2017 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 20, 2017

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 (CUs)**

History of Sample Redesigns

- **New sample of geographic areas and addresses selected <u>every decade (2010)</u>**
  - 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?)**

Concepts

- Target Population:

  U.S. non-institutional civilian population

- Consumer Unit
  - person or a group of persons in a household related by blood, marriage, adoption, or other legal arrangements
  - OR who are unrelated but pool their incomes to make joint expenditure decisions
  - Same as households approximately 98% of time

Concepts (continued)

- Old (pre-2010 census) Sampling Frame:

  List of Households from which we draw our sample
  - Unit Frame: Regular households (80%)
  - Area Frame: Rural households (10%)
  - Permit Frame: New construction (9%)
  - Group Quarters: (<1%)

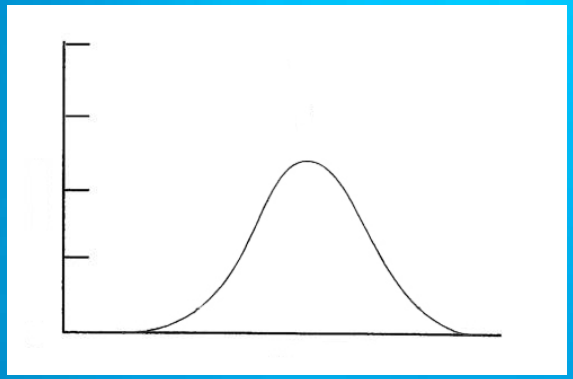
- Since 2015: Census Master Address File (MAF)
  - (based on 2010 Census, with biannual updates from the United States Postal Service)
  - Group Quarters (<1%)

Sampling Process

Sample Selection – Overview

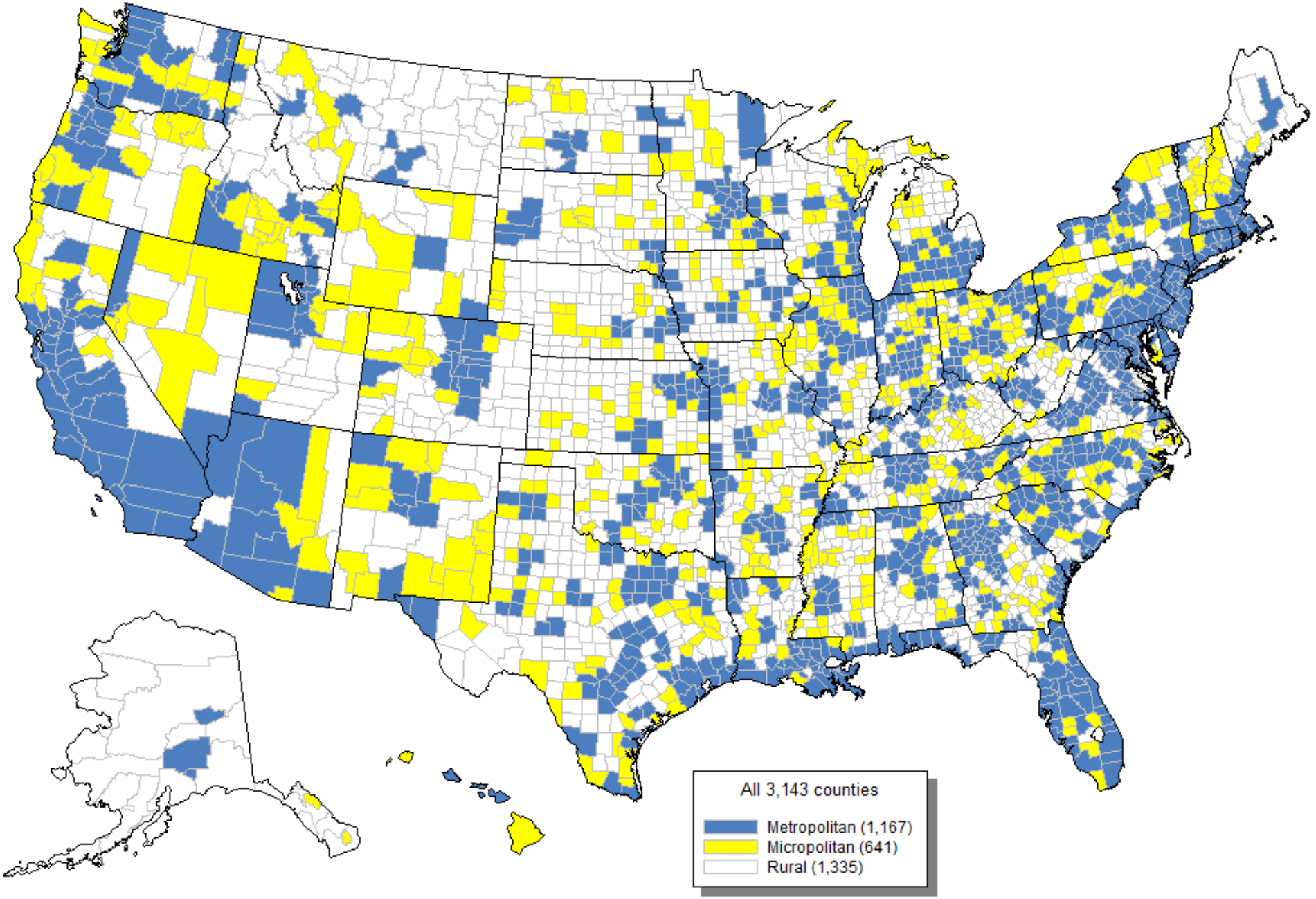
- Geographic areas are randomly selected to represent the total U.S.
- Households are randomly selected to represent the geographic areas
- Guiding principle:

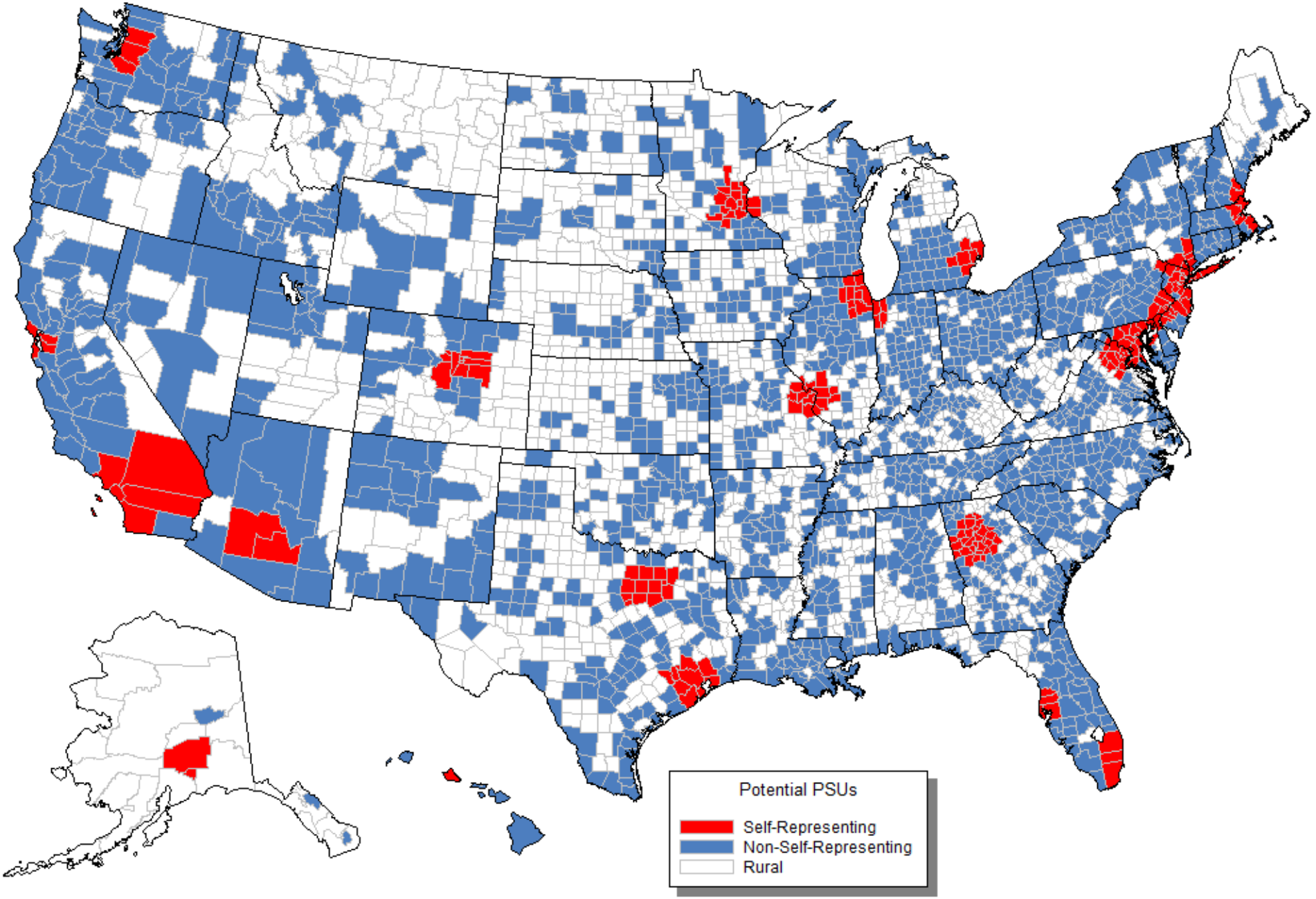
  "Randomness ensures representativeness."



PSU Definitions

- PSU: Primary Sampling Unit
  - Counties are geographically grouped together to become units for sample selection
- CBSA: Core Based Statistical Areas (~old MSA)
  - Counties are grouped together into geographic entities called core based statistical areas (CBSA's) by Office of Management and Budget
  - **Metropolitan** – one or more counties centered around urban area of > 50,000 people
  - **Micropolitan** – one or more counties centered around urban area of 10,000 - 50,000 people
- Over 3,140 counties and county-equivalents in the U.S.
- Over 900 CBSAs defined by OMB


All 3,143 counties
Metropolitan (1,167)
Micropolitan (641)
Rural (1,335)


Potential PSUs
Self-Representing
Non-Self-Representing
Rural

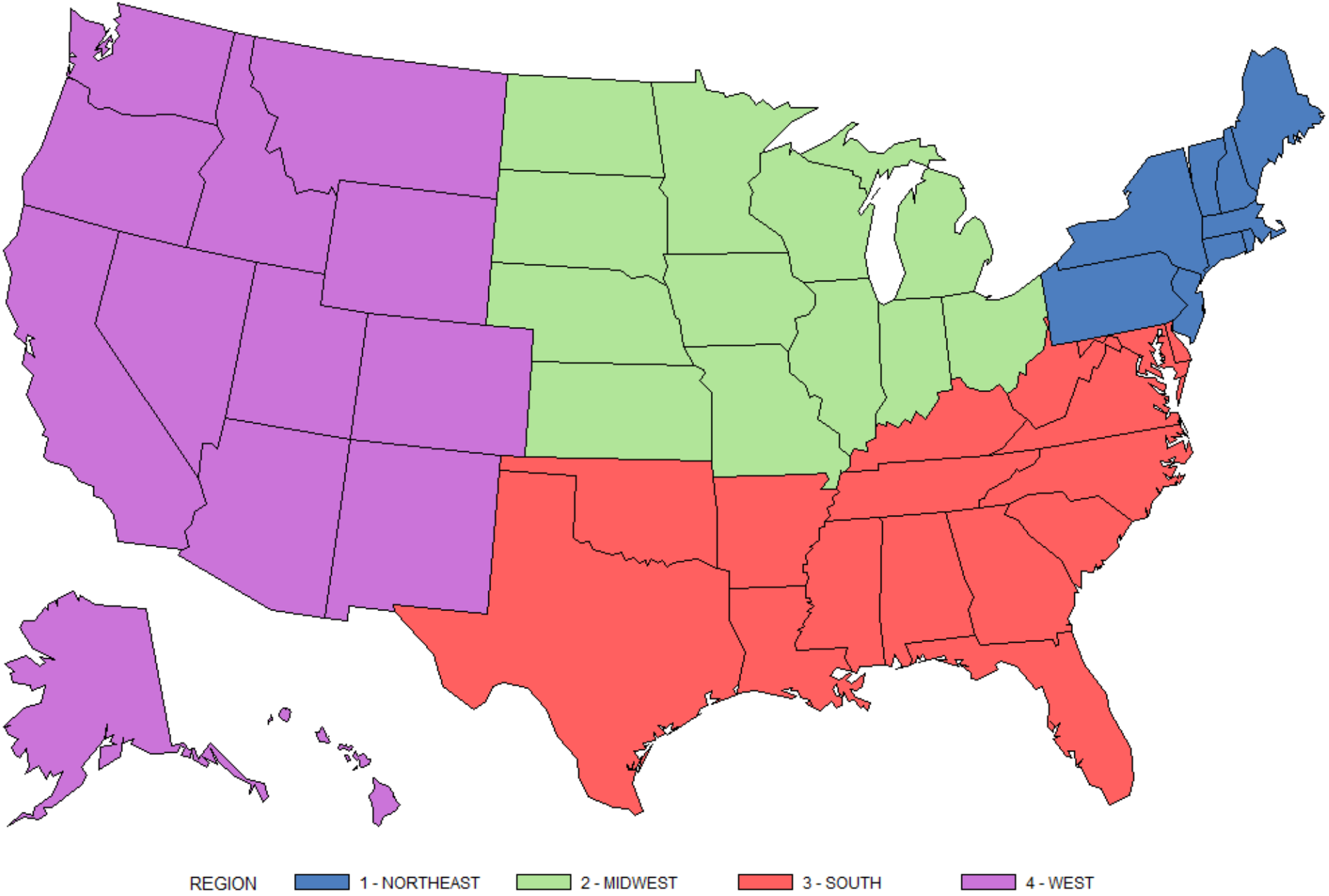
Selection of PSUs (2010-Census Design)

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

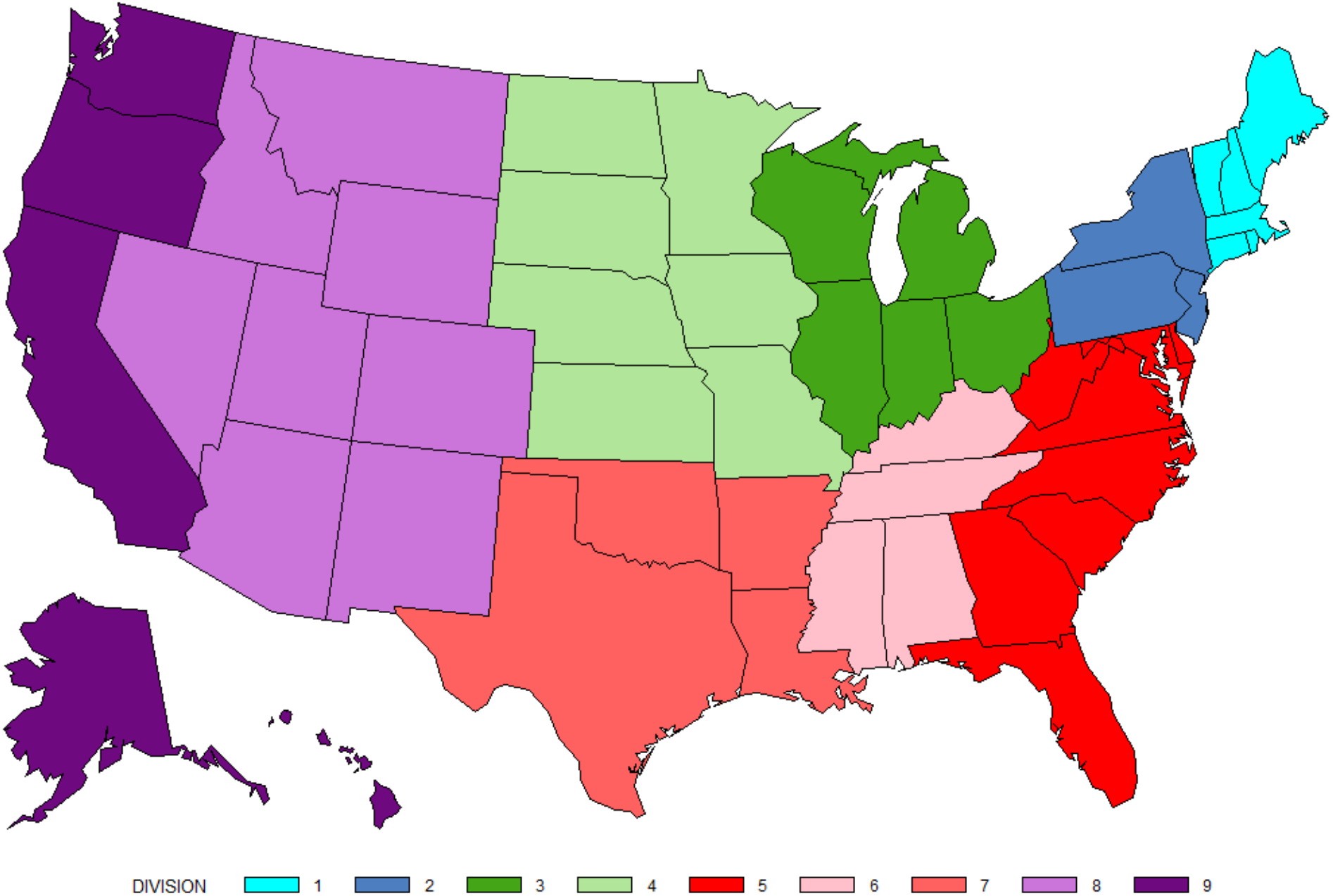
2010 Census-based Sample Selection
CPI – 75 PSUs; CE – 91 PSUs

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

The Four Census Regions



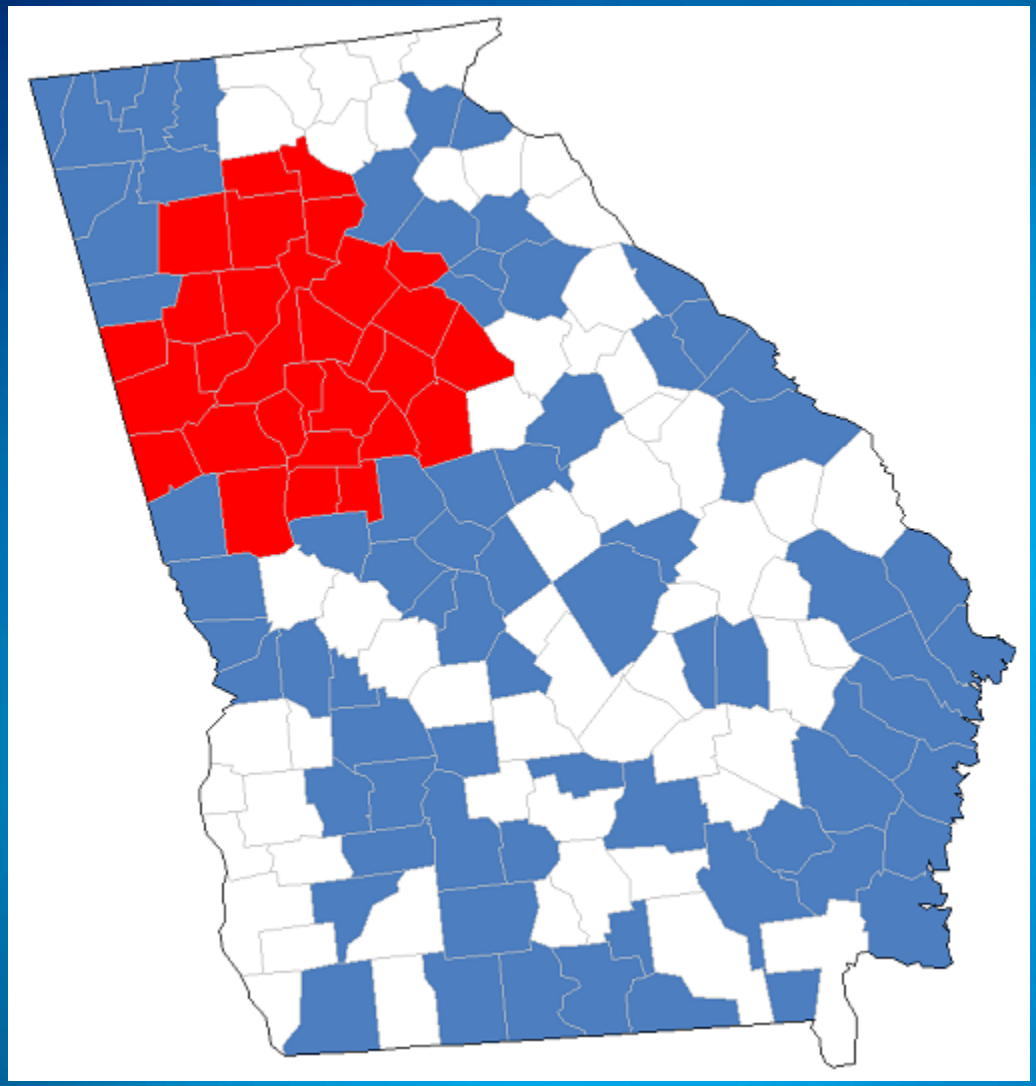
The Nine Census Divisions



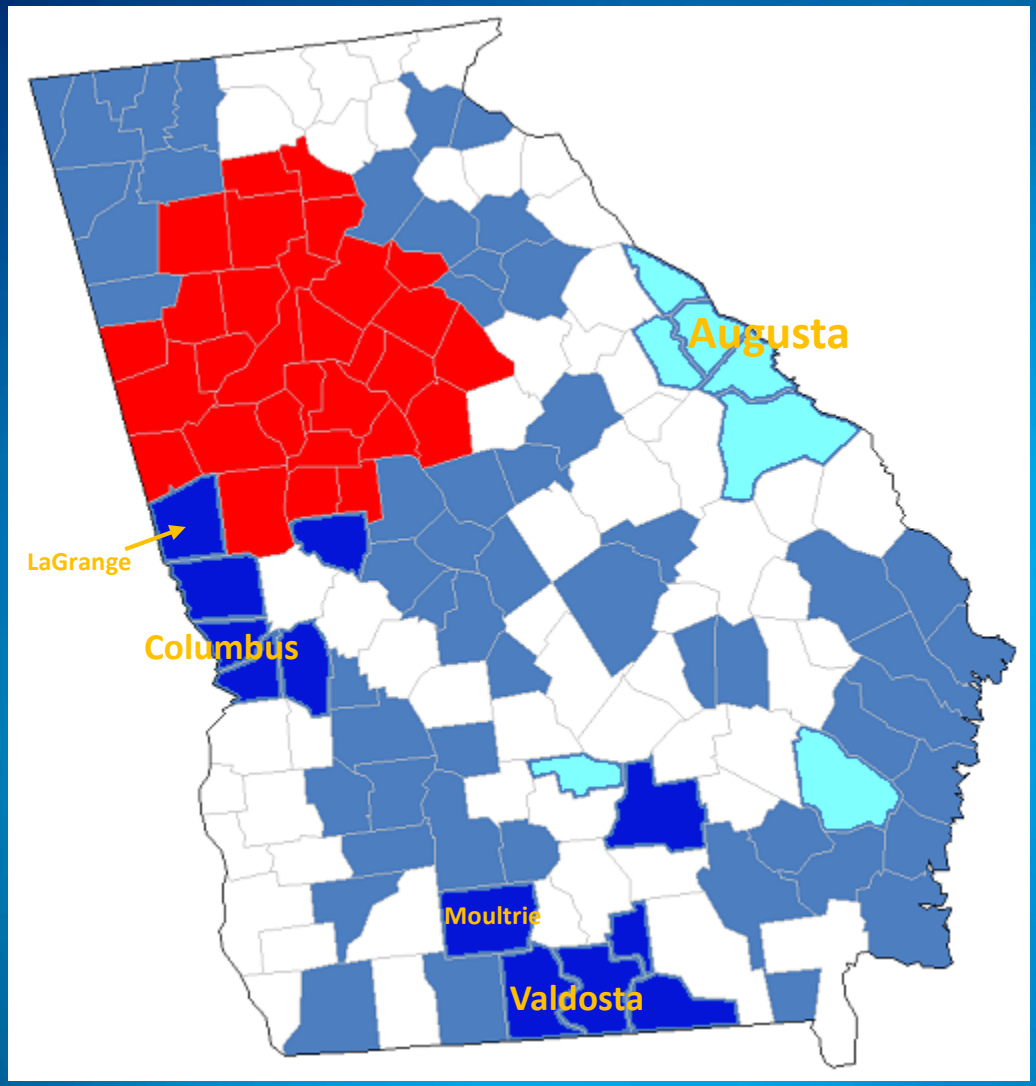
2010 Census-based Sample Selection
CPI – 75 PSUs; CE – 91 PSUs

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



Hypothetical PSU Selection (continued)

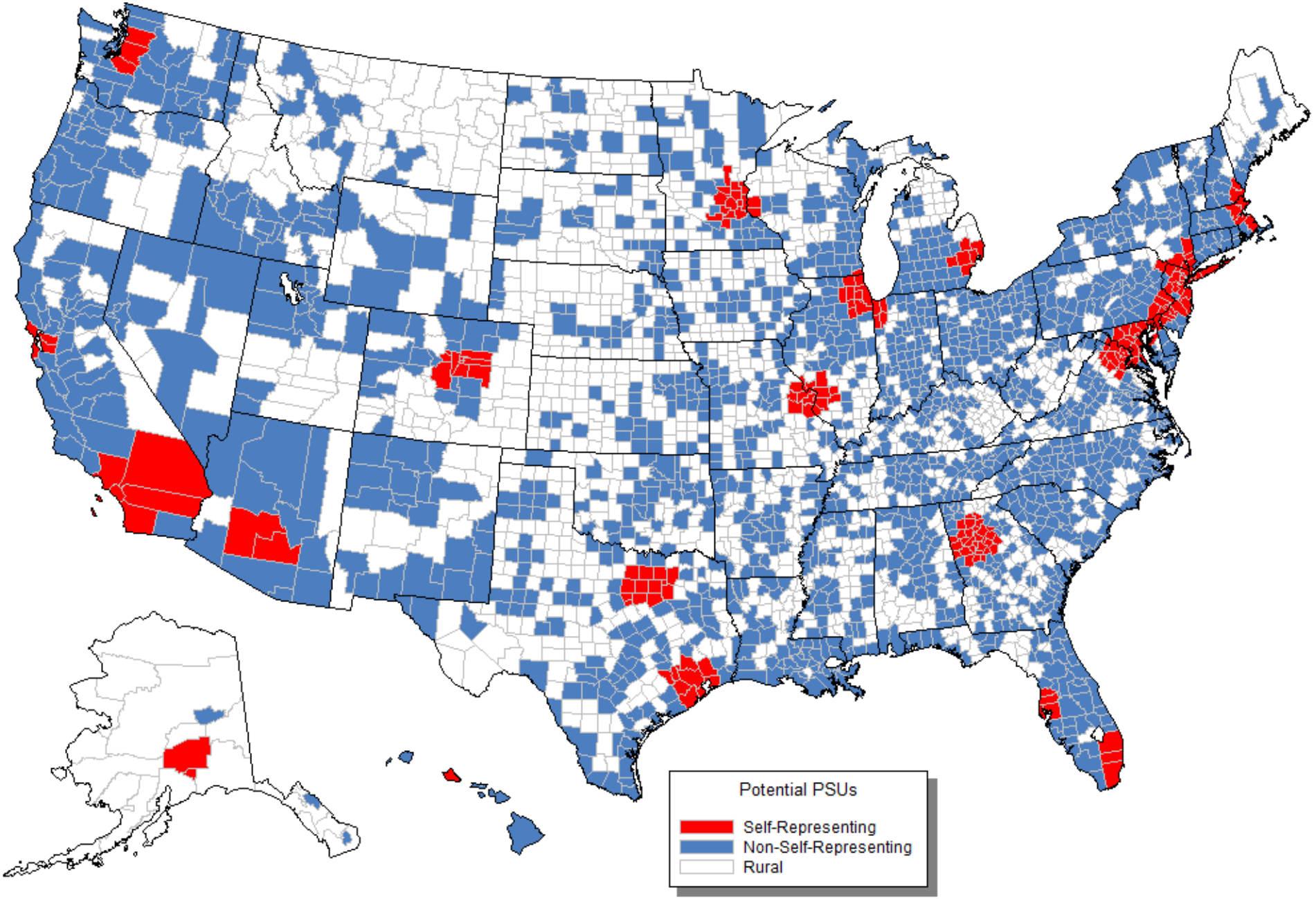
CBSA	2010 Population	Probability of Selection
Augusta, GA-SC	564,873	0.92208
Jessup, GA	30,099	0.04913
Fitzgerald, GA	17,634	0.02879
Total	**612,606**	**1.00000**

CBSA	2010 Population	Probability of Selection
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
Total	**616,504**	**1.00000**

Hypothetical PSU Selection (continued)

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

	CBSA	2010 Population	Probability of Selection
	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
	Total	**616,504**	**1.00000**


Potential PSUs
Self-Representing
Non-Self-Representing
Rural

Number of Households

- **Allocate Target Sample to PSUs**
  - Target size: ~7,000 interviewed households
    - Based on Finite Budget
    - For Diary Survey per year
    - For Interview Survey per quarter
  - 6,600 to Households used jointly by CE and CPI
    - for CPI cost-weight calculations
    - 23 **S**elf-Representing PSUs
    - 52 **N**on-Self-Representing PSUs
  - 400 to CE Households
    - 16 **R**ural PSUs

Number of Households (continued)

- **Target Sample Size**
  - 7,000 interviewed households per year (Diary)
  - 7,000 interviewed households per quarter (Interview, interviews #2-5 only)

- **Target Sample Yield**
  - 14,000 weekly diaries per year (=7,000 x 2)
  - 28,000 quarterly interviews per year (=7,000 x 4)

Number of Households (continued)

- **Local Target Sample Size**
  - Allocate 7,000 interviewed households to individual PSUs, proportional to each stratum's population
  - Minimizes CE's nationwide variance

Translate Addresses into Interviewed Households

- 80% “eligibility” rate: (most of the missing 20% are unoccupied)
- 70% response rate
- 56% “participation” rate (0.56 = 0.80×0.70)

Translate Interviewed Households into Addresses (continued)

PSU		Interviewed households	Addresses	%
S11A	Boston	169	322	52
S12A	New York City	195	286	68
S12B	Philadelphia	220	420	52
S35A	Washington, DC	212	335	63
S35C	Atlanta	182	291	63
	etc.	*etc.*	*etc.*	
Total		**7,000**	**12,000**	

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)

Select a Random Sample of Households (Continued)

- --- D --- I --- D --- I --- D --- I --- D --- I --- D --- I --- D --- I --- *etc.*

- D=Diary, I=Interview

- Each “D” and “I” has enough sample to cover the next 10 years

Weighting Process

Weighting Process

Base Weight Calculation: Real-World Example (Self-Representing PSU)

- S49A (Los Angeles): population 12,828,837
  - MAF counts 4,500,000 housing units
  - 470 addresses needed for each survey
    - based on estimated response rates specific to Los Angeles
    - proportional to 12,000 addresses needed for all PSUs
  - "Take Every" = 4,500,000 / 470 ≈ 9,575
    (every 4,788th address when considering both surveys)
- Stratum population also 12,828,837
  (self-representing PSU)
- PSU Weight = 1 (for any self-representing PSU)
- Base Weight = "Take Every" * PSU Weight
  ≈ 9,575 * 1 ≈ 9,575

Weighting Process

Base Weight Calculation: Hypothetical Example (Non-Self-Representing PSU)

- PSU Population 538,200
  - MAF counts 224,250 housing units
  - 115 addresses needed for each survey
    - based on estimated response rates specific to this PSU
    - proportional to 12,000 addresses needed for all PSUs
  - "Take Every" = 224,250 / 115 ≈ 1,950
    (every 975th address when considering both surveys)
- 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 (Continued)

- Base Weight (~10,000)
  - 9,999 CUs + Self
- 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

Weighting Process (Continued)

- **Final Weight**
  - Variable FINLWT21
  - Base Weight * Weighting Control Factor * Non-interview Adjustment Factor * 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

Any Questions?

Contact Information

Brian T. Nix

Mathematical Statistician

Statistical Methods Division

www.bls.gov/cex

202-691-6877

Nix.Brian@bls.gov