

# Experimental Weights for Estimating State Expenditures Using CE Public Use Microdata

Susan L. King

Mathematical Statistician

OPLC/SMD

Survey Methods Symposium and CE Microdata  
Users' Workshop

July 14, 2016

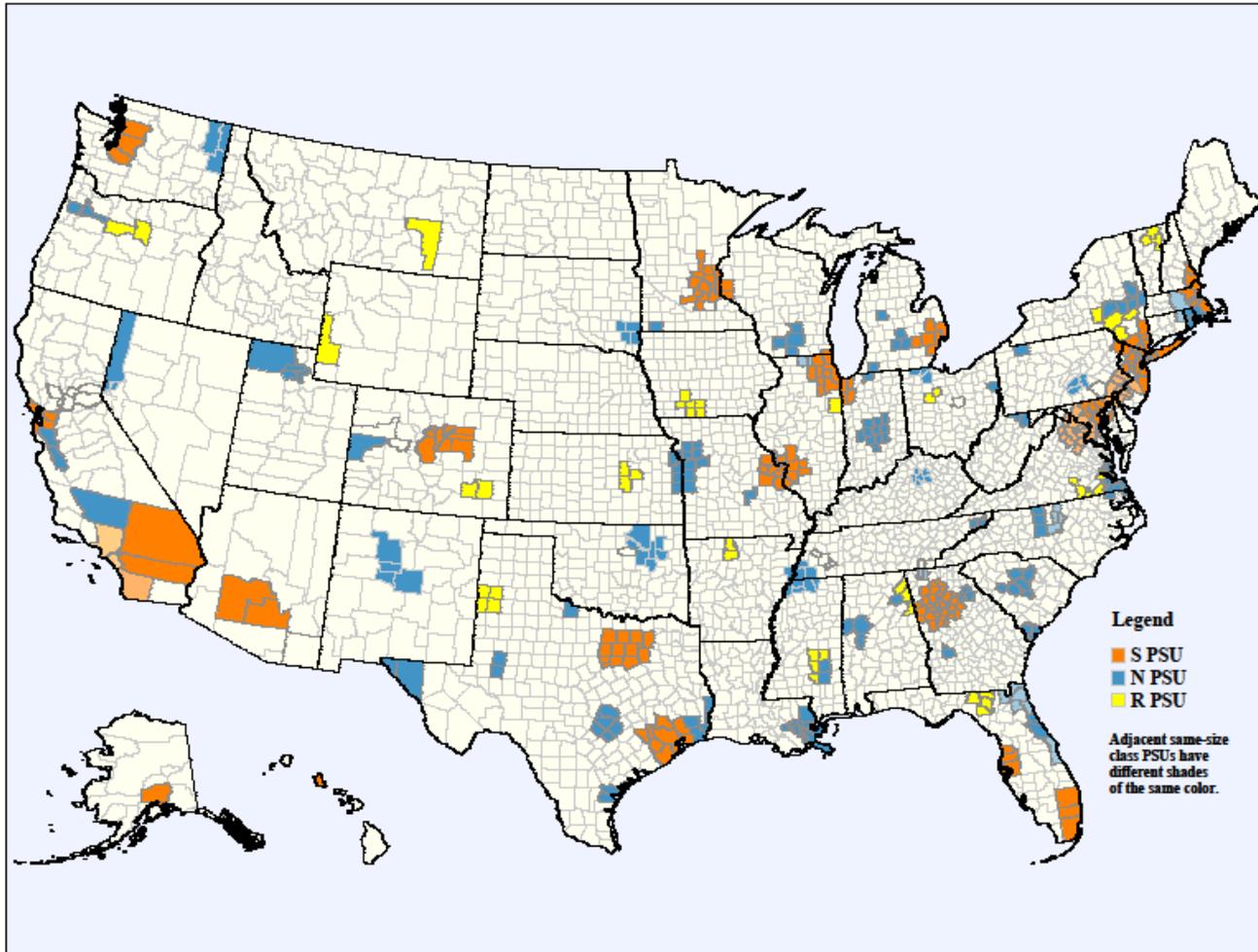


# Objective

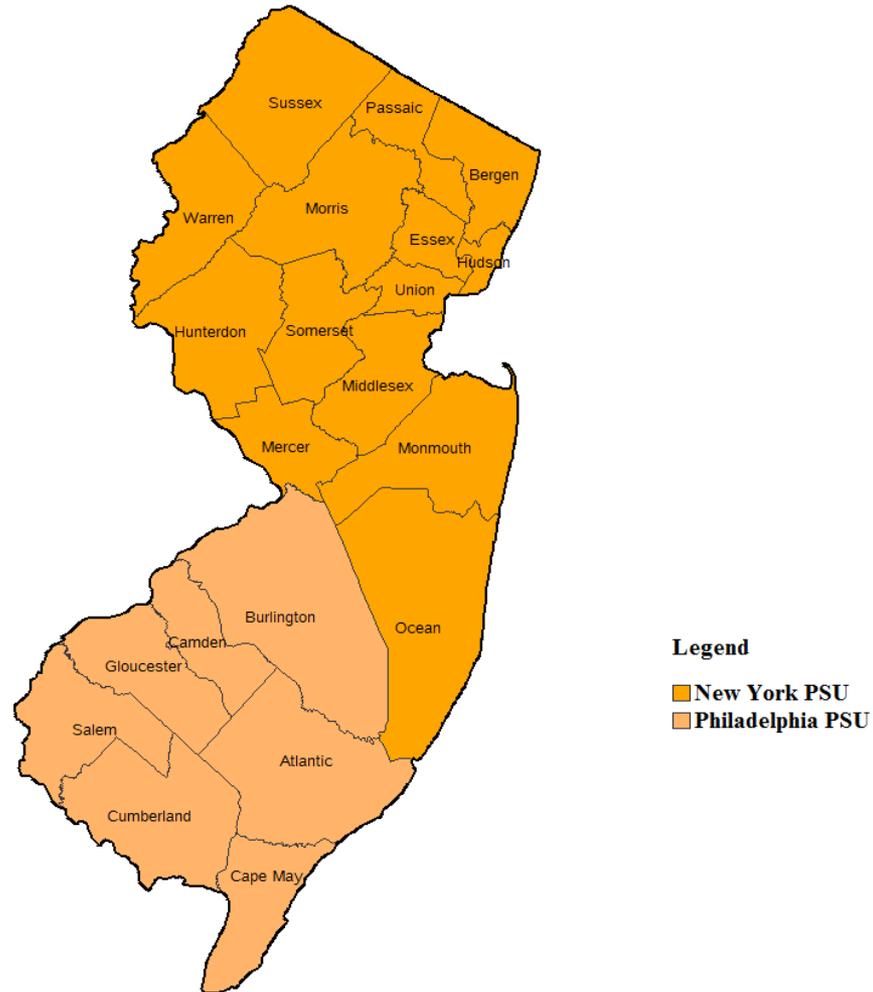
- Provide an extra set of weights on Public Use Microdata to allow the user to make state estimates for their research projects
- Requires different techniques than if CE were to provide state expenditure estimates for selected expenditure categories
- Research in progress



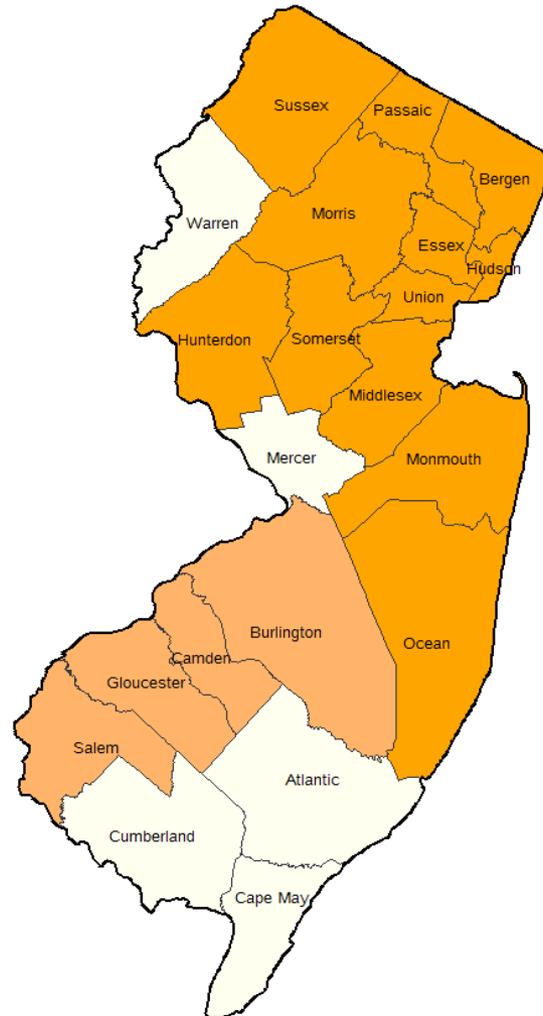
# Simulated Design 2010 CE PSUs



# New Jersey Design 2000



# New Jersey Design 2010



## Legend

- New York PSU
- Philadelphia PSU
- Not Sampled Design 2010

# Road Map

- Use CE data from 2013 and quarter 1 of 2014
- Calculate expenditure estimates using every county in New Jersey, Design 2000
- Drop Warren County , re-weight, calculate expenditure estimates
- Drop South New Jersey, re-weight, calculate expenditure estimates
- Drop all 5 counties, re-weight, calculate expenditure estimates, Design 2010



# PSUs in Stratum X348

	PSU	Population
	Charlotte, NC-SC	1,114,808
	Charleston-North Charleston, SC	549,033
	Fayetteville-Fort Bragg, NC	302,963
✓	<b>Savannah, GA</b>	<b>293,00</b>
	Columbus, GA-AL	274,624
	Naples, FL	251,377
	Gastonia, NC	190,365
	Albany, GA	157,833
	Decatur, AL	145,867
	Warner-Robbins, GA	134,433
	Total	3,121,303



# Diary Base Weights Design 2000

$$CED_{BW} = \frac{1}{\text{probability selecting PSU}} \times (2 \times \text{Within PSU Sampling Interval})$$

For a Self-Representing PSU, Philadelphia

$$\begin{aligned} CED_{BW} &= 1 \times (2 \times 4,110.03) \\ &= 8,220.06 \end{aligned}$$

For a Non-Self Representing PSU, X348

$$\begin{aligned} CED_{BW} &= \frac{3,121,303}{293,000} \times (2 \times 639.38) \\ &= 13,622.52 \end{aligned}$$

# Interview Base Weights Design 2000

$$CEQ_{BW} = CED_{BW} \times \text{CEQ sub sampling factor}$$

For Self Representing PSU Philadelphia

$$\begin{aligned} CEQ_{BW} &= 8220.06 \times 1.0644 \\ &= 8,749.33 \end{aligned}$$

For a Non-Self Representing PSU, X348

$$\begin{aligned} CEQ_{BW} &= 13,622.52 \times 1.0471 \\ &= 14,264.14 \end{aligned}$$

# Other Weight Adjustments

- Control Factor Weight
  - ▶ Adjust for multiple housing units when one housing unit was expected
- Non-Interview Adjustment Weight
  - ▶ Cell collapsing procedure that accounts for refusal to participate
- Calibration Adjustment Weight
  - ▶ Adjusts weights to known population counts

# Variance

- Balanced Repeated Replication (BRR)
- Hadamard Matrix
- 43 Strata (Rows)
  - ▶ PSU/Half Samples are assigned to rows
  - ▶ Balanced by population
- 44 Replicates (Columns)

# Road Map for State Estimates

- Assign Census Tracts in NJ to strata
- Match strata assignment to CE data
- Find selection probabilities for each strata
  - ▶  $\frac{\text{Population of Census Tracts with CE Data in Strata}}{\text{Population of Census Tracts in Strata}}$
  - ▶ Base Weight =  $\frac{1}{\text{Probability of Selection}}$  x Within PSU Sampling Interval
- Calibration
- Variance – Jackknife

# Census Tracts

- Small, relatively permanent, contiguous areas within a County or equivalent entity
- Locally updated before decennial census
- Optimum size 4,000 people but can range from 1,200 to 8,000 people
- Vary in geographical size
- Maintained from Census to Census



# Stratification Objective

- To Minimize Survey Variance
  - ▶ Tracks within each stratification cluster are homogenous with respect to expenditures
  - ▶ Variability between stratification clusters
  - ▶ Stratification cluster populations should be approximately equal ( $\pm 10\%$ )
- Constrained clustering problem – solved using heuristic algorithm

# Variables Homogeneity

- Variables -from 5 year American Community Survey (ACS) estimates, 2010 -2014
  - ▶ Median Census Tract household property value
  - ▶ Median Census Tract household income
- Correlate with expenditures

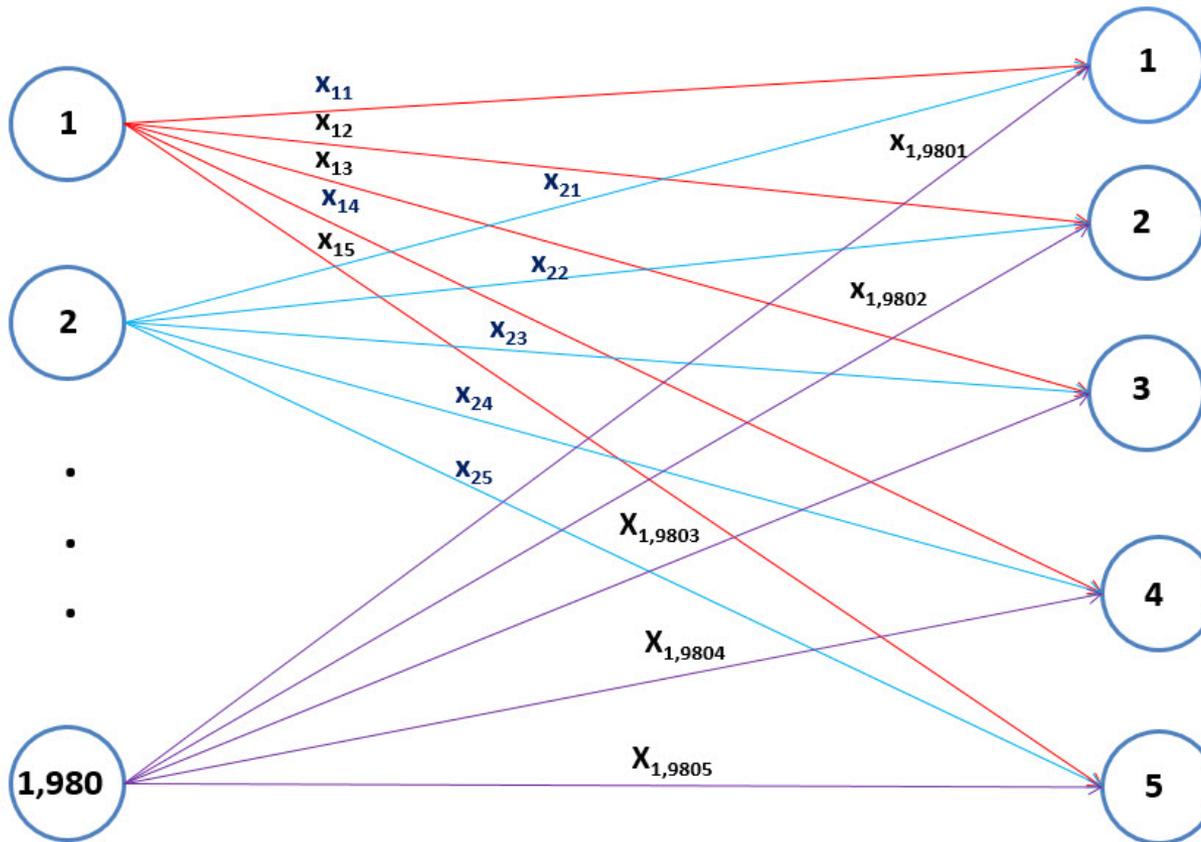
# Step 1 – Relaxed Clustering

- The number of stratification clusters are determined *a priori*—5 stratification clusters
- Standardize the variables
- Ignoring the balanced population constraint, use k-means clustering to assign Census Tracts to stratification clusters (PROC FASTCLUS)
- Cluster centers are used in Step 2

# Step 2-Assignment Model

Census Track

Stratification Cluster



# Step 2 – Euclidean Distances

- Calculate the Euclidean distance between each Census Tract and the stratification cluster center from k-means clustering
- These distances are the objective function coefficients

# Step 2 – Optimization Model

$$\text{Minimize} \quad \sum_{i=1}^{1,980} \sum_{j=1}^5 c_{ij} x_{ij}$$

$$\text{subject to:} \quad \sum_{j=1}^5 x_{ij} = 1 \quad \text{for every } i$$

$$\sum_{i=1}^{1,980} p_i x_{ij} \geq 1,589,693 \quad \text{for every } j$$

$$\sum_{i=1}^{1,980} p_i x_{ij} \leq 1,942,958 \quad \text{for every } j$$

$$x_{ij} = 0 \text{ or } 1$$

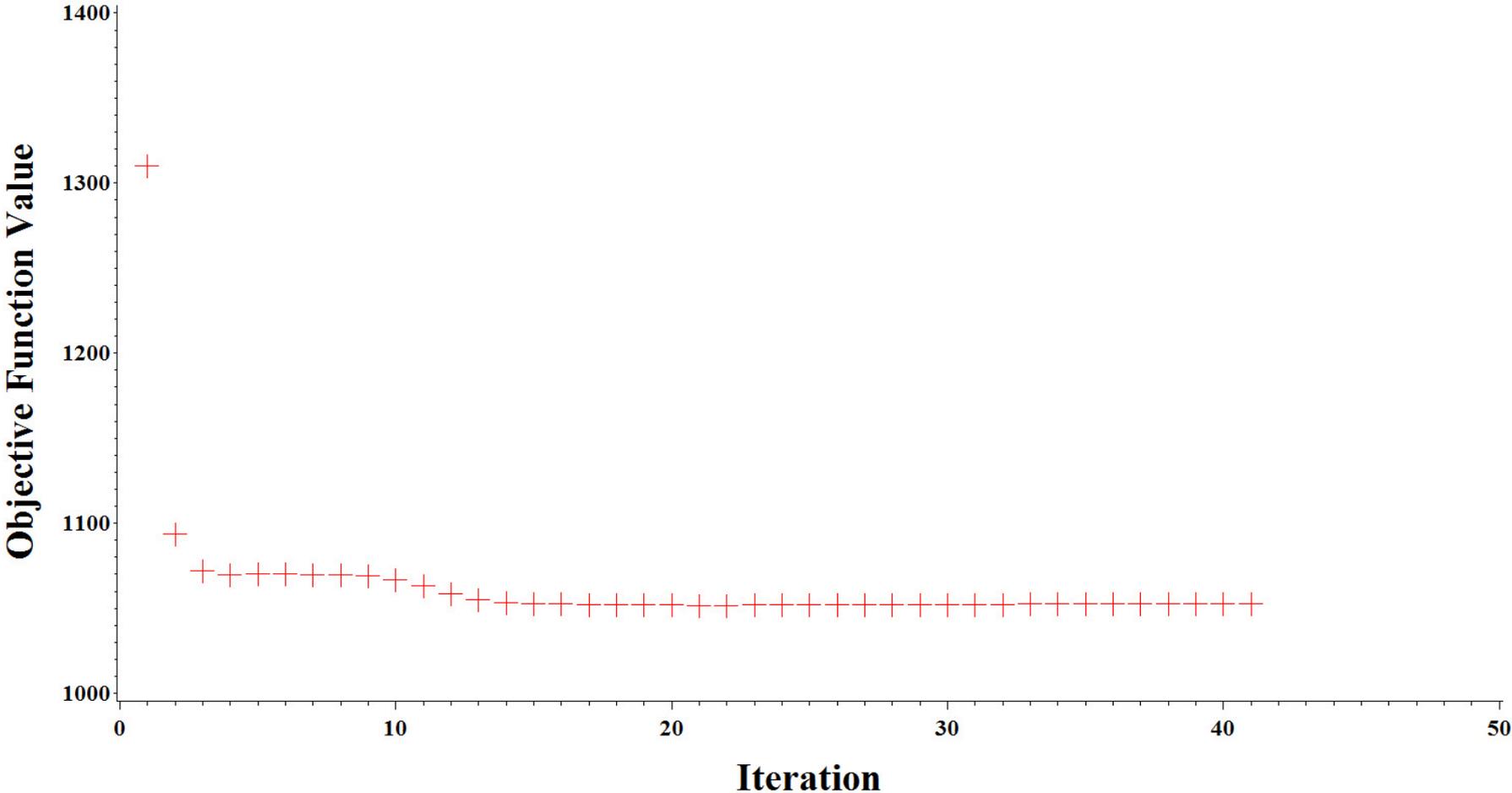
# Step 3 – Reoptimize

- New cluster centers are calculated by averaging the median household property value and median household income for each stratification cluster
- Re-optimize using the new cluster centers
- Iterate between Step 2 and Step 3 until the stopping criteria is met

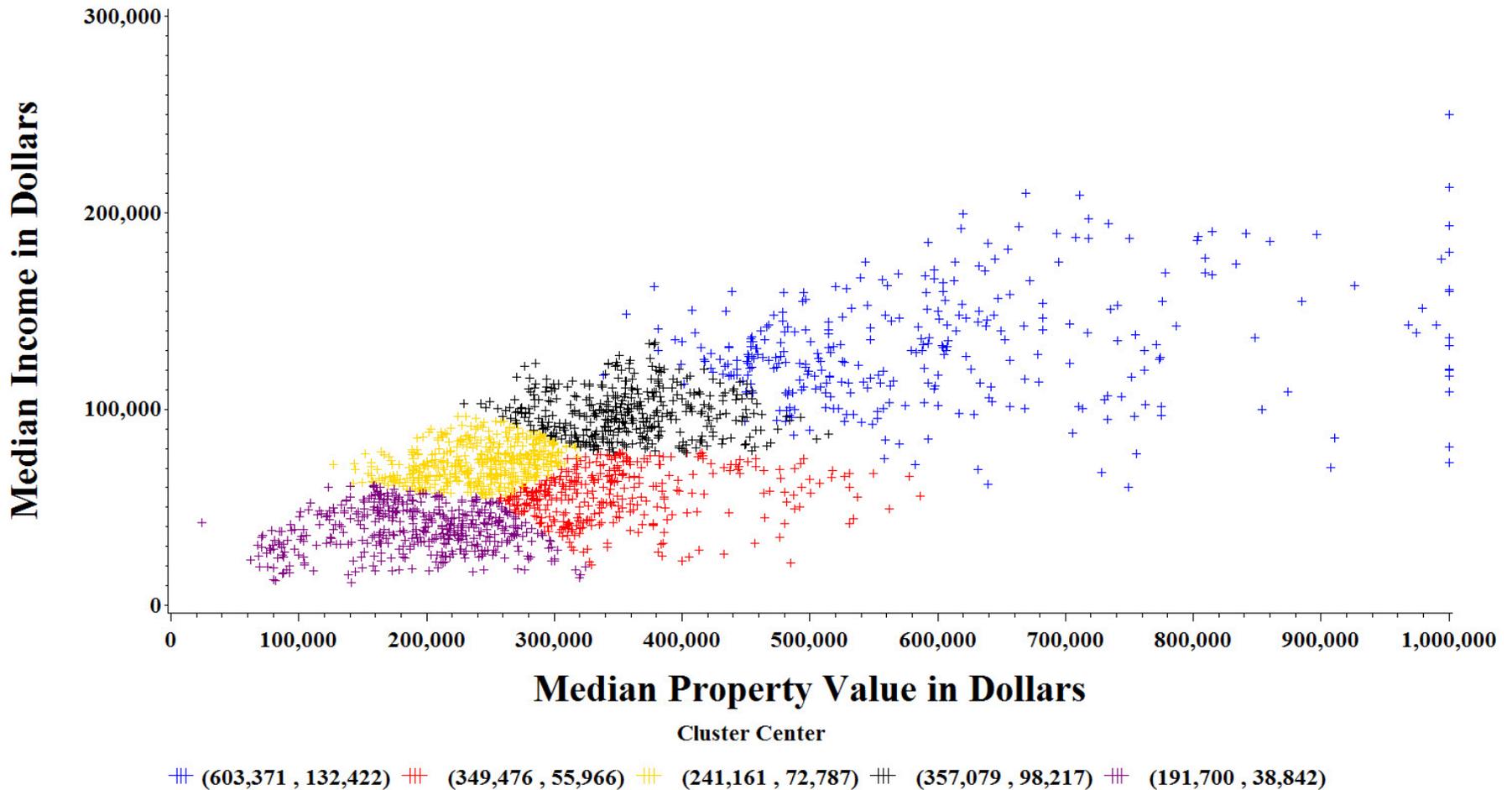
# Stopping Criteria

- Either stop when the Trace ( $W$ ) does not change for two consecutive iterations or the objective function value does not change for two consecutive iterations

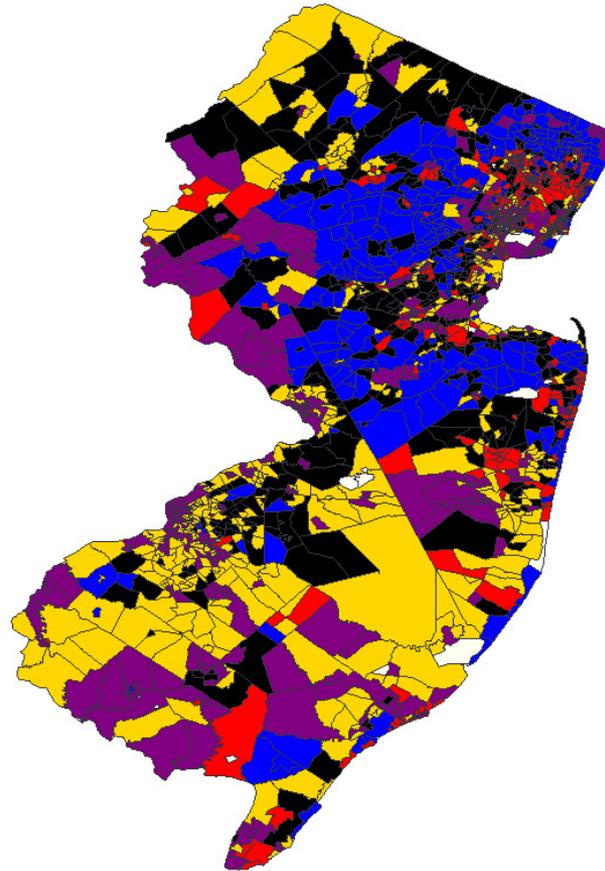
# Objective Function Value



# Stratification Clusters



# New Jersey Tracts



■ (603,371 , 132,422) ■ (349,476 , 55,966) ■ (241,161 , 72,787) ■ (357,079 , 98,217) ■ (191,700 , 38,842) □ No Information

# Census Tract Probability of Selection Results



# Interview Dropped County Distribution 2013

County	1	2	3	4	5	Total
Atlantic	0	10	5	0	21	36
Cape May	3	2	8	0	3	16
Cumberland	0	0	0	0	12	12
Mercer	5	0	21	6	2	34
Warren	0	0	1	1	4	6
Total	8	12	35	7	42	104

1 = (603,371 , 132,422)    2 = (349,476 , 55,966)    3 = (241,161 , 72,787)  
 4 = (357,079 , 98,217)    5 = (191,700 , 38,842)



# Interview Weight 2013

Cluster	Stratum Population	CE Population Design 2000	Weight 2000	CE Population Design 2010	Weight 2010
1	1,589,709	904,864	1.76	862,917	1.84
2	1,594,559	727,875	2.19	711,235	2.24
3	1,942,045	731,213	2.66	560,602	3.46
4	1,941,308	892,770	2.17	837,651	2.32
5	1,764,005	803,377	2.20	595,635	2.96

1 = (603,371 , 132,422)    2 = (349,476 , 55,966)    3 = (241,161 , 72,787)  
 4 = (357,079 , 98,217)    5 = (191,700 , 38,842)



# Diary Dropped County Distribution 2013

County	1	2	3	4	5	Total
Atlantic	0	0	3	0	8	11
Cape May	2	5	4	1	2	14
Cumberland	0	0	0	0	10	10
Mercer	1	0	0	4	6	11
Warren	0	2	0	2	4	8
Total	3	7	7	7	30	54

1 = (603,371 , 132,422)    2 = (349,476 , 55,966)    3 = (241,161 , 72,787)  
 4 = (357,079 , 98,217)    5 = (191,700 , 38,842)



# Diary Weight 2013

Cluster	Stratum Population	CE Population Design 2000	Weight 2000	CE Population Design 2010	Weight 2010
1	1,589,709	334,466	4.75	329,543	4.82
2	1,594,559	366,387	4.35	332,180	4.80
3	1,942,045	358,921	5.41	334,770	5.80
4	1,941,308	483,004	4.02	442,854	4.38
5	1,764,005	435,804	4.05	272,422	6.48

1 = (603,371 , 132,422)    2 = (349,476 , 55,966)    3 = (241,161 , 72,787)  
 4 = (357,079 , 98,217)    5 = (191,700 , 38,842)



# Calibration

$$\text{Minimize} \quad \sum_{i=1}^n (c_i - x_i)^2$$

$$\text{Subject to:} \quad \sum_{i=1}^n w_{ij} x_i = b_j \quad \text{for } j = 1, 2, \dots, 16$$

$$x_i > 0$$

$$0.5c_i \leq x_i \leq 4c_i$$

where:

$c_i$  = base weight (new first stage weight \* old within PSU sampling interval);

$w_{ij}$  = the number of members in  $CU_i$  with demographic characteristic  $j$ ;

$b_j$  = the U.S. population for demographic characteristic  $j$ ;

$n$  = the number of  $CU_i$ ;

$x_i$  = final calibration weight for  $CU_i$ , the variable to be calculated.

# Calibration Constraints

- Total New Jersey consumer units
- Total New Jersey homeowner consumer units
- Total New Jersey population by age
  - ▶ Age 14-24
  - ▶ Age 25-34
  - ▶ Age 45-54
  - ▶ Age 55-64
  - ▶ Age 65-74
  - ▶ Age 75 +

# Diary Non-Black Quarter 2, 2013

	Total							
	CUs	14-24	25-34	35-44	45-54	55-64	65-74	75 +
Respondent	92	24	28	32	34	33	30	2
New Jersey Population	3,281,493	1,117,446	879,869	985,401	1,143,750	968,234	607,199	542,355
Raw Weight	35,668	46,560	31,424	30,794	33,640	29,340	20,240	271,178



# Diary Black Quarter 2, 2013

	14-24	25-34	35-44	45-54	55-64	65-75	75+	Tenure
Respondent	1	6	6	1	12	0	2	56
New Jersey Population	257,920	166,169	139,739	197,040	140,715	65,489	45,032	2,146,012
Raw Weight	257,920	27,695	23,290	197,040	11,726	.	22,516	38,322



# Variance

## ■ Jackknife

- ▶ Compute the mean with all tracts,  $\hat{\theta}$
- ▶ Calculate the replicate means by dropping one tract at a time,  $\hat{\theta}_r$
- ▶ Calculate the variance,

$$\hat{V}(\hat{\theta}) = \sum_{r=1}^R \left( \frac{\text{number of tracks} - 1}{\text{number of tracks}} \right) (\hat{\theta}_r - \hat{\theta})^2$$

# Results



# Total Expenditures

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	62,709.25	65,165.19	67,621.13	2,455.94
Drop Warren	62,697.71	65,178.68	67,659.65	2,480.97
Drop South NJ	62,364.95	64,966.18	67,567.41	2,601.23
Design 2010	62,574.06	65,277.34	67,980.62	2,703.28



# Food

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	7,305.45	7,841.58	8,377.71	536.13
Drop Warren	7,252.07	7,790.83	8,329.59	538.76
Drop South NJ	7,159.70	7,700.97	8,242.25	541.27
Design 2010	7,166.61	7,734.60	8,302.59	567.99



# Alcoholic Beverages

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	493.71	587.57	681.42	93.85
Drop Warren	485.51	580.55	675.60	95.04
Drop South NJ	452.92	545.16	637.40	92.24
Design 2010	481.14	588.56	695.98	107.42



# Housing

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	24,182.27	25,063.75	25,945.24	881.49
Drop Warren	24,233.71	25,111.87	25,990.04	878.17
Drop South NJ	24,128.19	24,962.33	25,796.47	834.14
Design 2010	24,212.05	25,074.47	25,936.89	862.42



# Apparel and Services

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	2,227.76	2,600.12	2,972.47	372.35
Drop Warren	2,282.07	2,654.93	3,027.79	372.86
Drop South NJ	2,197.96	2,570.28	2,942.60	372.32
Design 2010	2,240.11	2,631.03	3,021.96	390.92



# Transportation

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	8,383.41	9,176.92	9,970.43	793.51
Drop Warren	8,410.67	9,242.94	10,075.21	832.27
Drop South NJ	8,334.92	9,205.52	10,076.13	870.61
Design 2010	8,210.63	9,117.73	10,024.83	907.10



# Health Expenses

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	3,626.70	3,856.83	4,086.97	230.13
Drop Warren	3,643.11	3,874.87	4,106.63	231.76
Drop South NJ	3,681.21	3,928.08	4,174.95	246.87
Design 2010	3,679.82	3,944.87	4,209.92	265.05



# Entertainment

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	2,386.01	2,583.90	2,781.80	197.89
Drop Warren	2,318.02	2,499.35	2,680.68	181.33
Drop South NJ	2,364.98	2,596.37	2,827.76	231.39
Design 2010	2,296.40	2,490.97	2,685.55	194.58



# Reading

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	86.83	109.99	133.14	23.15
Drop Warren	87.09	110.02	132.94	22.92
Drop South NJ	83.75	106.95	130.16	23.21
Design 2010	84.49	107.86	131.22	23.36



# Education

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	2,175.50	3,237.84	4,300.19	1,062.35
Drop Warren	2,165.89	3,253.35	4,340.81	1,087.46
Drop South NJ	2,046.52	3,375.65	4,704.79	1,329.14
Design 2010	2,138.44	3,502.35	4,866.27	1,363.92



# Tobacco Products and Smoking Supplies

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	243.36	288.71	334.07	45.35
Drop Warren	238.35	284.45	330.55	46.10
Drop South NJ	252.36	302.91	353.46	50.55
Design 2010	232.25	276.71	321.16	44.46

# Miscellaneous

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	463.09	546.90	630.70	83.80
Drop Warren	421.38	495.00	568.63	73.62
Drop South NJ	465.11	557.96	650.81	92.85
Design 2010	437.02	526.69	616.36	89.67



# Cash Contributions

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	1,338.78	1,543.40	1,748.01	204.61
Drop Warren	1,331.32	1,535.66	1,739.99	204.33
Drop South NJ	1,246.62	1,453.15	1,659.67	206.53
Design 2010	1,274.78	1,505.43	1,736.08	230.65



# Personal Insurance and Pensions

Design	Lower Bound (\$)	Mean (\$)	Upper Bound (\$)	Standard Error (\$)
Design 2000	6,471.60	6,949.70	7,427.80	478.10
Drop Warren	6,493.27	6,967.66	7,442.05	474.39
Drop South NJ	6,435.51	6,917.65	7,399.78	482.14
Design 2010	6,533.30	7,029.53	7,525.76	496.23



# Comparison to ACS

- ACS Public Use Microdata Sample (ACS-PUMS)
  - ▶ Sample in every state
  - ▶ Ask questions on housing, providing comparisons to CE
    - CE ask more detailed questions
  - ▶ CE calibrates to Current Population Survey (CPS)
  - ▶ ACS–PUMS calibrates to Census Population Estimates (PEP)

# Preliminary Expenditure Comparison between CE and ACS

Expenditure	National			New Jersey		
	CE (\$)	ACS (\$)	CE/ACS	CE (\$)	ACS (\$)	CE/ACS
Electricity and Natural Gas	1,814.16	2,117.00	0.86	2,308.24	2,667.57	0.87
Rented Dwellings	3,323.61	4,220.06	0.79	4,507.20	5,167.57	0.87



# Conclusions

- Means for Drop Warren, Drop South NJ, and Design 2010
  - ▶ are close to the mean for Design 2000 (truth)
  - ▶ are within the bounds for Design 2000 (truth)
- A good first step and a promising approach

# Future Work

- Application to other states
  - ▶ What other problems may arise
  - ▶ Threshold on interviews
- Calibration
  - ▶ Population and Tenure are well populated
  - ▶ Should age be replaced or further collapsed?
- Additional comparisons to other surveys

# Contact Information

**Susan L. King**

Mathematical Statistician

OPLC/SMD

202-691-6895

[king.susan@bls.gov](mailto:king.susan@bls.gov)

