Using Census Public Use Microdata Areas (PUMAs) as Primary Sampling Units in Area Probability Household Surveys

Joe McMichael

Patrick Chen

Acknowledgement

- **The authors would like to thank our colleagues, Dr. Rachel Harter and Dr. Akhil Vaish for their help in preparation of this presentation.**

- **Part of the work for this study was funded by Energy Information Administration (EIA), Department of Energy under 2015 RECS Contract Nos. DE-EI-0000515.**

- **The views expressed in this presentation do not necessarily reflect the official policies of the EIA, Department of Energy, nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government.**

Outline

- **PUMA and PUMA Statistics**
- **Brief Review of Area Probability Household Survey Design**
- **Benefits of Using PUMA as PSU**
- **Concerns of Using PUMA as PSU**
- **Simulation Studies and Methods to Address Concerns of Using PUMA PSUs**
- **Conclusions**

PUMA and PUMA Statistics

- What is a PUMA?
  - ➢ Public Use Microdata Area
  - ➢ Tabulation and dissemination of decennial census and American Community Survey (ACS) Public Use Microdata Sample (PUMS) data.
- How PUMAs are formed in the 2010 Census
  - ➢ Nested in States or equivalent entities
  - ➢ Counties & equivalent entities and census tracts are geographic building blocks
  - ➢ At least 100,000 persons throughout the decades

PUMA and PUMA Statistics

	Estimated Occupied Housing Units		Land Area (Square Miles)	
	County	**PUMA**	**County**	**PUMA**
Minimum	39	24,484	2.0	1.4
P1	414	29,503	26.0	3.2
P25	4,367	41,515	430.7	37.4
P50	10,014	46,918	615.6	134.5
P75	25,840	56,363	924.0	947.7
P99	475,913	83,527	8,139.0	20,674.7
Maximum	3,241,204	120,193	145,504.9	438,781.1
N	3,143	2,351	3,143.0	2,351.0
Mean	37,135	49,645	1,123.7	1,502.3
Sum	116,716,292	116,716,292	3,531,925.0	3,531,925.0

Brief Review of Area Probability Household Survey Design

- Multi-stage cluster designs are employed
- Primary sampling units (PSUs) are selected at the first stage
- Smaller geographical areas or secondary sampling units (SSUs) are selected at the second stage
- PSU and SSU samples are selected using PPS sampling method
- Households/persons are selected at the third or fourth stage
- Counties or combinations of contiguous counties are commonly used as PSUs

Brief Review of Area Probability Household Survey Design (cont.)

Disadvantages of Using County PSUs:

- Collapsing small counties
- Large variation in the size measure for probability proportional to size (PPS) sampling
- Unequal weighting caused by certainty PSUs

Using PUMAs as PSUs

- Benefits of Using PUMA PSUs
  - A single PUMA can be used as a PSU
  - Smaller variation in size measure
  - More accurate size measure can be calculated from micro data
  - Improvement on design and stratification using micro data at PUMA level
  - Improvement in weighting using micro data (poststratification adjustment)

- Drawback of Using PUMA PSUs
  - PUMA definition may be changed in next decennial census

Concerns of Using PUMAs as PSUs

- Do PUMA PSUs have similar heterogeneity as county PSUs?
- Will PUMA PSUs cover core-based statistical areas represented by certainty county PSUs?
- Will PUMA PSUs increase field data collection costs?

Addressing the Concern of Heterogeneity

- **Large geographical areas have higher heterogeneity and smaller ICC than small geographical areas**
- **75% of PUMAs are smaller than 75% of counties**
- **Compared the within cluster variance for proportion variables for both PUMAs and counties**

$$Var\,(w) = \sum_{i}^{n} \frac{k_i p_i (1 - p_i)}{K - n},$$

where n is number of clusters, k_i is the number of sampling units within each cluster, K is the total number of sampling units in all clusters

Addressing the Concern of Heterogeneity (cont.)

Proportion Variable	Estimate	Within County Variance (VarC)	Within PUMA Variance (VarP)	Relative Diff ((VarP-VarC)/VarC)
Household Income <$50k	47.33%	23.87%	23.26%	-2.56%
Households in Poverty	15.37%	12.71%	12.44%	-2.12%
Persons Aged 65 and Older	5.60%	5.26%	5.25%	-0.19%
Persons Did Not Move in 12 Months	84.89%	12.67%	12.59%	-0.63%
Persons Now Married	50.97%	24.63%	24.35%	-1.14%
Persons 25 Years Old with Bachelors or Greater	22.91%	17.02%	16.56%	-2.70%
Hispanic	16.62%	11.09%	10.24%	-7.66%
African American	12.57%	9.34%	8.36%	-10.49%
Housing Units Detached	61.68%	21.34%	20.42%	-4.31%
Housing Units Rented	35.06%	21.59%	20.82%	-3.57%
Housing Units Using Gas as Main Heating	54.04%	18.82%	18.60%	-1.17%
Housing Units >=3 Bedrooms	59.96%	22.95%	22.13%	-3.57%

Addressing the Concern of CBSA Coverage

Conducted a Simulation Study to Assess the Coverage of PUMA PSU Sample on Core Based Statistical Areas (CBSAs)

- **Frame: PUMAs from 2010 Decennial Census**
- **Selection Method: Stratified PPS systematic sample**
- **Stratification: 19 RECS geographical domains**
- **Sample Size: total 200 PSUs**
- **Size Measure: Number of HUs in 2010 Decennial Census**
- **Sorting Variables:**
  - **Sort Trial 1: 2005 RECS certainty county indicator**
  - **Sort Trial 2: Density (Total HU/Land Area)**
  - **Sort Trial 3: 2005 RECS certainty county indicator and density**
- **Iterations: 1,000**
- **Probability of 20 largest CBSAs being included in 1,000 samples**

Addressing the Concern of CBSA Coverage (cont.)

CBSA	Number of Counties	# of Housing Units (2013)	Probability Sorting Trial 1	Probability Sorting Trial 2	Probability Sorting Trial 3
New York-Newark-Jersey City, NY-NJ-PA	25	7,821,586	1.00	1.00	1.00
Los Angeles-Long Beach-Anaheim, CA	2	4,522,188	1.00	1.00	1.00
Chicago-Naperville-Elgin, IL-IN-WI	14	3,791,572	1.00	1.00	1.00
Dallas-Fort Worth-Arlington, TX	13	2,602,427	1.00	1.00	0.99
Miami-Fort Lauderdale-West Palm Beach, FL	3	2,476,108	1.00	1.00	1.00
Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	11	2,438,169	0.98	0.98	0.98
Houston-The Woodlands-Sugar Land, TX	9	2,387,366	0.99	1.00	0.99
Washington-Arlington-Alexandria, DC-VA-MD-WV	24	2,278,746	0.99	0.99	0.99
Atlanta-Sandy Springs-Roswell, GA	29	2,190,417	0.99	0.99	0.98
Boston-Cambridge-Newton, MA-NH	7	1,889,080	0.98	0.97	0.99
Detroit-Warren-Dearborn, MI	6	1,887,874	0.97	0.95	0.97
Phoenix-Mesa-Scottsdale, AZ	2	1,832,428	1.00	0.99	1.00
San Francisco-Oakland-Hayward, CA	5	1,756,620	0.97	0.98	0.98
Riverside-San Bernardino-Ontario, CA	2	1,514,203	0.96	0.97	0.96
Seattle-Tacoma-Bellevue, WA	3	1,490,977	1.00	0.98	1.00
Minneapolis-St. Paul-Bloomington, MN-WI	16	1,405,948	0.98	0.99	0.99
Tampa-St. Petersburg-Clearwater, FL	4	1,361,831	0.88	0.88	0.88
St. Louis, MO-IL	15	1,230,506	0.91	0.93	0.94
San Diego-Carlsbad, CA	1	1,176,718	0.90	0.92	0.91
Baltimore-Columbia-Towson, MD	7	1,142,286	0.84	0.86	0.85
Average			0.97	0.97	0.97

Addressing the Concern of Data Collection Costs

Conducted a Simulation Study to Assess Whether PUMA PSUs Have Higher Field Costs

- Frame: PUMAs and counties from 2010 Decennial Census
- Selection Method: Stratified PPS systematic sample
- Stratification: 19 RECS geographical domains
- PSU Sample Size: 200 PUMA PSUs and 200 county PSUs
- SSU Sample Size: 4 census block groups (CBGs) per PSU
- Size Measure: Number of HUs in 2010 Decennial Census
- Sorting Variables: None
- Iterations: 1,000
- Calculating and comparing
  - Average CBG pair-wise travel distance within PSUs
  - Average CBG pair-wise travel distance within various distance thresholds

Addressing the Concern of Data Collection Costs (cont.)

Average CBG Pair-Wise Travel Distance within PSUs (miles)

Statistics	County	PUMA
Mean	13.83	13.79
10 Percentile	3.10	1.28
25 Percentile	6.04	2.47
Median	11.23	5.10
75 Percentile	18.53	13.01
90 Percentile	27.54	31.25

Addressing the Concern of Data Collection Costs (cont.)

Average CBG Pair-Wise Travel Distances within Distance Thresholds (miles)

	Within 10 Miles		Within 50 Miles		Within 70 Miles	
Statistics	County	PUMA	County	PUMA	County	PUMA
Mean	5.81	4.84	23.33	21.94	34.82	33.32
10 Percentile	2.09	1.33	5.78	3.45	7.42	4.69
25 Percentile	3.72	2.51	11.48	9.07	15.43	13.31
Median	5.98	4.59	21.75	20.38	32.33	30.76
75 Percentile	8.04	7.13	34.76	33.91	53.61	52.50
90 Percentile	9.21	8.82	43.76	43.36	66.73	66.25

Conclusions

Using PUMA as PSUs is a viable alternative

- PUMAs have similar heterogeneity as counties
- PUMA PSUs have very good coverage of major CBSAs
- PUMA PSUs will likely decrease field costs (cost neutral at worst)
- PUMA PSUs have several advantages compared to county PSUs
- 2015 Residential Energy Consumption Survey
- FDA Tobacco User Panel Survey

Contact Information

Patrick Chen
Senior Research Statistician
919-541-6309
pchen@rti.org

Joe McMichael
Research Statistician
919-485-5519
mcmichael.@rti.org