A "First-Cut" at Forming Rural PSUs October2012

Susan L. King[1]

Abstract
After every decennial census, many surveys including the Consumer Expenditure Survey (CE) redefine their primary sampling units (PSUs), which are small sets of adjacent counties. CE conducts expenditure surveys in metropolitan, micropolitan, and rural areas in the United States. For metropolitan and micropolitan areas, the PSUs are the U.S. Office of Management and Budget's "core-based statistical areas" (CBSAs). Counties which are not in a metropolitan or micropolitan CBSA are rural and CE must group these counties into a PSU. For CE, rural PSUs are small clusters of adjacent counties that are required to have a minimum population of 7,500 people and a maximum area of 3,000 square miles. Unlike metropolitan and micropolitan CBSA's, rural PSU's are also required to be within a state boundary. Using an adjacency matrix and zero-one integer linear programming, a "first-cut" assignment of rural counties to a PSU is made. Since the algorithm does not account for geographical obstacles such as rivers and mountains, input from field representatives is used in the final assignment.

Key Words: PSU, zero-one integer linear programming, adjacency, Consumer Expenditure Survey

1. INTRODUCTION

Every ten years the Consumer Price Index (CPI) and the Consumer Expenditure Survey (CE) redefine their primary sampling units (PSUs) using the latest population estimates from the decennial census. Both surveys share the same metropolitan and micropolitan PSUs because CPI uses CE's expenditure estimates for its survey weights. CE and CPI's metropolitan and micropolitan PSUs are the U.S. Office of Management and Budget's (OMB) "core-based statistical areas" (CBSAs). A CBSA is a cluster of adjacent counties having an urban "core" of 10,000 or more people and the counties are socioeconomically tied to the core as measured by residents' commuting patterns. The counties in a CBSA may cross state boundaries. Counties which are not part of a metropolitan or micropolitan CBSA are rural. CE collects household expenditure data in all three geographic areas (metropolitan, micropolitan, and rural), whereas CPI collects data in only two of the three geographic regions (metropolitan and micropolitan). Since OMB does not group rural counties into small clusters of adjacent counties appropriate for PSUs, CE must define its own rural PSUs. CE requires a rural PSU to be within a state boundary, have a minimum population of 7,500 people, to be smaller than 3,000 square miles, and adjacent to other rural counties. Using an adjacency matrix and zero-one integer linear programming (Rardin), a "first-cut" assignment of rural counties to a PSU is made. Since the algorithm does not account for geographical obstacles such as rivers and mountains, input from field representatives is used in the final assignment.

Alabama and South Carolina are the states selected to illustrate the algorithm. Alabama has 24 rural counties to assign to PSUs and the average rural county land area is 792

[1] king.susan@bls.gov – U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. NE, Room 3650, Washington, DC 20212

square miles. Alabama is of manageable size to visualize the problem complexity, but its adjacency matrix is too large for a landscape page. Therefore, the example adjacency matrix is from South Carolina. All tables and figures are in the Appendix.

2. DATA

The data are a listing of rural counties by state, county FIPS code, county name, population, and land area in square miles. All of the data is publicly available. The rural counties are found using OMB's CBSAs for the 2000 Census. The county population and land area are from the U.S. Census Bureau's State & County QuickFacts for 2010 (http://quickfacts.census.gov/qfd/index.html). The algorithms were programmed in SAS®9.2 and two SAS® boundary files, CNTYNAME and USCOUNTY are used. Other software and boundary files could be used.

3. METHODS

After finding the set of rural counties in a state, the next step is to find rural county patterns that satisfy the state boundary, adjacency, population and land area constraints. The state boundary requirement is easily satisfied by treating each state as a separate problem and running the optimization by state.

3.1 Identifying Adjacent Counties

The set of feasible patterns is developed using an adjacency matrix. A county boundary file, such as the USCOUNTY file, is used to determine whether a particular county shares at least one common boundary point with another rural county. Each county in a state has multiple coordinates or boundary points, which are used to draw a map. The level of detail varies by boundary file. For each state, the adjacency algorithm progresses through the list of rural county boundary points using a nested do-loop that compares all possible pairs of rural counties. If two counties have identical coordinates, the two counties are adjacent.

The adjacency relationship between rural counties is marked in a matrix. Table 1 shows the adjacency matrix for South Carolina. The rural South Carolina counties are listed in the first column and top row of the matrix. Every element in the matrix is a "0" or a "1". A "1" indicates that two rural counties are adjacent and a "0" indicates that two rural counties are non-adjacent. For example, there is a "1" at the intersection of the Allendale row and the Bamberg column to indicate their adjacency, but there is a "0" at the intersection of the Allendale row and the Chesterfield column to indicate their non-adjacency. The matrix is symmetric and every county is adjacent to itself as indicated by the "1" along the matrix diagonal. Rural counties that are not adjacent to other rural counties are identified by having a row and column sum equal to one. For example, both the row and column sum for Lee and Chesterfield counties are one, indicating that these rural counties have no adjacent rural counties. The relationship between adjacent counties is not transitive. As an example, Marion is adjacent to Williamsburg, and Williamsburg is adjacent to Clarendon, but Marion is not adjacent to Clarendon.

3.2 Constructing Patterns or Potential PSUs

Using the adjacency matrix, a set of patterns or potential PSUs is formed. Patterns are based on a reference county and its adjacent rural counties. A reference county can be thought of as the PSU's "core" or "hub." Every pattern is required to have at least two counties. Patterns are found using combinations. If a reference county has n adjacent rural counties and k is a subset of n, then the total number of patterns for a single reference county is $\sum_{k=1}^{n}\binom{n}{k} = 2^n - 1$. Each rural county is treated as a reference county and the complete universe of patterns is the concatenation of each rural county's $2^n - 1$ patterns.

Figure 1 shows the rural counties in Alabama. Bullock County does not have any adjacent rural counties and is removed from analysis until the final map is created. Pattern formation is illustrated from the point of view of Cleburne and Clay counties. Cleburne County is in northeast Alabama and shares a border with Georgia. Cleburne is adjacent to Cherokee, Clay, and Randolph. The following seven patterns (2^3-1) are possible.

1. Cleburne, Clay
2. Cleburne, Cherokee
3. Cleburne, Randolph
4. Cleburne, Clay, Cherokee
5. Cleburne, Clay, Randolph
6. Cleburne, Cherokee, Randolph
7. Cleburne, Clay, Cherokee, Randolph

Clay County is adjacent to Randolph and Cleburne counties and generates an additional three patterns.

8. Clay, Cleburne
9. Clay, Randolph
10. Clay, Cleburne, Randolph

Patterns 8 and 10 are identical to patterns 1 and 5. Only one set is retained.

After finding the patterns for all counties in a state, the next step is to find the patterns that meet both the population and land area constraint. Patterns which fail to meet both of these constraints as well as duplicate patterns are deleted. The reference county method generated 261 rural county patterns for Alabama. All of the patterns meet the population constraint. After removing redundant patterns and patterns which failed to meet the land area constraint, there were 90 distinct patterns in Alabama.

3.3 Optimization Algorithm

Mathematical optimization, more specifically zero-one integer linear programming, is used to assign counties to PSUs such that each county is assigned to only one PSU. Each pattern is a candidate PSU. The set of patterns, a_{ij}, is a matrix where i is the county and j is a pattern. If a_{ij}=1 then county i is in pattern j. Otherwise, if county i is not in pattern j, then a_{ij}=0. The decision variable, x_j, equals 1 if pattern j is selected. Otherwise, x_j =

0. The constraint indicates that each county can be assigned to only one pattern. For the given set of patterns, the objective function minimizes the number of PSUs and thus reduces sample collection cost. Minimizing the number of PSUs increases the number of counties assigned to a PSU, and the total land area of the PSU will be closer to bound of 3,000 square miles. The patterns are not weighted so there is no cost coefficient in the objective function. Let n be the number of rural counties and m be the number of patterns. The zero-one linear integer programming model is:

$$\text{Minimize} \quad \sum_{j=1}^{m} x_j$$

$$\textit{Subject to}: \quad \sum_{j=1}^{m} a_{ij} x_j = 1, \quad \text{for every } i = 1, 2, ..., n$$

$$x_j = 0 \text{ or } 1$$

where:

$$x_j \quad \begin{cases} \\ \text{if pattern j is selected} \end{cases}$$

Alternative optimal solutions occur in zero-one integer programing and are abundant for this problem. An alternative optimal solution has the same number of PSUs but different county assignments to a PSU. Different solutions can be found visually, or with different software and solvers.

4. RESULTS

Two alternative optimal "first-cut " solutions for assigning the 23 rural counties to eight PSUs are shown in Figures 2 and 3 and their corresponding Tables 2 and 3 in the Appendix. These solutions are created using different SAS® solvers and other optimal solutions are possible. The 24th county, Bullock, has no adjacent rural counties and is a single county PSU. All of the rural counties have a population above 7,500 people. The size and spatial arrangement of counties can complicate assignments. For example, in the southwestern corner, Clarke is adjacent to both Monroe, and Washington counties and all three counties have a land area greater than 1,000 square miles. Consequently, all three counties cannot be in the same PSU. Perry and Marengo counties are also in the

southwest corner of the state. Perry County has only one adjacent rural county: Marengo County and these two counties must be in the same PSU.

The grouping of Randolph, Clay, Cleburne, and Cherokee counties in the Northeastern region of Alabama is the same on both maps and is possible because Cleburne is adjacent to Randolph, Clay, and Cherokee counties.

The algorithm builds PSUs based on a reference county and its adjacent counties. If the counties are chained, it may be possible to combine PSUs, reducing the total number of PSUs and the sample collection cost. To illustrate chaining, suppose DeKalb, Cherokee, Cleburne, Randolph, and Chambers were rural counties and met the population and land area constraints. These counties are chained. DeKalb and Chambers counties have only one adjacent rural county, whereas the other three counties have two adjacent rural counties. The algorithm would produce a two PSU solution. One PSU would have three counties: DeKalb, Cherokee, Cleburne or Cleburne, Randolph, and Chambers. The second PSU would have the remaining two counties. After reviewing the PSU assignments, the decision maker might want to combine these two PSUs into one PSU. The new pattern could be added to the list and the model re-optimized or the decision maker could make the adjustment manually. Not originally including this pattern is a limitation of the reference county method.

The decision maker may want to adjust the PSU assignments. For example to reduce transportation cost, it is desirable for the PSUs to be round. The decision maker might want to retain PSUs 2, 3, 4, and 8 from Figure 2 and PSUs 6 and 7 from Figure 3. Conecuh, Escambia, and Monroe counties could be placed into a PSU. Butler, Crenshaw and Covington would be in another PSU. Consequently, the assignments from the zero-one programming algorithm are only a "first cut" solution.

The number of rural counties in Alabama is an optimal size to visualize the problem complexity. The rural county assignments are neither obvious nor too cumbersome for the reader to attempt the assignments by hand and thereby, gain an appreciation of the algorithm. The assignments become more difficult as the number of counties increase and as the county populations decrease, triggering the population constraint. Additional constraints might be added such as an upper bound on the population, balancing the population or land area in a PSU, or travel distance restrictions.

5. EXTENSION TO OTHER STATES

Although a few states in the eastern United States do not have rural counties, most states must assign rural counties to a PSU. In the eastern United States, the counties are irregularly shaped and with the exception of Maine, the counties are less than 1,500 square miles and are sufficiently populated for automatic assignment. A few eastern states have counties with small land area. In the western United States it is more difficult to assign rural counties to PSUs due to counties with large land area and sparse populations. Many counties have a land area greater than 3,000 square miles. Even if the county land area is smaller than 3,000 square miles, the combined land area of two adjacent counties may be greater than the 3,000 square mile land area constraint. Also the spatial arrangement of counties with sparse populations and different land areas affects the assignment. In many cases, the "problem counties" can be removed for later assignment by a decision maker, and the assignment algorithm successfully run on the remaining counties. Sometimes the population and land area constraints can be relaxed

and an assignment made. CE decided to place counties with no adjacent rural counties, counties with a land area greater than 3,000 square miles, and counties which are difficult or impossible to group with another county into their own PSU.

6. CONCLUSIONS

Zero-one integer linear programming is used to make the assignment of rural counties to PSUs. The objective is to minimize the number of PSUs subject to the population constraints, land area constraints, and a constraint that requires each county to be assigned to only one PSU. Possible PSU patterns are determined *apriori* based on adjacency. Usually, there are alternative optimal solutions. With a colored map and a listing of the PSU assignments it is easier for the human eye to perceive assignment modifications. The map and county listing initiate the discussion on PSU adjustments due to the spatial arrangements of the counties, mountains, bridges, and other barriers. The algorithm is automatic in the eastern United States, whereas in the western states user intervention may be required. Therefore, the PSU groups from the PSU assignment algorithm are a "first cut" solution.

7. ACKNOWLEDGEMENTS

The views expressed in this paper are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

Rardin, Ronald L. 1998. Optimization in Operations Research. Prentice Hall, Inc., Upper Saddle River, NJ.

Appendix: Figures and Tables

Table 1: Adjacency Matrix for South Carolina

County Name	ABBEVILLE	ALLENDALE	BAMBERG	BARNWELL	CHESTERFIELD	CLARENDON	HAMPTON	LEE	MC CORMICK	MARION	WILLIAMSBURG
ABBEVILLE	1	0	0	0	0	0	0	0	1	0	0
ALLENDALE	0	1	1	1	0	0	1	0	0	0	0
BAMBERG	0	1	1	1	0	0	1	0	0	0	0
BARNWELL	0	1	1	1	0	0	0	0	0	0	0
CHESTERFIELD	0	0	0	0	1	0	0	0	0	0	0
CLARENDON	0	0	0	0	0	1	0	0	0	0	1
HAMPTON	0	1	1	0	0	0	1	0	0	0	0
LEE	0	0	0	0	0	0	0	1	0	0	0
MC CORMICK	1	0	0	0	0	0	0	0	1	0	0
MARION	0	0	0	0	0	0	0	0	0	1	1
WILLIAMSBURG	0	0	0	0	0	1	0	0	0	1	1

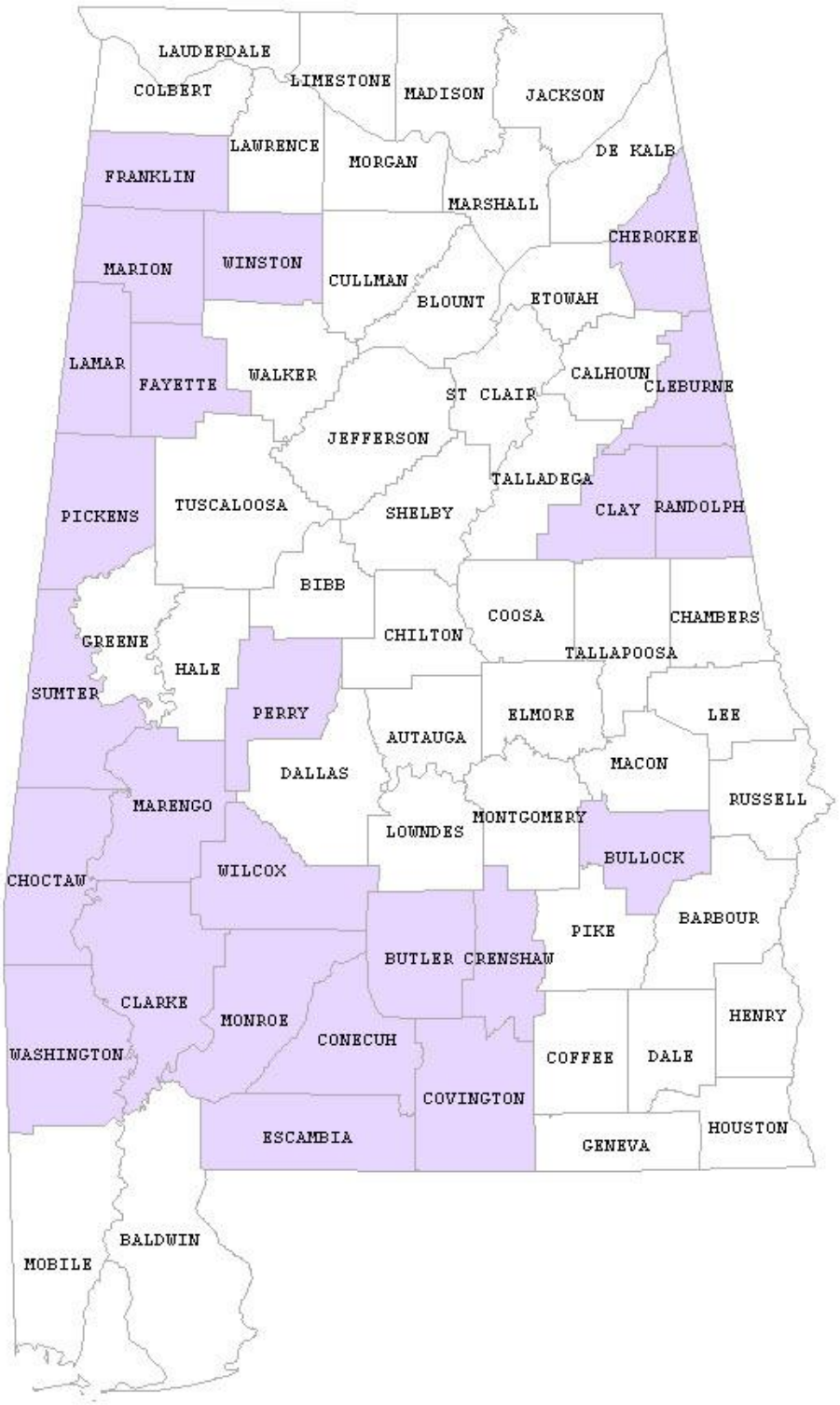

Figure 1: The rural counties in Alabama are shaded in lavender. The goal is to assign adjacent rural counties to a PSU such that the number of PSUs is minimized, the population is larger than 7,500 people and the land area is less than 3,000 square miles. Since Bullock County has no adjacent rural counties it will be placed in its own PSU.

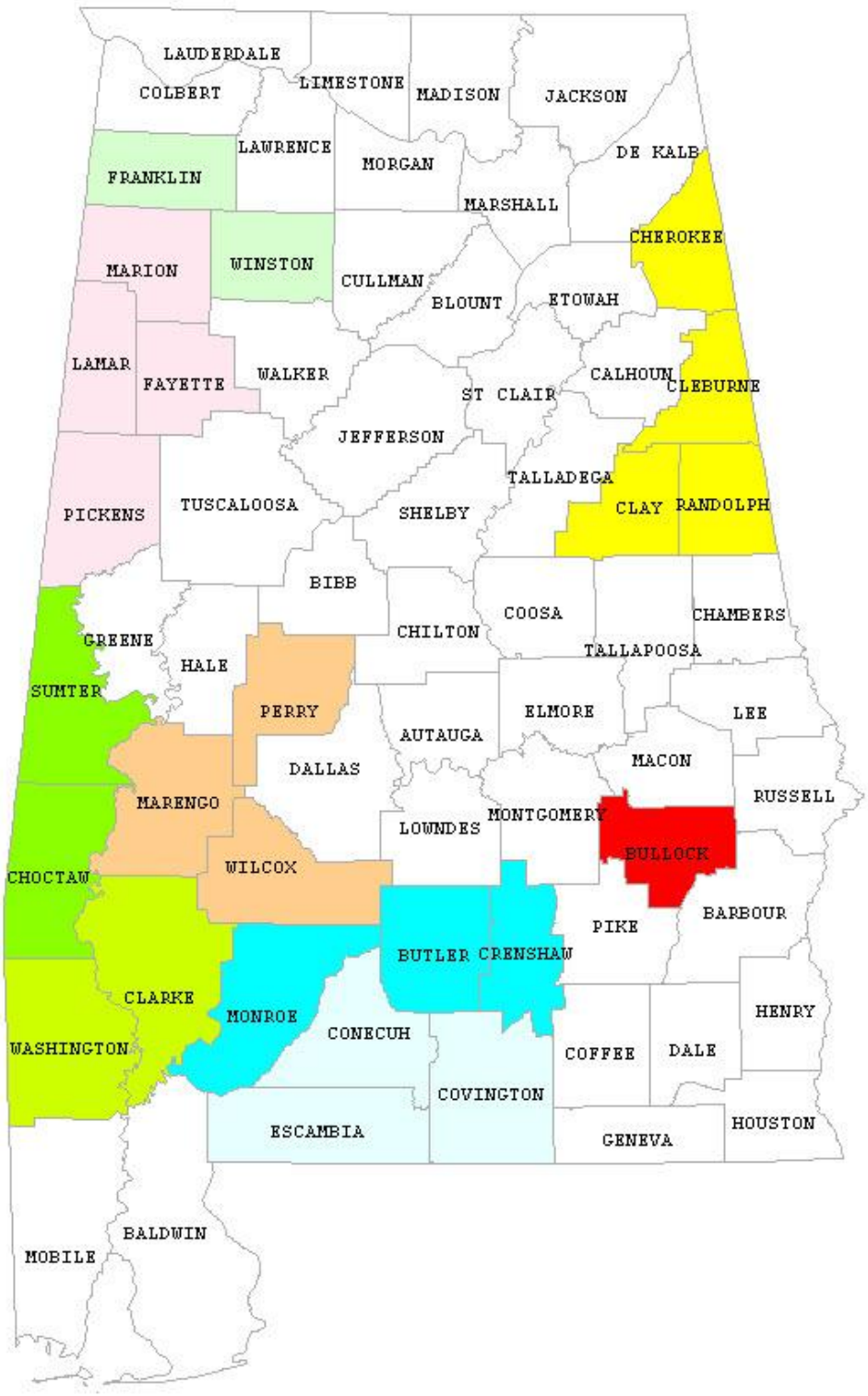

Figure 2: This map shows one combination of county assignments to nine PSUs. Bullock County is always in its own PSU. PROC OPTMILP was the optimization solver used to make the assignments.

Table 2: PSU Assignments using PROC OPTMILP

PSU	County Name	Population	Square Miles
1	BUTLER	20,090	776.87
	CRENSHAW	13,754	609.58
	MONROE	22,553	1,025.85
Total		56,397	2,412.30
2	CHOCTAW	14,055	913.51
	SUMTER	13,266	904.94
Total		27,321	1,818.45
3	CLARKE	26,304	1,238.38
	WASHINGTON	17,204	1,080.66
Total		43,508	2,319.04
4	CHEROKEE	24,545	553.12
	CLAY	13,809	605.07
	CLEBURNE	14,799	560.21
	RANDOLPH	22,620	581.05
Total		75,773	2,299.45
5	CONECUH	13,066	850.79
	COVINGTON	36,856	1,033.82
	ESCAMBIA	37,490	947.38
Total		87,412	2,831.99
6	FAYETTE	17,691	627.66
	LAMAR	14,295	604.85
	MARION	29,465	741.41
	PICKENS	19,524	881.42
Total		80,975	2,855.34
7	FRANKLIN	30,801	635.64
	WINSTON	23,974	614.44
Total		54,775	1,250.08
8	MARENGO	21,055	977.04
	PERRY	10,643	719.48
	WILCOX	12,803	888.68
Total		44,501	2,585.20
9	BULLOCK	10,796	625.01
Total		10,796	625.01

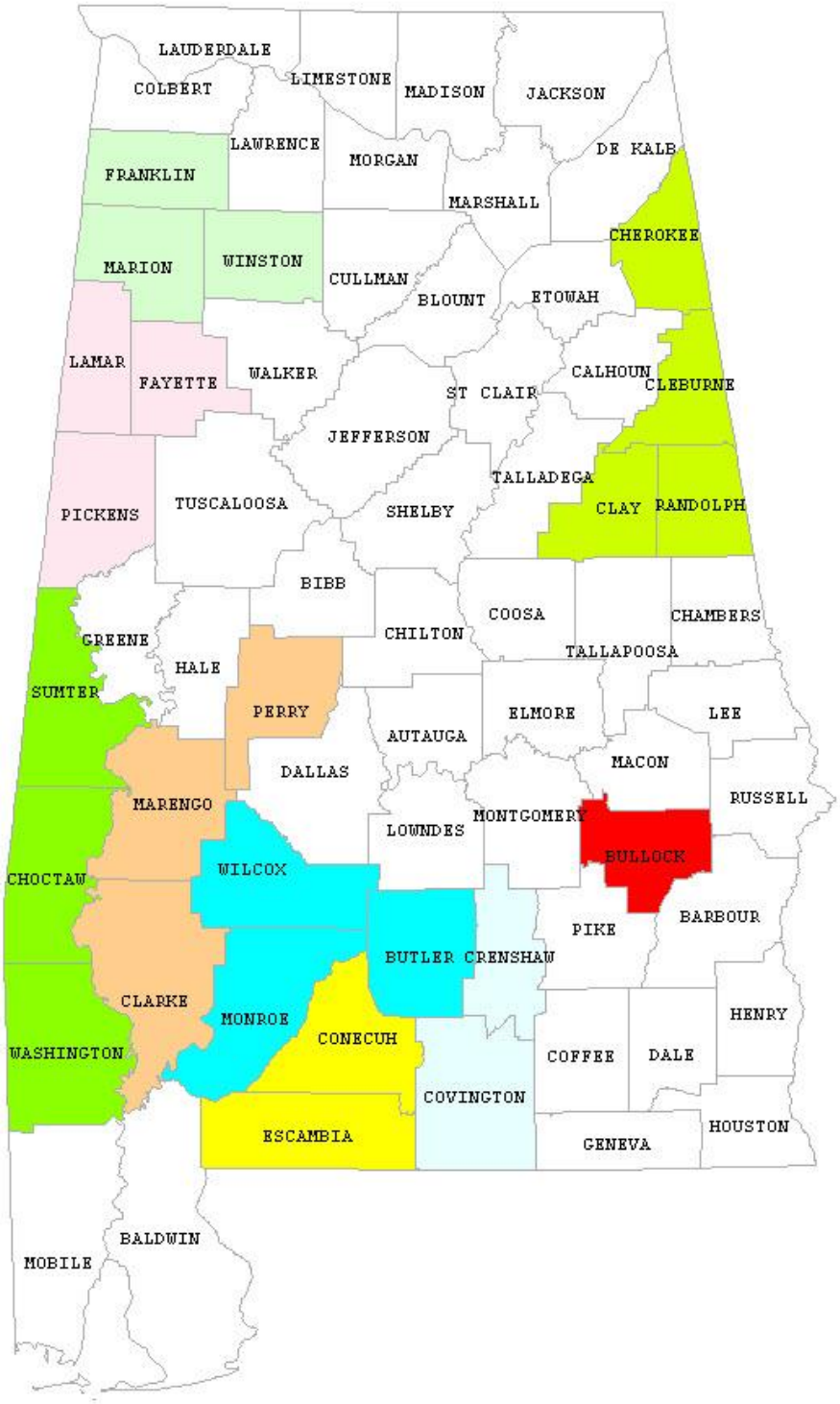

Figure 3: This map shows a second assignment of counties to nine PSUs. Bullock County is always in its own PSU. PROC LP was the optimization solver used to make the assignments.

Table 3: PSU Assignments using PROC LP

PSU	County Name	Population	Square Miles
1	BUTLER	20,090	776.87
	MONROE	22,553	1,025.85
	WILCOX	12,803	888.68
Total		55,446	2,691.40
2	CHOCTAW	14,055	913.51
	SUMTER	13,266	904.94
	WASHINGTON	17,204	1,080.66
Total		44,525	2,899.11
3	CHEROKEE	24,545	553.12
	CLAY	13,809	605.07
	CLEBURNE	14,799	560.21
	RANDOLPH	22,620	581.05
Total		75,773	2,299.45
4	CONECUH	13,066	850.79
	ESCAMBIA	37,490	947.38
Total		50,556	1,798.17
5	COVINGTON	36,856	1,033.82
	CRENSHAW	13,754	609.58
Total		50,610	1,643.40
6	FAYETTE	17,691	627.66
	LAMAR	14,295	604.85
	PICKENS	19,524	881.42
Total		51,510	2,113.93
7	FRANKLIN	30,801	635.64
	MARION	29,465	741.41
	WINSTON	23,974	614.44
Total		84,240	1,991.49
8	CLARKE	26,304	1,238.38
	MARENGO	21,055	977.04
	PERRY	10,643	719.48
Total		58,002	2,934.90
9	BULLOCK	10,796	625.01
Total		10,796	625.01