2019 CE Survey Microdata Users' Workshop

Sampling Methods and Derivation of Sampling Weights

Brian T. Nix

Consumer Expenditure Survey
Statistical Methods Division
Bureau of Labor Statistics

July 18, 2019

Overview

- History and Concepts
- Sample Selection
  - Define PSUs
  - Stratify and Select a Sample of PSUs
  - Stratify and Select a Sample of Households
- Weighting the Households

History of Sample Redesigns

- New sample of geographic areas selected every decade
  - **1980 Census-Based Sample Design (1986–1995)**
  - **1990 Census-Based Sample Design (1996–2004)**
  - **2000 Census-Based Sample Design (2005–2014)**
  - **2010 Census-Based Sample Design (2015–2024)**
  - **2020 Census-Based Sample Design (2025–2034???)**

Concepts

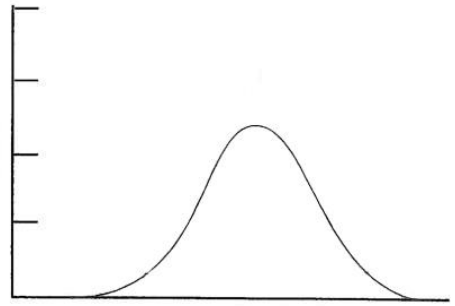
- Target Population
  - U.S. non-institutional civilian population
- PSU (Primary Sampling Unit)
  - Geographic area used for sampling
  - Cluster of contiguous counties
    - (between 2 and 5 counties on average)
- CU (Consumer Unit)

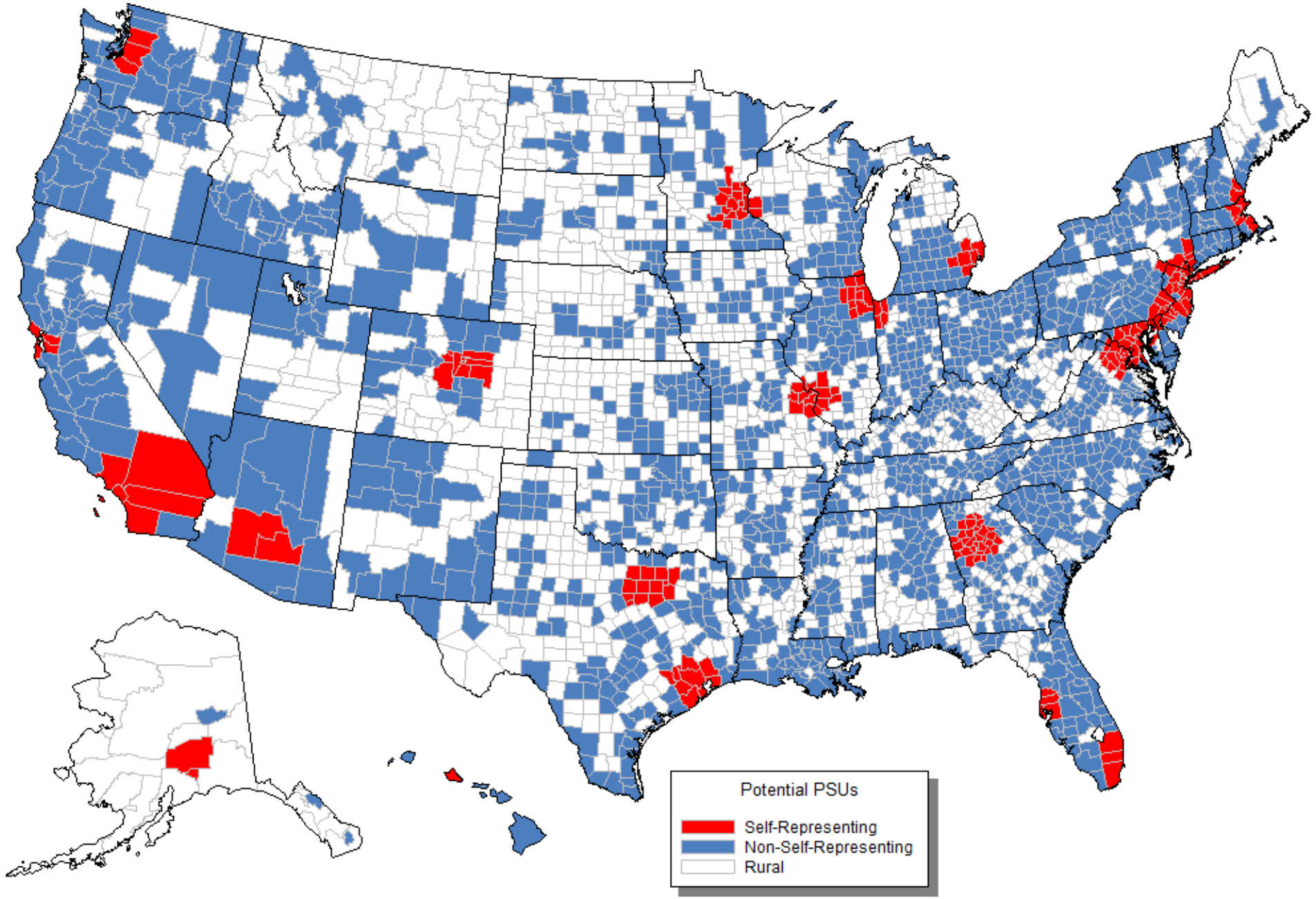
≈ Household

Sample Selection – Overview

- Geographic areas are randomly selected to represent the total U.S.
- Households are randomly selected to represent the geographic areas
- Guiding principle:

  *"**Randomness ensures representativeness**."*

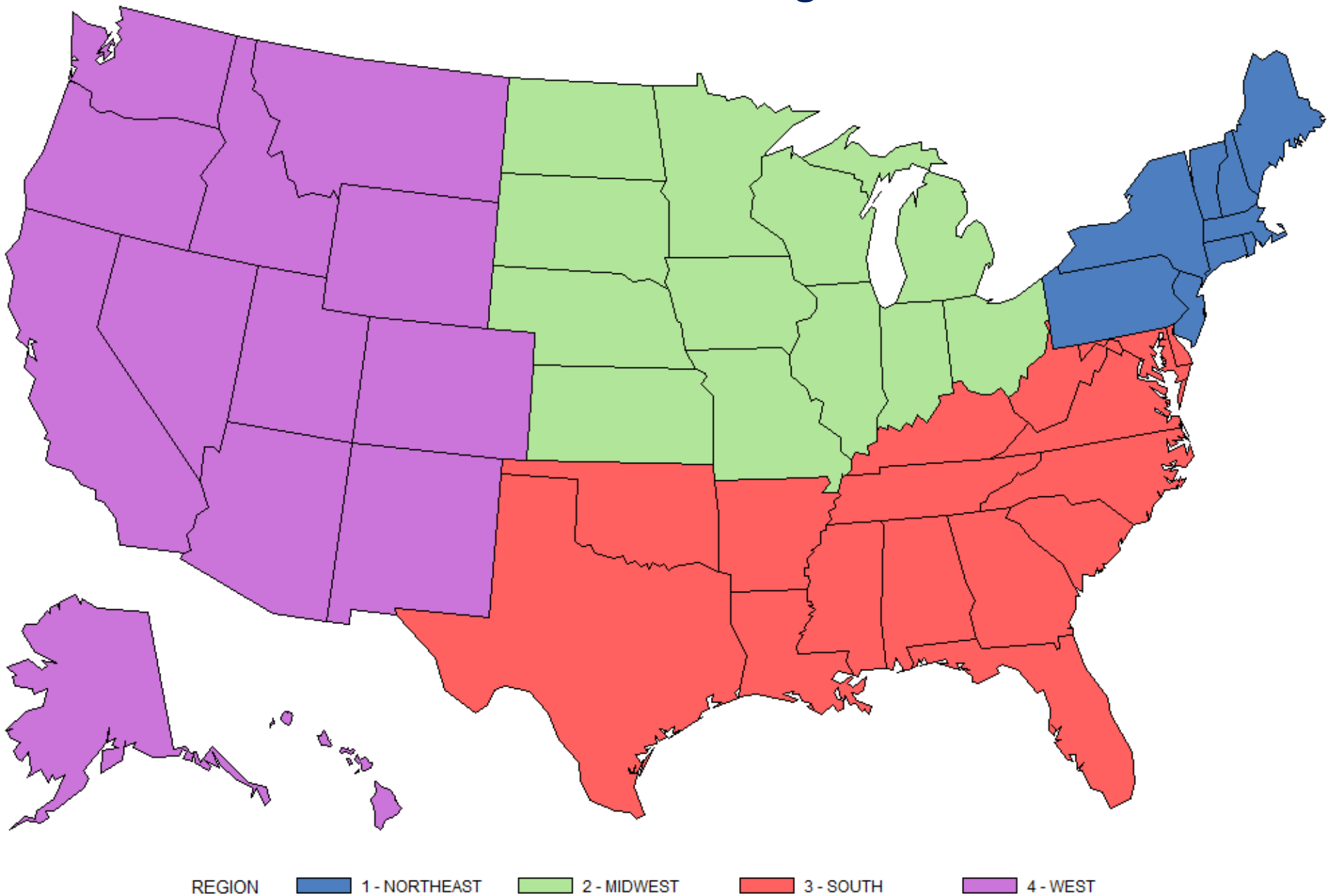



Potential PSUs
Self-Representing
Non-Self-Representing
Rural

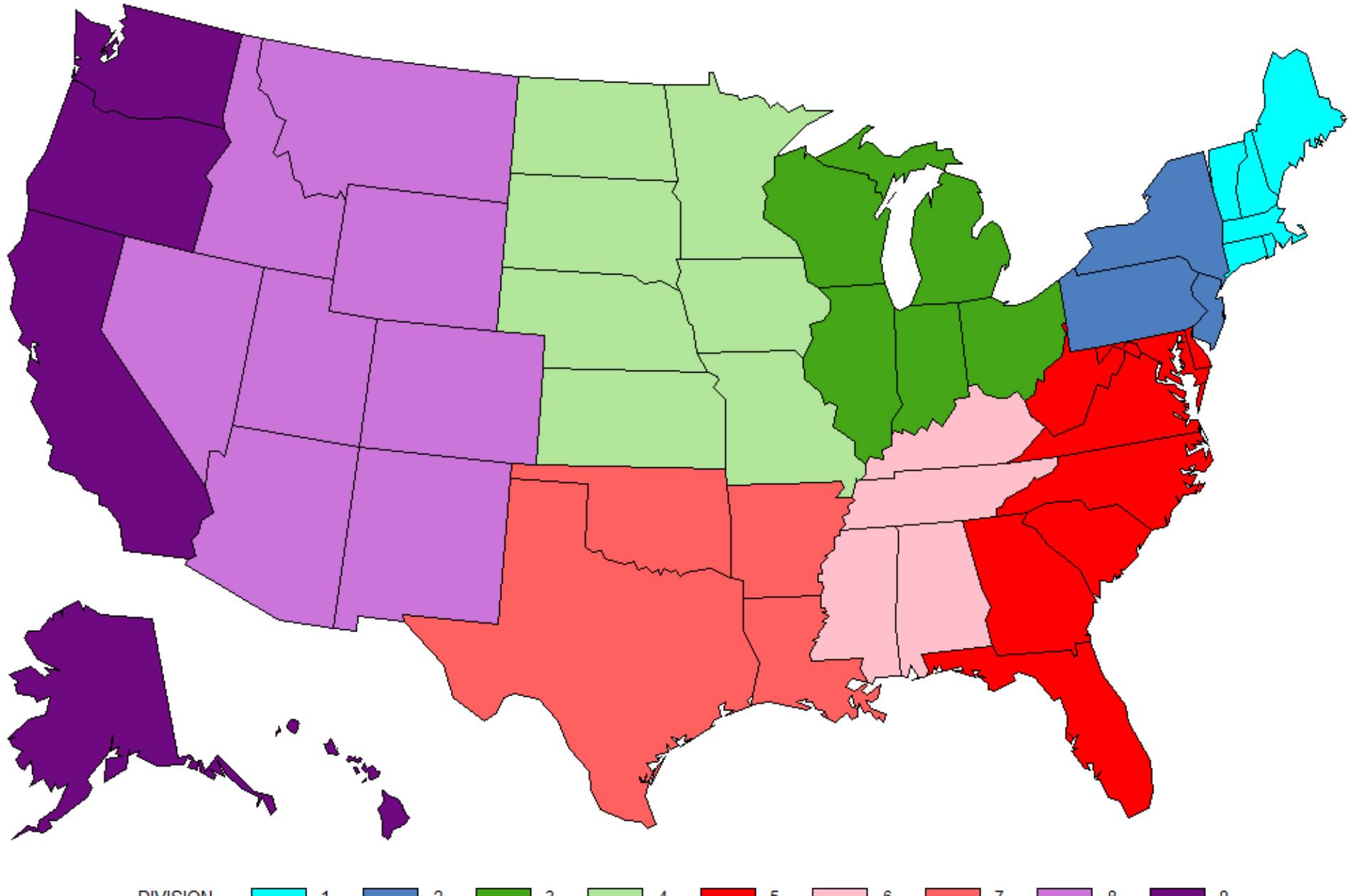
Selection of PSUs

PSU class	Description	CBSA/ Non-CBSA	Population Total	Examples
S	Self-Representing	Metropolitan (urban)	Greater Than or Equal to 2,500,000	S11A Boston MA S49D Seattle WA
N	Non-Self-Representing	Metro- or Micropolitan (urban)	Less Than 2,500,000	*Suppressed*
R	Rural *(also not Self-Representing)*	Non-CBSA (rural)		*Suppressed*

The Four Census Regions



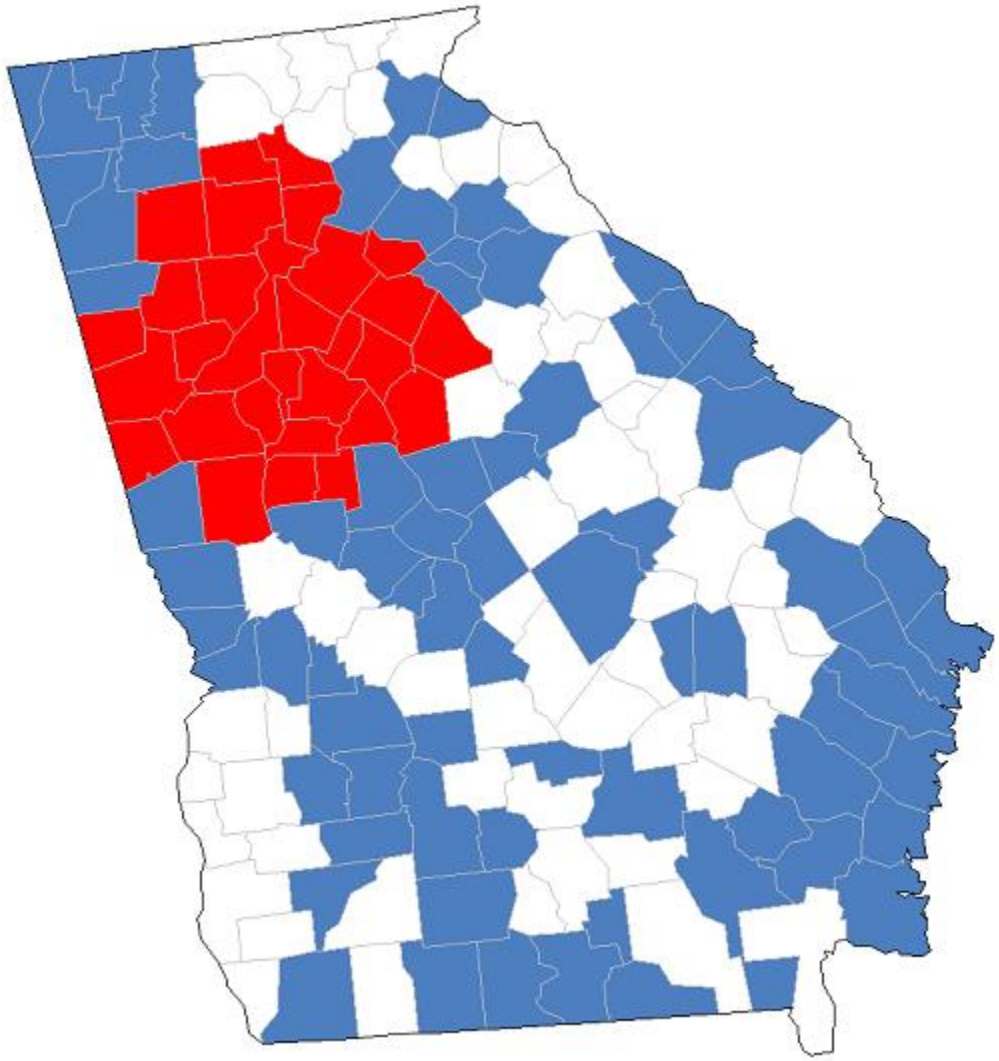
The Nine Census Divisions



Sample Selection:
CPI – 75 PSUs; CE – 91 PSUs

PSU Size	Region/Division									Total
	Northeast		Midwest		South			West		
	01	02	03	04	05	06	07	08	09	
S	1	2	2	2	5	0	2	2	7	23
N	2	4	8	4	12	6	8	4	4	52
R	1	1	2	2	2	2	2	3	1	16
Total	4	7	12	8	19	8	12	9	12	91

Hypothetical PSU Selection



Hypothetical PSU Selection (continued)

	CBSA	2010 Population	Probability of Selection
✓	Augusta, GA-SC	564,873	0.9221
	Jessup, GA	30,099	0.0491
	Fitzgerald, GA	17,634	0.0288
	Total	**612,606**	**1.0000**

	CBSA	2010 Population	Probability of Selection
	Columbus, GA-AL	294,865	0.4783
	Valdosta, GA	139,588	0.2264
✓	LaGrange, GA	67,044	0.1088
	Moultrie, GA	45,498	0.0738
	Douglas, GA	42,356	0.0687
	Thomaston, GA	27,153	0.0440
	Total	**616,504**	**1.0000**

Number of Addresses

- **Local Target Sample Size**
  - Allocate 12,000 addresses in each survey to individual PSUs, proportional to each stratum's population
  - Minimizes CE's nationwide variance

Number of Addresses (continued)

Given the values of p_i and r_i for every index area i, find the values of n_i that	
Minimize	$\sum_{i=1}^{91} \left(\frac{n_i r_i}{NR} - \frac{p_i}{p} \right)^2$
Subject to:	$\sum_{i=1}^{91} n_i = 12{,}000$
	$n_i \geq 0$, for $i = 1$ *to* 91

Number of Addresses (continued)

where

- p_i = population of the *i*-th index area;
- r_i = productivity rate (eligibility rate times the response rate) of the *i*-th index area;
- n_i = number of addresses allocated to *i*-th index area;
- p = $\sum_{i=1}^{41} p_i$ is the population of the United States;
- $n_i r_i$ = expected number of interviewed households in the i-th index area;
- NR = $\sum_{i \in USA} n_i r_i$ is the expected number of interviewed households nationwide.

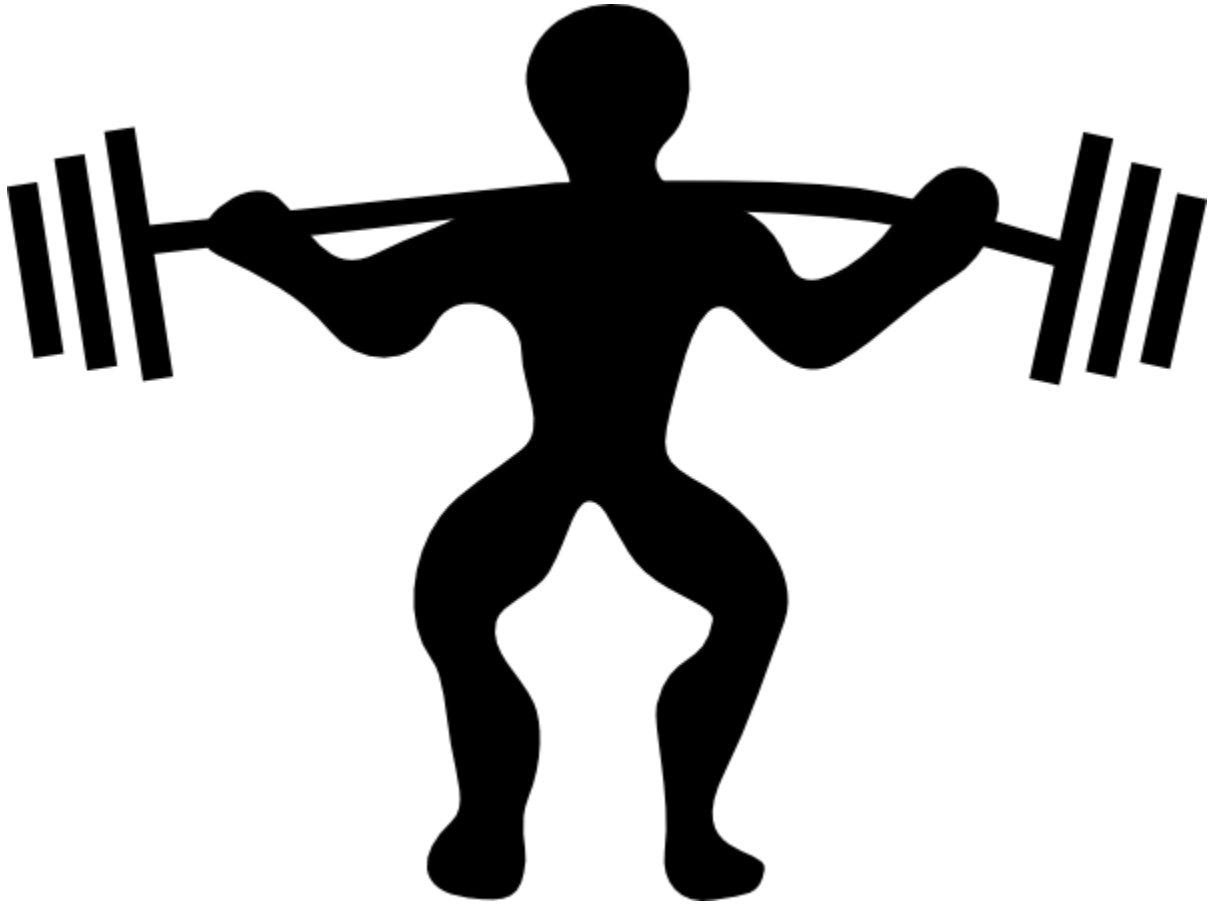
Calculating the Productivity Rate

- 83% **Eligibility** rate
  - (most of the missing 17% are unoccupied)
- 60% **Response** rate
- 50% **Productivity** rate ($0.50 \approx 0.83 \times 0.60$)

Selecting a Random Sample of Households

- Sort households from poor to rich based on information from Decennial Census and ACS
- Compute the sampling interval for each PSU
  - Sampling interval = (# addresses in sampling frame) ÷ (# addresses in CE sample)
- Typical sampling intervals:
  - Every 1,000th address (N and R PSUs)
  - Every 5,000th address (S PSUs)

Weighting Process



Weighting Process

- Base Weight (~10,000)
  - Household + 9,999 others
- Non-interview Adjustment Factor (~1.75)
  - Type A: Refusal to Participate
- Calibration Adjustment Factor
  - Adjusts sample estimate to CPS Totals
  - About 1.15 for Interview Survey

Weighting Process: Calculating the Base Weight

(using hypothetical values)

- PSU Population 538,200
  - MAF counts 224,250 housing units
  - 115 addresses allocated for each survey
  - "Take Every" = 224,250 / 115 ≈ 1,950
- Stratum population 2,800,000
- PSU Weight = 2,800,000 / 538,200 ≈ 5.2025
- Base Weight = "Take Every" * PSU Weight

  ≈ 1,950 * 5.2025 = 10,145

Weighting Process: Calculating the Final Weight

- Variable FINLWT21

- = Base Weight
  - x Non-Interview Adjustment Factor
  - x Calibration Adjustment Factor

- Around 20,000 for Interview Survey, 40,000 for Diary Survey

Conclusion

- Both Sample Design and Weighting Work Together to Produce:
  - Unbiased Estimates of U.S. Expenditures
  - Subject to Allotted CE Budget

Contact Information

Brian T. Nix
Mathematical Statistician
Statistical Methods Division
www.bls.gov/cex
202-691-6877
Nix.Brian@bls.gov