

# 2018 CE Survey Microdata Users' Workshop

# Sampling Methods and Derivation of Sampling Weights

---

**Brian T. Nix**

**Division of Price Statistical Methods  
Bureau of Labor Statistics**

**July 19, 2018**

# 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

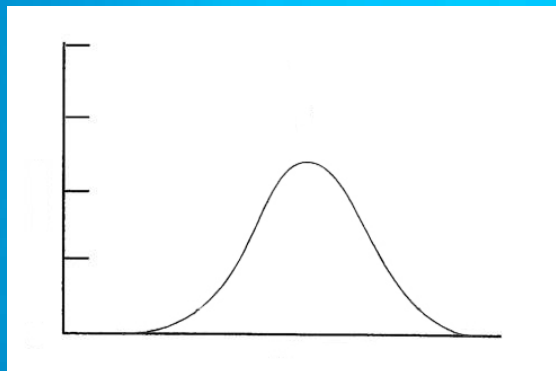
## ■ Sampling Frame

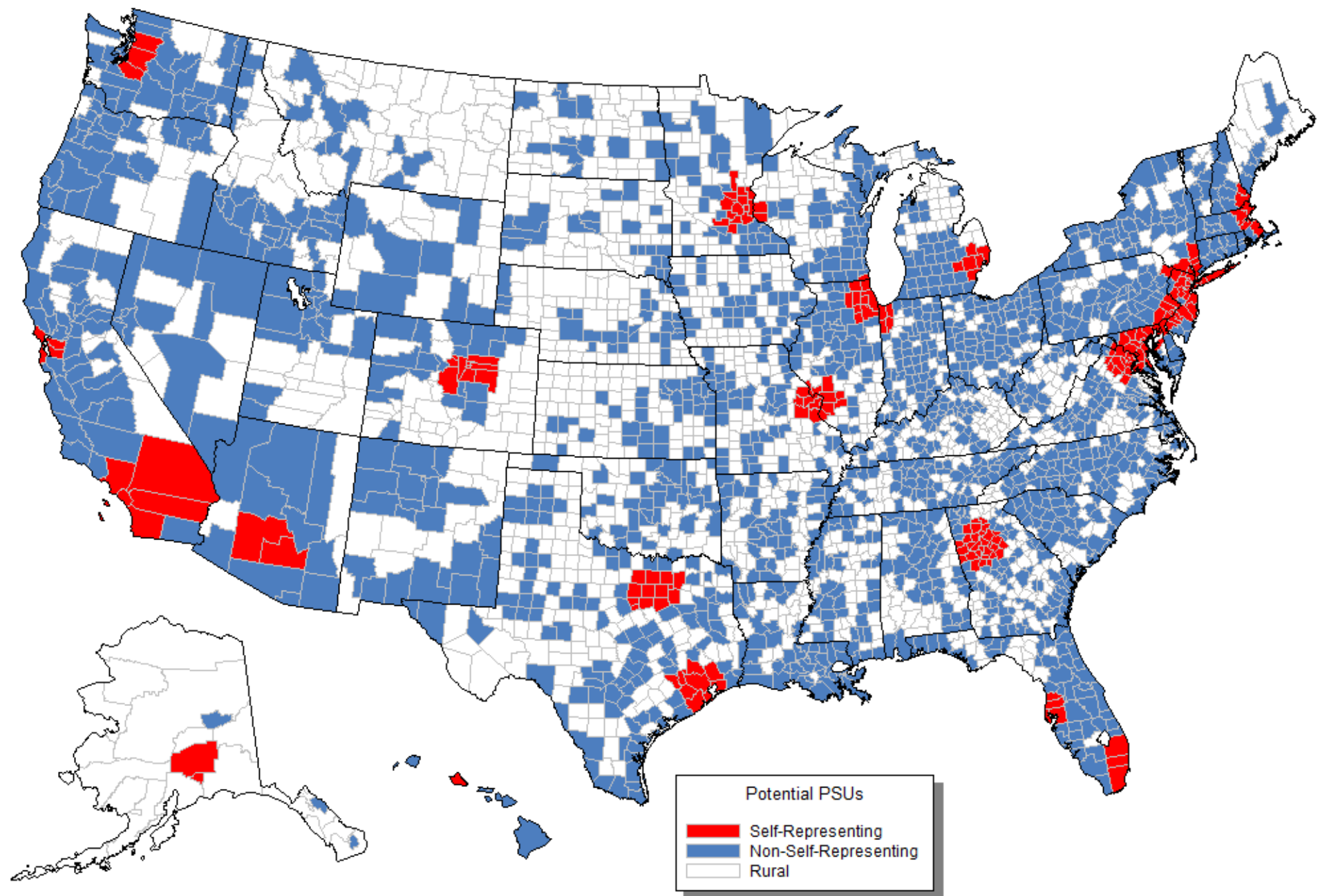
- List of households from which we draw our sample
- Based on 2010 Census (“Master Address File”)
- Biannual updates from U.S. Postal Service (twice a year)

# 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.”*



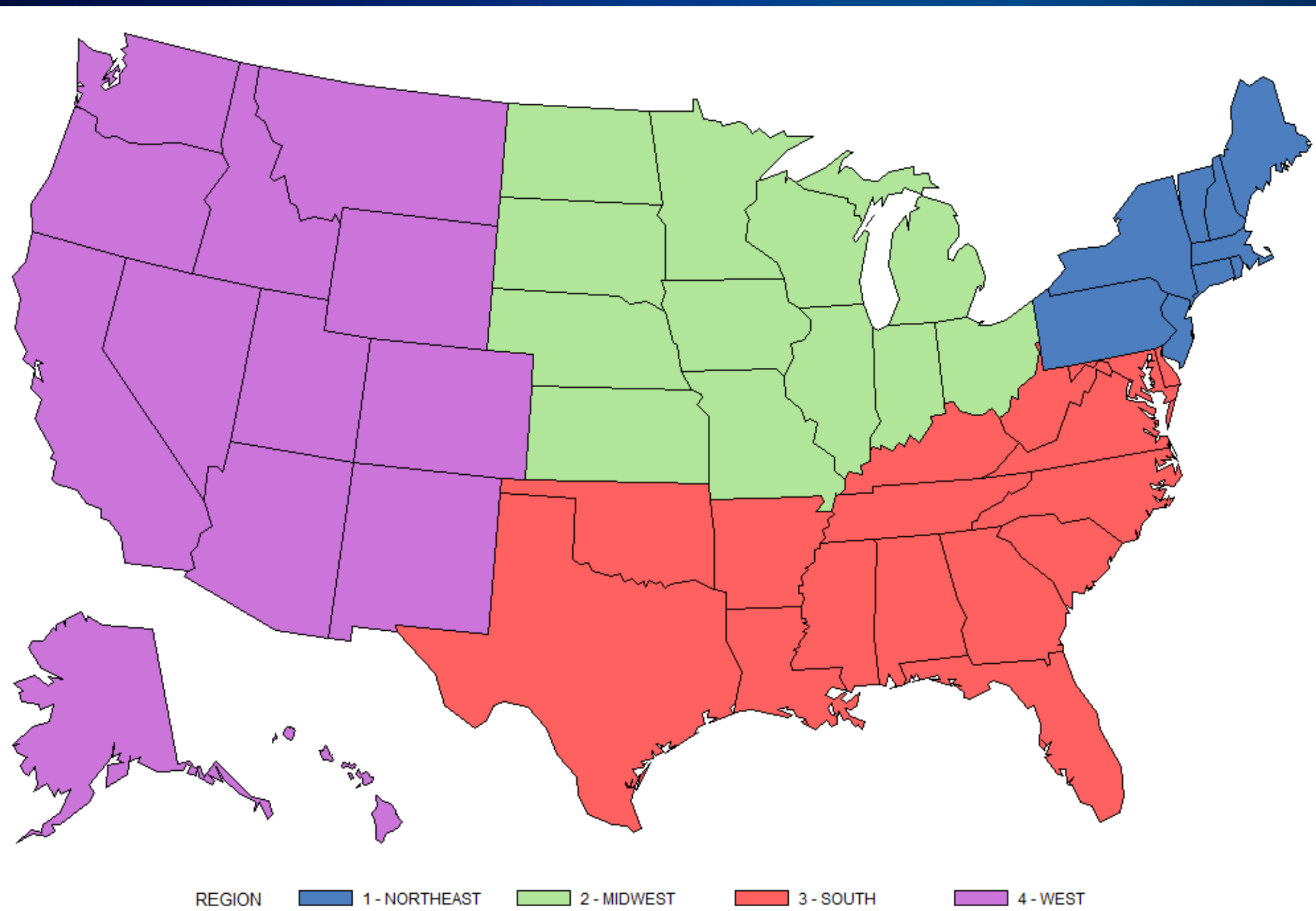


# Selection of PSUs

PSU class	Description	CBSA/ Non-CBSA	Population Total	Examples
S	Self-Representing	Metropolitan (urban)	More Than 2,500,000	S11A Boston MA S49D Seattle WA
N	Non-Self-Representing	Metro- or Micropolitan (urban)	Less Than 2,500,000	<i>Topcoded</i>
R	Rural <i>(also not Self-Representing)</i>	Non-CBSA (rural)		<i>Topcoded</i>

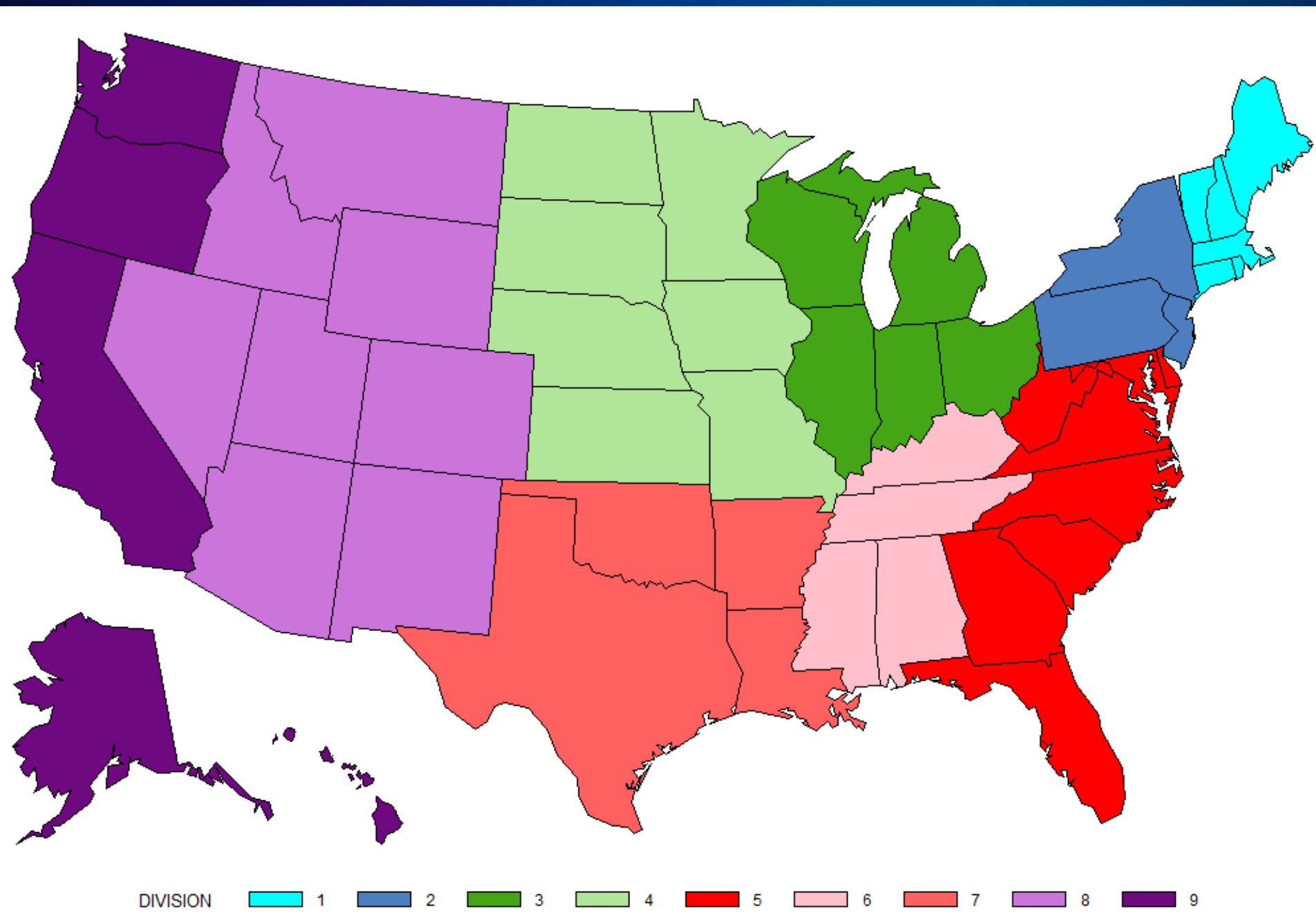


# The Four Census Regions





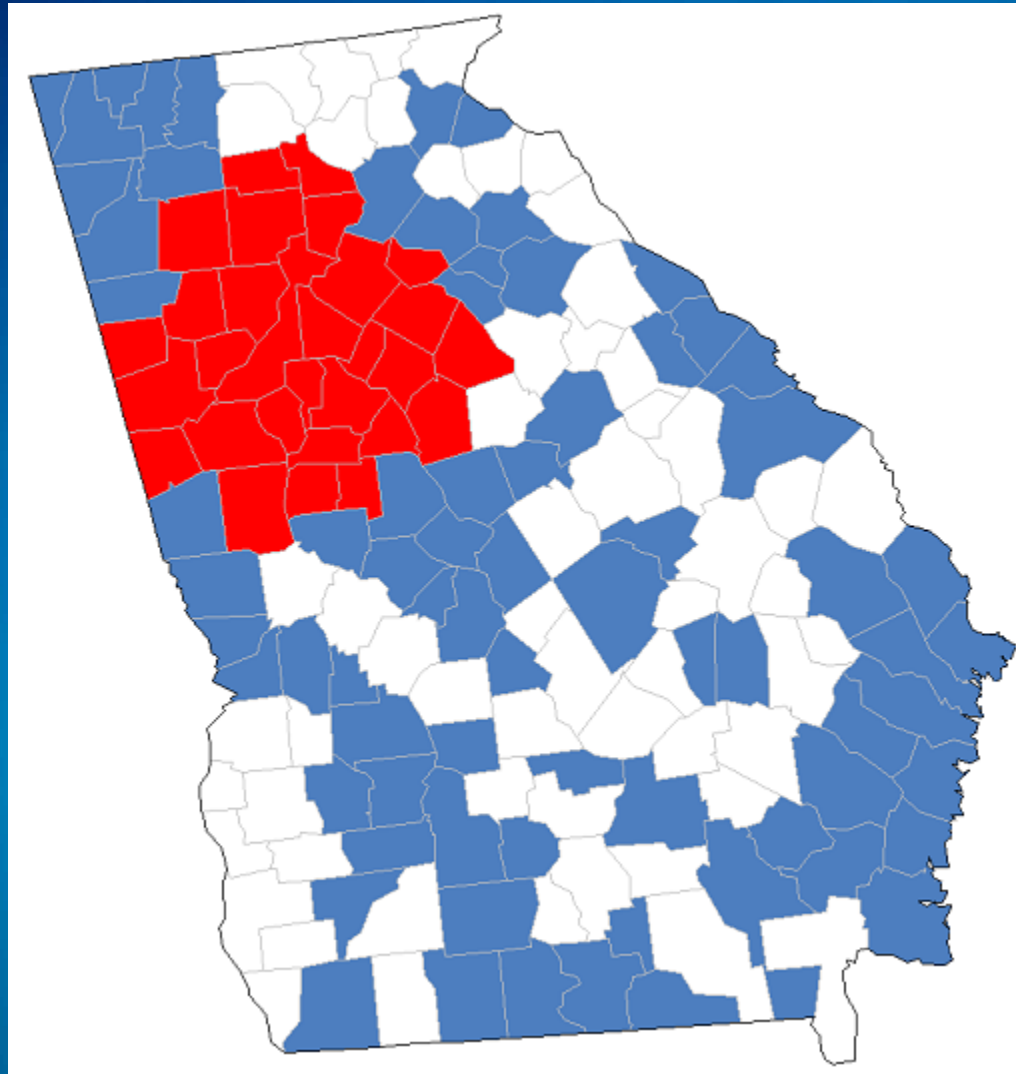
# The Nine Census Divisions



## 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)

<b>CBSA</b>	<b>2010 Population</b>	<b>Probability of Selection</b>
✓ Augusta, GA-SC	564,873	0.92208
Jessup, GA	30,099	0.04913
Fitzgerald, GA	17,634	0.02879
<b>Total</b>	<b>612,606</b>	<b>1.00000</b>

<b>CBSA</b>	<b>2010 Population</b>	<b>Probability of Selection</b>
Columbus, GA-AL	294,865	0.47829
Valdosta, GA	139,588	0.22642
✓ LaGrange, GA	67,044	0.10875
Moultrie, GA	45,498	0.07380
Douglas, GA	42,356	0.06870
Thomaston, GA	27,153	0.04404
<b>Total</b>	<b>616,504</b>	<b>1.00000</b>

# Number of Addresses

- **Local Target Sample Size**

- Allocate 12,000 addresses 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:

$$\sum_{i=1}^{91} n_i = 12,000$$

$$n_i \geq 0, \text{ for } i = 1 \text{ to } 91$$

# Translate Addresses into Interviewed Households

- 83% “eligibility” rate: (most of the missing 17% are unoccupied)
- 60% response rate
- 50% “participation” rate ( $0.50 \approx 0.83 \times 0.60$ )

# Number of Households from Number of Addresses

<u>PSU</u>	<u>Population</u>	<u>Addresses</u>	<u>Interviewed Households</u>
S11A Boston	4.6 million	193	80
S12A New York City	19.6 million	770	389
S12B Philadelphia	6.0 million	191	109
S35A Washington, DC	5.6 million	188	103
S35C Atlanta	5.3 million	170	95
<u>etc.</u>		<u>etc.</u>	<u>etc.</u>
<b>Total</b>	<b>308 million</b>	<b>12,000</b>	<b>6,350</b>



# Select a Random Sample of Households (Mechanics)

- **Sort households from poor to rich based on information from Decennial Census and ACS:**
  - Number of people in household
  - Tenure (owner, renter)
  - Market value of home (owners)
  - Monthly rent (renters)

# Select a Random Sample of Households (Continued)

- 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
- **Weighting Control Factor (~1.00)**  
Apartment Building instead of a House
- **Non-interview Adjustment Factor (~1.50)**  
Type A: Refusal to Participate
- **Calibration Adjustment Factor (~1.15)**  
Adjusts sample estimate to CPS Totals

# Base Weight

## A Hypothetical Example: (Non-Self-Representing PSU)

- PSU Population 538,200
  - MAF counts 224,250 housing units
  - 115 addresses allocated for each survey
  - “Take Every” =  $224,250 / 115 \approx 1,950$
- Stratum population 2,800,000
- PSU Weight =  $2,800,000 / 538,200 \approx 5.2025$
- Base Weight = “Take Every” \* PSU Weight  
 $\approx 1,950 * 5.2025 = 10,145$

# Final Weight

- **Variable FINLWT21**
- **= Base Weight**
  - x Weighting Control Factor**
  - x Non-Interview Adjustment Factor**
  - x Calibration Adjustment Factor**
- **~15,000 to 20,000**

# Conclusion

**Both Sample Design and Weighting  
Work Together to Produce:**

- Best Estimates of U.S. Expenditures
- Subject to Allotted CE Budget

# Contact Information

---

**Brian T. Nix**

**Mathematical Statistician  
Statistical Methods Division**

***[www.bls.gov/cex](http://www.bls.gov/cex)***

**202-691-6877**

**Nix.Brian@bls.gov**