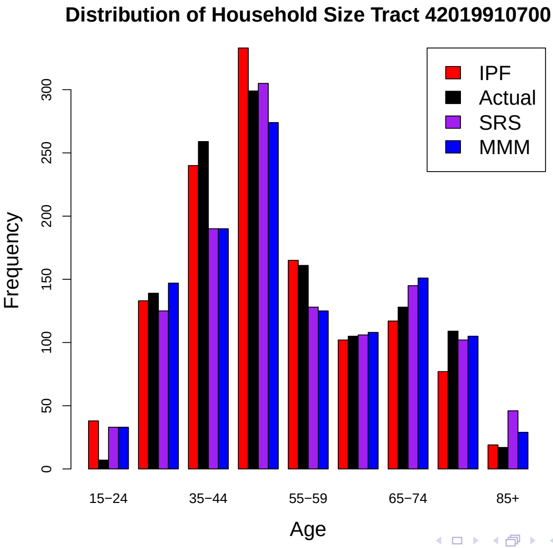


MMM is superior at matching household counts alone



Comparing SRS and MMM for Butler Co.
Synthetic Populations

IPF matches better when we add more variables

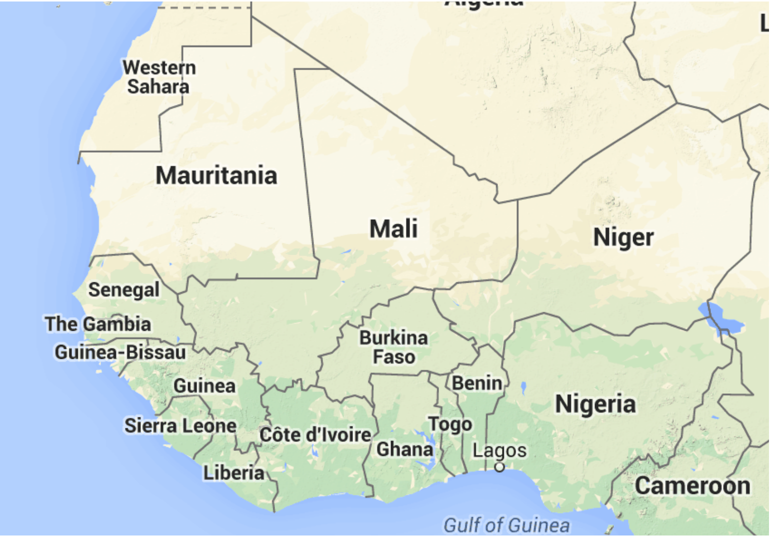


We also implemented methods for Sierra Leone

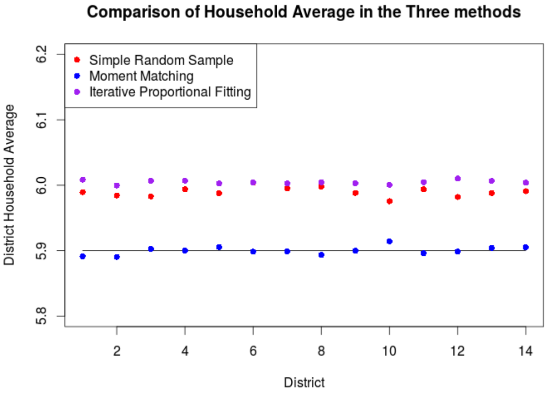
- The country that started it all for us

- Made up of 14 separate districts.
- Lack summary tables for each district
- Data: Household size average, Household head gender distribution

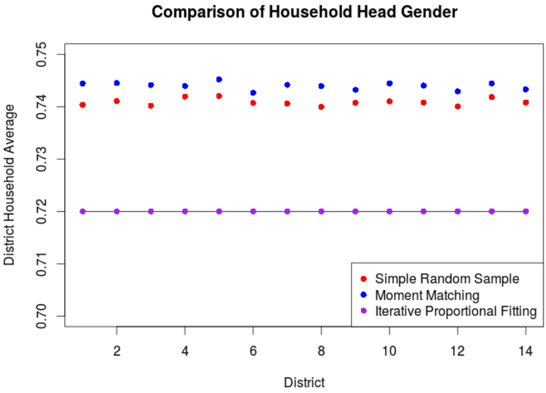
Map of Western Africa



MMM Does the best job matching the HH Average



But IPF can match more than one variable



MMM better matches Household Size, but IPF Can handle more variables

Butler

	MSE Age	MSE HH Size
IPF	4419	0.049
SRS	5964	0.071
MMM	5984	0.003

Sierra Leone

	MSE HH Size	MSE Head Ratio
IPF	0.391	0.0001
SRS	0.336	0.078
MMM	0.022	0.089

We would like to extend these methods

- Use multiple moments (e.g. variance)
- Use multiple variables for MMM
- Explore other, scalable options
  - Bayesian Hierarchical/Density Trees
- Records which are completely synthetic

We would like to extend these methods

- Use multiple moments (e.g. variance)
- Use multiple variables for MMM
- Explore other, scalable options
  - Bayesian Hierarchical/Density Trees
- Records which are completely synthetic

Generate the world!

Thank you!

Questions?