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

Joe McMichael, Patrick Chen
RTI International,[1] 3040 Cornwallis Road, Research Triangle Park, NC 27709

Abstract

Multistage cluster designs are employed in area probability household surveys. Samples of primary sampling units (PSUs) are selected at the first stage; within a PSU, smaller geographical areas are selected at subsequent stages, and households are selected at the final stage. Many area probability household surveys use single counties, or a combination of contiguous counties, as the PSU. The Census Public Use Microdata Areas (PUMAs) are defined for the dissemination of Census public use microdata and American Community Survey microdata. We used the PUMAs as PSUs in two national surveys. In this paper, we describe the benefits of using PUMAs as PSUs instead of counties. We compared the within-cluster variances for proportion estimates between counties and PUMAs. We also present the results of two simulation studies designed to address concerns with using PUMAs as PSUs (field costs and coverage of major metropolitan areas). Using PUMAs as PSUs is a viable option for in-person household surveys.

Key Words: Household Survey, Primary Sampling Units, PUMA PSU, Area Sampling, Multistage Sample Design

1. Introduction

In in-person household surveys, to reduce data collection costs, multistage cluster design has been commonly used. Samples of primary sampling units (PSUs) are selected at the first stage. Within a PSU, smaller geographical areas, secondary sampling units (SSUs) are selected at the second stage. Households or persons within households are selected at subsequent stages. In many large national surveys, a single county or a group of contiguous counties are employed as PSUs, for example, National Health Interview Survey, National Crime Victimization Survey, Survey of Income and Program Participation, Medicare Current Beneficiary Survey, and National Health and Nutrition Examination Survey. Typically, PSU samples are selected using the probability proportional to size measure (PPS) method; the size measure could be the number of housing units (Hus) or number of persons depending on the target survey population. Using a single county or a group of contiguous counties as PSUs (county-type PSUs) has some limitations or disadvantages. The first is that small counties need to be collapsed to meet minimum size requirements. The second is that county-type PSUs have large variation in the size measure used for PPS sampling. The size measure is subject to more errors because of large variation in the size measure. The inaccuracy of the size measure may compromise the self-weighting feature, thereby increasing variance of estimates. The third limitation is that when some large counties are included in the sample as certainty PSUs, unequal weighting issues arise. Although unequal weighting can be compensated for by selecting more SSUs, it may still increase the unequal weighting effect.

[1] RTI International is a registered trademark and a trade name of Research Triangle Institute.

The Census Public Use Microdata Areas (PUMAs) are statistical areas defined for the tabulation and dissemination of decennial census and American Community Survey data (U.S. Census, 2010). The 2010 Census PUMAs cover the entirety of the United States and Puerto Rico and are nested in states or equivalent entities. The geographic building blocks of PUMAs are counties and census tracts; large counties are split into multiple PUMAs, and small contiguous counties are combined into one PUMA. Each PUMA has at least 100,000 persons. There are 2,351 PUMAs defined in the 2010 Census and 3,143 counties. Table 1 compares the number of estimated occupied HUs and land area (in square miles) between counties and PUMAs. Counties have much larger variation in the number of estimated occupied HUs than PUMAs. The majority of PUMAs are geographically smaller than counties; only the top 25% of PUMAs are geographically larger than the top 25% of counties. Some PUMAs are very large geographically; the largest one covers 438,781 square miles.

Table 1. Comparison of Counties and PUMAs

	Estimated Occupied Housing Units		Land Area (Square Miles)	
	County	**PUMA**	**County**	**PUMA**
Min	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
Max	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

PUMAs have the same features as counties to be used as PSUs. We explored the feasibility of using PUMAs as PSUs (PUMA-type PSUs) and found that there are several advantages of using PUMA-type PSUs:

- A single PUMA can be readily used as a PSU—no collapsing is need.
- PUMAs have smaller variation in the size measure.
- For surveys for which the size measure is not readily available at county level, the size measure can be calculated from the microdata at PUMA level.
- There is rich information in the microdata at PUMA level—it can be used in the PSU stratification to improve sample design and in the poststratification weight adjustment to improve survey estimates.

One drawback of using PUMA-type PSUs is that some PUMAs may not be consistently defined across decennial censuses. There are three major concerns of using PUMA-type PSUs. Large counties are split into multiple PUMAs, and most PUMAs are geographically smaller than counties. In general, the larger the area is, the higher the heterogeneity is. Higher heterogeneity is desirable; it results in smaller intracluster correlation (Kish, 1965).

The first concern is: Do PUMA-type PSUs have similar heterogeneity as county-type PSUs? When using county-type PSUs, some large counties are included in the sample as certainty PSUs; likely no PUMAs are included in the sample as certainty PSUs when PUMA-type PSUs are used. The second concern is: Can PUMA-type PSU samples cover major core-based statistical areas (CBSAs) represented by certainty county–type PSUs? Because some PUMAs are geographically large, the third concern is: Will using PUMA-type PSUs increase field data collection costs? We address these three concerns in this paper.

2. Methods

2.1 Do PUMA-type PSUs Have Similar Heterogeneity as County-type PSUs?

Sampling units within a cluster tend to be similar, while sampling units in different clusters vary more; thus, the ratio of between-cluster variances over within-cluster variance is large in clustering sampling, thereby decreasing precision (Lohr, 1999). In choosing clusters, samplers like to have clusters with smaller homogeneity or higher heterogeneity to alleviate clustering effect. To assess whether PUMA-type PSUs have similar heterogeneity as county-type PSUs, we need to calculate and compare the within-cluster variances between PUMAs and counties. We chose 12 proportion estimates that are available for both counties and PUMAs and calculated the within-cluster variance (McVay, 1947) with an improvement to account for different cluster sizes as shown in formula (1):

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

where *n* is number of clusters, k_i is the number of sampling units within i^{th} cluster, p_i is the proportion estimate in i^{th} cluster, and *K* is the total number of sampling units in all clusters.

2.2 Can PUMA-type PSU Samples Cover Major CBSAs Represented by Certainty County–type PSUs?

To address whether PUMA-type PSU samples will cover CBSAs represented by certainty county–type PSUs, we conducted a simulation study. We followed the 2005 Residential Energy Consumption Survey (RECS) design where county-type PSUs were selected from 19 geographical domains and selected 200 PUMA-type PSUs using the stratified PPS method. The estimated number of HUs were used as the size measure in PPS sampling. We sorted the PUMA-type PSU frame in three ways before selecting PSU samples:

- Sorting Trial 1: By 2005 RECS certainty county status
- Sorting Trial 2: By density defined by total number of HUs/land area
- Sorting Trial 3: By both 2005 certainty county status and density

We repeated 1,000 times and calculated the probabilities of the 20 largest CBSAs being included in the 1,000 PSU samples.

2.3 Will Using PUMA-type PSUs Increase Field Data Collection Costs?

We conducted another simulation study to address the field cost concern. Similar to the simulation study described in Section 2.2, we selected stratified PUMA- and county-type PSU samples, 200 PSUs for each type. Within each selected PSU, four census block groups (CBGs) were selected using PPS sampling method as SSUs. The PSU frames and SSU frames were not sorted before samples were selected. We repeated 1,000 times. We calculated the pairwise distances between CBGs, which measures the distance between centroids of two CBGs. For one PSU and SSU sample, we first calculated pairwise distances for all possible CBG pairs within PSUs and within several distance ranges. We then calculated the mean and percentiles of the pairwise distances. Across 1,000 samples, we calculated the average of mean and average of percentiles and compared them between PUMA-type PSUs and county-type PSUs. Shorter pairwise distances suggest lower field costs.

3. Results and Discussion

The within-cluster variances for 12 proportion estimates at county and PUMA level were calculated and are presented in Table 2 (located at the end of the paper). Among them, six are characteristics of persons and six are characteristics of HUs. For all 12 estimates, the within-cluster variances for PUMAs are smaller than the within-cluster variances for counties. The relative differences vary from -0.19% to -10.49%, and the average relative difference is -3.34%. The results reflect the factor that most PUMAs are geographically smaller than counties. However, the differences are small; thus, we believe that PUMAs have similar heterogeneity as counties.

The probabilities of including the 20 largest CBSAs in the 1,000 PUMA-type PSU samples are displayed in Table 3 (located at the end of the paper). The five largest CBSAs have almost 100% coverage. The average probabilities for the 20 largest CBSAs are 97% for all three sorting scenarios. PUMA-type PSUs cover the major metropolitan areas represented by certainty county–type PSUs very well if county-type PSUs are used.

The simulation results to address field cost concerns are shown in Tables 4 and 5 (located at the end of the paper). Table 4 shows the average CBG pairwise distances within PSUs for mean, 10th percentile, 25th percentile, median, 75th percentile, and 90th percentile across 1,000 PUMA-type and county-type PSU samples. The within-PSU CBG pairwise distances can be used to estimate the field costs as if one field interviewer handles data collection for one PSU. The average CBG pairwise distances for PUMA-type PSUs are shorter than county-type PSUs for the mean and up to the 85th percentile. For the top 15th percentile, PUMA-type PSUs have longer CBG pairwise distances than county-type PSUs. Table 5 shows the average CBG pairwise distances within 10-, 50-, and 70-mile ranges. The average CBG pairwise distances within certain distance ranges provide information on how to efficiently assign work to field interviews across PSUs. As shown in Table 5, PUMA-type PSUs have shorter average CBG pairwise distances than county-type PSUs for all three distance ranges. Results from Tables 4 and 5 suggest that using PUMA-type PSUs has lower field costs than using county-type PSUs. While we think the average pairwise travel distances between CBGs is a fair indication of the relative difference in field data collection costs between County-type PSUs and PUMA-type PSUs, it does not capture specific cost factors related to efficiencies gained by PSUs being in close proximity or SSUs being very far apart. For example, neighboring PSUs could share field staff while spatially large PSUs may require additional staff to limit long travel times. We think more work can be done in this area. A more sophisticated simulation study that can generalize the cost components associated with within and between PSU/SSU travel could help the survey industry better use limited resources.

4. Conclusions

Our studies show that PUMAs have similar within-cluster heterogeneity as counties; thus, using PUMA-type PSUs will have minimum impact on increasing intracluster correlation coefficient, although most PUMAs are geographically smaller than counties. The simulation studies indicate that PUMA-type PSUs have a very good coverage of major CBSAs represented by certainty county PSUs, and using PUMA-type PSUs likely reduces field data collection costs, or at worst are cost neutral compared to using county-type PSUs. In addition, there are several benefits and advantages of using PUMA-type PSUs: PUMAs can be readily used, PUMAs have smaller variation in size measures, and rich information in the ACS microdata at PUMA level can be used to improve the sample design and survey estimates. Thus, we conclude that using PUMAs as PSUs for in-person household surveys is a viable option survey planners need to consider. We have employed PUMA-type PSUs for the 2015 RECS and FDA Tobacco User Panel Survey.

Acknowledgment

Part of this work was funded by Energy Information Administration (EIA), Department of Energy, under Contract No. DE-EI-0000515. The authors would like to thank our colleagues, Dr. Rachel Harter and Dr. Akhil Vaish, for help and advice with this paper.

Disclaimer

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

References

Kish, L. 1965. *Survey sampling*. John Wiley & Sons, Inc.

Lohr, S. L. 1999. *Sampling: Design and analysis*. Duxbury Press.

McVay, F. E. 1947. Sampling methods applied to estimating numbers of commercial orchards in the commercial peach area. *Journal of the American Statistical Association, 42*(240), 533-540.

U.S. Census. 2010. *Final Public Use Microdata Area (PUMA) criteria and guidelines for the 2010 Census and the American Community Survey*. Available at http://www2.census.gov/geo/pdfs/reference/puma/2010_puma_guidelines.pdf

Table 2. Comparison of Within-Cluster Variance Between PUMA-type PSUs and County-type PSUs for Proportion Estimates

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 Bachelor’s 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%

Table 3. Probabilities of 20 Largest CBSAs Being Included in the 1,000 PSU Samples

CBSA	Number of Counties	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

Table 4. Average CBG Pairwise Travel Distances Within PSUs (miles)

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

Table 5. Average CBG Pairwise Travel Distances With Distance Ranges (miles)

	Within 10 Miles		**Within 50 Miles**		**Within 70 Miles**	
Statistics	County-type PSU	PUMA-type PSU	County-type PSU	PUMA-type PSU	County-type PSU	PUMA-type PSU
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