Using Census Data to Define Estimation Areas for the American Community Survey: A Case Study

Joseph C. Powers, US Census Bureau
Alfredo Navarro, US Census Bureau
4H054B DSSD, Washington DC 20233-7613,
joseph.c.powers@census.gov

Abstract

In January 2005, the American Community Survey (ACS) expanded to sample all 3,219 counties in the U.S. and Puerto Rico. The ACS weighting and estimation methodology requires estimation areas to meet a minimum population size so that the observed sample size is big enough to produce estimates with adequate reliability. Counties below the threshold size must be grouped or clustered prior to estimation. A simple method groups the counties based on adjacency and then assess all the clusters using a predefined criterion. A better, automated algorithm was also developed. The algorithm is an iterative method that uses a set of Census long form characteristics to define a similarity index based on the Euclidean distance metric. This paper describes the naïve method, the algorithm, and a statistical assessment. The results of the two systems are compared for Puerto Rico and Texas.

Keywords: Reliability, Naïve Clusters, Statistical Clustering, Compactness, Weighting, Estimation Areas

1 Introduction

The American Community Survey (ACS), as the replacement for the Census long form, is an essential part of the 2010 Census redesign. The ACS is a continuous survey with sample cases every month. While the ongoing nature of the ACS will provide more timely estimates of long form characteristics, this improvement comes at the cost of smaller samples. For Census 2000, about 17 million housing units were selected to receive the long form, but the ACS selects about 3 million addresses annually. To maintain a comparable level of data quality, areas with a household population of 65,000 or more will have estimates based on a single year of data, but smaller areas will require averaging over multiple years.

Estimation areas are the geographic level where the weights are computed; that is, a different set of weights will be derived to use within each estimation area. For the ACS, the fundamental building block of the estimation area is the county. All estimation areas contain one or more counties, and every county belongs to exactly one estimation area. Because the stability of the weights depends on having sufficient sample, estimation areas must meet a population size criterion.[1] The population size criterion, or threshold size, depends on the state's sampling rate, the state's unweighted response rate, and a minimum of 400 interviews, a standard set for the Census 2000 long-form. That is, for state j, unweighted response rate R_j, and sampling rate S_j: $\text{ThresholdPop}_j = \lceil \frac{400}{R_j S_j} \rceil$. The thresholds range from 14,673 for North Dakota to 31,320 for Florida.

Because the ACS must produce annual estimates for areas with a population of at least 65,000, these counties must form their own estimation areas.[2] Counties with population between the threshold and 65,000 may be their own estimation area, or they may be clustered with counties below the threshold. However, no estimation area may contain more than one county with more than the threshold population. Additionally, because the ACS must produce annual state-level estimates, estimation areas may not cross state lines.

In addition to meeting the threshold size, estima-

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

[1] The two exceptions to this rule occur in Massachusetts and Puerto Rico. In Massachusetts, the only two counties below 65,000 are Dukes and Nantucket. They form an estimation area even though their combined population is 23,554, about 4,800 people fewer than the state threshold. The exception in Puerto Rico is 124 people shy of the threshold and is discussed in more detail below.

[2] The sole exception to this rule occurs in Hawai'i. Since Kalawao County, a former leper colony with only 147 people, does not have a county seat, and is administered by the Hawai'i Department of Health, it was combined with Maui County instead of Kauai County beacuse Kalawao is physically connected to Moloka'i, one of the islands of Maui County. However, Maui County is larger than 65,000.

tion areas should be consistent with another principle: The best estimation area for a county is the county itself. In practice, this principle often conflicts with the size criterion. When a county cannot be its own estimation area, it should be combined with counties as similar to it as possible. In other words, each estimation area should be as homogenous as possible. As the dissimilarity in the counties of an estimation area increases, the less well the weights represent the individual counties.

In many surveys, estimation areas are formed from contiguous groups of counties. While the adage that birds of a feather flock together is apt, this practice is also often justified by cost constraints. For instance, many surveys sample a few metropolitan statistical areas (MSA) to represent the nation. In these surveys, it makes sense to force the estimation areas to be contiguous, since this also allows for estimates at the MSA level. Since the ACS includes every county, and must produce estimates at the state level, one of the aims of this study was to determine if requiring estimation areas to be physically connected is worth the sacrifice in homogeneity.

2 Measuring Homogeneity

Creating homogenous clusters of counties requires some means of comparing counties, and a basis of comparison.

2.1 Characteristics to be Compared

The characteristics used for comparison are:

1. An estimated poverty rate of the housing unit population.
2. The percent of the housing unit population living in rural areas.
3. The percentage of the housing unit population renting.
4. The demographics by sex, age, race and ethnicity of the housing unit population.
5. The latitude and longitude of the county's physical centroid.

With the exceptions of items (1) and (5), all of the population characteristics are short-form Census items and were computed directly from the housing unit returns. Because the weights applied to the long-form estimates are controlled to population totals that include residents of group quarters, counties with nonzero group quarters populations may have different estimated poverty rates, since this estimate is used for the housing unit population.

The age dimension of the demographic information is broken down into six categories: 0–5, 6–17, 18–25, 26–54, 55–64, and 65 or older. The race and ethnicity dimension differs between Puerto Rico and the States. Stateside, the race and ethnicity categories are: non–Hispanic Whites or Other[3], non–Hispanic Black or African American, non–Hispanic American Indian or Alaskan Native (AIAN), non–Hispanic Asian, non–Hispanic Native Hawaiian or Other Pacific Islander (NHOPI), and Hispanic of any race. Because over 98% of the population of Puerto Rico is Hispanic, Hispanics are grouped with the non-Hispanics of their race. Also, the race dimension in Puerto Rico had the three categories White, Black, and Other.

The Stateside demographic array has twice the dimensions of the Puerto Rico array, and in many states, numerous parts of it will be sparse. That is, many states have very few people of some race groups. To increase numerical stability, the state level demographics are run through a collapsing rule. Any race group with less than 10% of the population was collapsed with the race group with the closest estimated national undercount rate. For example, a state with a small Hispanic population would form a combined group of Blacks and Hispanics, and then if all groups were more than ten percent, the collapsing would stop.

Latitude and longitude of the county's centroid were included to encourage the formation of physically close clusters, without introducing the complexity of treating adjacency in terms of graph theory. Again to increase numerical stability of the calculations, the centroid latitudes and longitudes for each county in the state were scaled to lie in the unit interval [0, 1]. Separately for each dimension, the minimum value for the state was subtracted from the coordinate of each county, and then the difference was divided by the range of values in the state. $\text{TransformedLat}_j = \frac{\text{latitude}_j - \min(\text{latitudes})}{\max(\text{latitudes}) - \min(\text{latitudes})}$ That is, the state's southernmost county's (northernmost) centroid latitude has a transformed value of zero (one), and likewise for the westernmost and easternmost counties' centroid longitudes respectively.

2.2 Method of Comparison

After grouping the counties into clusters which attain the population threshold, the natural questions

[3]That is, people who indicated they were not Hispanic and selected only "Some Other Race" as their race.

to consider are how good are they, and how do they compare to other proposed groupings. The answer to the second question is built from the answer to first: For any cluster, compare the cluster's characteristics of interest to each county's characteristics, and then summarize the comparison over the counties in the cluster. This is done in two steps:

1. For each characteristic calculate the absolute relative difference between the county's characteristic and the cluster's.

2. Calculate a weighted average of the relative differences, weighting the differences on the basis of importance of each characteristic.

The weighted average measures how similar the county is to its cluster. To measure the overall quality of the cluster, take the simple average of the weighted averages for each county in the cluster.

The first calculation is simple. For characteristic j, the corresponding relative difference (RD) is:

$$\mathrm{RD_j} = \begin{cases} \frac{|\mathrm{County_j} - \mathrm{Cluster_j}|}{\mathrm{Cluster_j}} & : \quad \mathrm{Cluster_j} \neq 0 \\ 0 & : \quad \mathrm{Cluster_j} = 0 \end{cases}$$

The second calculation is also simple, but depends on some judgment in setting the weights. After considering several possible sets, the weights were chosen so that the relative difference for poverty receives heavy weight, and so that the other variables receive a fair share of the remainder. The weights are one-half for poverty, one sixteenth for each of latitude and longitude, one eighth for each of percent renting and percent rural, and the remaining eighth divided equally among those demographic categories that are not collapsed.

3 Naïve Clusters

Many of the decisions made in how to go about creating the clustering algorithm originate in a simpler, more naïve procedure. The simplest method to form clusters is to manually designate several cluster systems based only on geography. That is, take a state map, label the counties with their populations, and have someone group the counties into clusters above the state's threshold size. This method suffers from being labor intensive and using little of the available data. Perhaps its greatest shortcoming is the unknowable subjective aesthetics of the person doing the grouping. To its credit, this method is an easy way to produce contiguous clusters, and is useful for exploratory analysis. For instance, the final choice for the weights applied to the relative differences was made after comparing several choices using cluster systems generated in this way.

4 Automated Clustering

Having decided how to measure the quality of a set of clusters, the next challenge is to generate collections of them and choose the best, preferably without human intervention. In the ideal case, it would be possible to construct all possible sets of clusters, eliminate those sets that contain clusters with populations below the threshold size, rank the survivors based on the mean weighted average relative differences, and choose the one with the smallest rank. As a practical matter, this is impossible–the number of clustering systems very quickly grows very large as the number of counties to cluster increases. For instance, Puerto Rico has 28 municipios (the county-equivalent governmental unit) that must be combined to form estimation areas, and another 40 municipios have between the threshold and 65,000 population. Some states have even more counties that must be clustered. Unfortunately there are 6,160,539,404,599,934,652,455 ($6.1605\ldots \times 10^{21} \approx 2^{72.38}$) distinct partitions of a set of just 28 elements. This number is larger than the largest integer most programs can store, so even to list the possible clustering systems would require considerable special processing. (For more details on Bell numbers, see Weisstein.)

Even if it were possible to enumerate every possible clustering system, computational difficulties would remain. Not every system is valid. That is, some will contain clusters that do not meet the population threshold. Other systems will contain a cluster with more than one county large enough to be its own cluster, and thus would also have to be removed. These would need to be screened out. Calculating the relative differences takes even more resources, and searching such a large dataset for the minimum is not trivial.

Instead of trying to optimize over many clustering systems at once, an alternative is to try and build a good clustering system heuristically. Our automatic method is an iterative method built on the idea that the best way to form a cluster is to start with a county under the threshold population, and either collapse it with a previous cluster, or to add similar counties until the cluster is above the threshold size. Then, the algorithm moves to the next undersize county not in a cluster until counties under the threshold are all clustered.

More formally, each county is a vector with components the characteristics of interest (poverty, percent rural, etc.). The vectors are then normalized so each has unit length. Counties are sorted in ascending order of population size, so that the counties

farthest from the threshold are clustered first. Then the algorithm passes through the list, oscillating between two states, the initialization state and the collection state, until all counties under the threshold are assigned to a cluster.

In the initialization state, the algorithm enters the following loop:

0. Delete from the list of the state's counties those counties with housing unit population of at least 65,000. Since these counties must be their own estimation areas, they are not be considered by the algorithm.

1. Pick out the smallest unclustered county and collapse it into a cluster of its own. If this county by itself is larger than the threshold size, go to 5. If there are no unclustered counties of any size, all counties must be assigned to clusters, so stop.

2. Calculate the distances between the singleton cluster and all clusters previously formed.

3. Calculate the distances between the singleton cluster and all counties under 65,000 that are not assigned to any cluster.

4. Identify the minimum distance.

   (a) If the minimum distance is from step 2, collapse the singleton into the cluster with this distance, and recompute and normalize the cluster's characteristics with the new county. What was the smallest unclustered county is now in a cluster above the threshold size, so go to step 1.

   (b) If the minimum distance is from step 3, combine the county with this distance with the singleton, compute and normalize the new cluster's characteristics, and enter the collection state. In the collection state, the algorithm enters the following loop:

      i. If the size of the current cluster under consideration is at least the threshold population, stop collecting counties for this cluster and go to step 1.

      ii. Calculate the distance between the cluster and all unclustered counties.

      iii. Collapse the nearest county into the cluster, recompute and normalize the cluster's characteristics.

      iv. Go to step 4.b.i.

5. If the algorithm has formed a cluster with one member, and that member is above the threshold size, all counties under the threshold size must have been combined into clusters, because the list of counties is sorted in ascending order by size. Stop the algorithm and assign any remaining counties to singleton clusters. That is, if any counties above the threshold and below 65,000 population remain, each of these counties forms its own cluster.

The distance calculated in the algorithm is a slight modification to simple Euclidean distance. First the simple Euclidean distance is calculated, and then it is adjusted using two independent adjustment factors. By the way the data have been transformed, Euclidean distance is equivalent to a similarity measure that treats all variables equally. The two adjustment factors encourage clusters to have properties that are not directly measured by the variables used.

The first adjustment factor is named the undersize discount factor because it makes the most sense to set it less than or equal to one. Distances to counties under the threshold size (i. e. those that will need to be clustered) are reduced (discounted). Distances between a singleton cluster and counties under the threshold size, and between a singleton cluster and clusters with all members under the threshold, are multiplied by the discount factor. Thus, when the undersize discount factor is smaller than one, the clustering algorithm is driven to prefer counties that will need to be clustered anyway. The result is to use fewer counties that are over the threshold size, in the hope that more counties over the threshold size end up being their own clusters.

The second adjustment factor is called the CBSA status penalty factor. In 2003, the Federal Office of Management and Budget revised the definitions of MSAs and created a new entity, called the Core Based Statistical Area (CBSA). Because there are many more CBSAs than MSAs, applying a discount when two proposed components of a cluster have the same CBSA is impractical. Instead, the CBSA penalty is applied to the distances between two objects with differing CBSA status. For example, when county Ψ is in some CBSA, the distances from Ψ to counties not in any CBSA are multiplied by the penalty factor, as are distances from Ψ to clusters that have both non-CBSA and CBSA members. Similarly, when a county is not in a CBSA, the penalty is applied to distances that would create a cluster of mixed CBSA status (to counties in some CBSA, and to clusters with all members in some CBSA).

A priori, it is impossible to know what values of the undersize discount and CBSA penalty factors will produce the best clustering system. Because the threshold population varies by state, and because the population by county differs from state to state, it is unreasonable to expect one undersize discount factor to be optimal for all states. Likewise, states differ in the number of CBSAs they contain so the CBSA status penalty factors will also differ. Nevertheless, by the nature of their construction, certain bounds on the possible values of the factors exist. Obviously, neither can be negative, since negative distances would not make sense. The CBSA status penalty factor could take on any value in the interval $(0, \infty)$; however values less than one act to encourage the formation of clusters with different CBSA status. The undersize discount is constrained to be in $(0, \infty)$, but values larger than one encourage the formation of clusters with counties over the threshold, when the aim is to discourage such clusters. In order to optimize these parameters, a discrete search of the "interesting" region of the parameter space was performed. The CBSA status penalty was allowed to vary from 1 to 1.5, inclusive, in increments of 0.02. At the same time, the undersize discount varied from 0.7 to 1.0, inclusive, also in increments of 0.02. The optimal pair of parameters is the smallest pair that produces the clustering system with the minimum mean weighted average relative difference, among the 416 possible systems.

5 Results

Initially, three sets of naïve clustering systems for each of Texas and Puerto Rico were created. Because they consisted of mostly contiguous parts, they do not differ by very much within a state. In the interest of brevity, only one naïve clustering system for each state level entity is presented. Similarly, only the automatic systems associated with the optimal combination of the adjustment factors are included. A summary of the automatic method for the nation is also included for comparing Puerto Rico and Texas to the nation as a whole.

Table 1 describes the counties to be clustered in Puerto Rico, Texas, and the nation as a whole, as well as describing the clusters with two or more members. The columns are:

0. The clustering system.
1. The housing unit population threshold.
2. The total housing unit population of the counties under the threshold.
3. A triage of the counties:
   (a.) The number of counties below the minimum cluster size.
   (b.) The number of counties above the threshold and below 65,000 housing unit population, i. e. potential donor counties.[4]
   (c.) The number of counties above 65,000 housing unit population.[5]
4. The number of clusters with two or more counties.
5. A breakdown of the clusters with two or more members:
   (a.) The number of clusters which have neither a county above the threshold nor mixed CBSA status.
   (b.) The number of clusters which have a county above the threshold as a member.[6]
   (c.) The number of clusters which have mixed CBSA status.
   (d.) The number of clusters which have both mixed CBSA status and a donor county as a member.

Several quantities not listed can be inferred from this table. For instance, the 3,219 counties in the United States and Puerto Rico form 2,006 estimation areas. The automatic method assigns over 72% of potential donors to their own estimation areas. In the case of Texas, the automatic method was much more likely to form clusters with mixed CBSA status. The naïve methods used ten fewer counties to form estimation areas in Puerto Rico, and fourteen fewer in Texas at the same time time that it made fewer clusters. This last point is of particular importance—the naïve method does a better job of keeping counties above the minimum threshold from joining with counties that must be clustered.

Table 2 describes the size of the clusters with more than one member. Because of the variability in the size of the estimation areas, in Texas the two–sided t–test p–value[7] for differences in average cluster population between the automatic and naïve methods is

[4] Nationally, there were 34,444,736 people in housing units living in these counties. Texas had 1,682,617 and Puerto Rico had 1,622,978.

[5] In total, 224,368,429 people living in housing units lived in these counties. Texas had 16,909,960 and Puerto Rico 1,609,012.

[6] The breakdown of housing unit population for these counties is: US 9,095,935, Texas 745,519, Puerto Rico 636,255.

[7] Assumes equal underlying variances, pools sample variances and uses Satterthwaite's approximation for degrees of freedom. F–tests of this assumption have p–values above 0.50.

Table 1: Basic State and Cluster System Characteristics

Level	MinCSize	PopLTT	Counties	N2+	Cluster Char.
Automatic Nationwide	Varies	18,591,944	1,595:886:738	627	210:245:275:103
Automatic Puerto Rico	28,185	529,846	28: 40: 10	20	4: 15: 3: 2
Naïve Puerto Rico				15	8: 5: 2: 0
Automatic Texas	28,596	1,698,134	165: 42: 47	50	17: 19: 22: 8
Naïve Texas				43	33: 5: 5: 0

Table 2: Multi–member Cluster Population Statistics

Level	N2+	Mean	Std. Dev.	Min.	Median	Max.
Automatic Nationwide	627	42,915.8984	16,680.7177	15,711	39,980	126,840
Automatic Puerto Rico	20	58,473.3023	14,198.8117	28,061	57,710	83,983
Naïve Puerto Rico	15	49,132.5333	12,504.0088	34,282	49,041	73,865
Automatic Texas	50	46,370.5815	15,330.3455	29,734	43,499	94,027
Naïve Texas	43	43,997.5349	14,828.5282	28,603	38,359	85,910

Table 3: Clustering System Scores

Level	N2+	Mean WRD	Std. Dev.
Auto. US	627	0.1293256354	0.0864796159
Auto. PR	20	0.1322528767	0.1094235704
Naïve PR	15	0.1398125127	0.0934742150
Auto. TX	50	0.1205622774	0.0711464338
Naïve TX	43	0.1806154261	0.0893082663

above 0.45. For Puerto Rico, the corresponding p–value is 0.051. On the average, the naïve method produced smaller clusters in Puerto Rico.[8]

By comparing Table 2 to Table 1, we can see that the smallest cluster from the automatic method in Puerto Rico is below the minimum cluster size. Initially, this cluster consisted of Culebra, Las Marías, and Jayuya, and was above the threshold size, but Culebra was manually removed and joined to the cluster containing Vieques, because Culebra and Vieques are both municipios of islands not physically connected to the rest of Puerto Rico. Las Marías and Jayuya together are just below the threshold size.

Table 3 summarizes the performance of the clustering methods. The Mean WRD Score column is the simple average of the weighted average relative differences for the clustering system. Here the question of interest is whether or not the automatic method produces clusters that are better than the naïve method. While the one–sided p–value for Puerto Rico is over 0.40, for Texas the p–value is about 0.00025.[9] In Texas, the typical cluster of counties from the automatic method is much more homogenous than for a cluster produced by hand.

The maps on the last page of this paper describe the estimation areas analyzed above. Desecheo and Mona, outlying islands to the west of Puerto Rico, are not shown, as they part of Mayagüez Municipio. Counties are numbered with their estimation area, except for counties that are their own estimation areas, which are labelled with zeros. The different estimation areas are also shown with different colors. Counties with red hatching were below their states' minimum size thresholds. Counties shaded grey had between the state threshold and 65,000 and were their own estimation areas, while those that are combined with other counties have the color of the estimation area. Counties above 65,000 housing unit population have no shading, hatching, or coloring because they were required to be their own estimation areas.

6 Conclusion

In general, the automatic method performs as well or better than a manual approach that strives for contiguity above all else. The automatic method is faster, i. e. it is possible to produce many more proposed systems algorithmically than with the manual approach in a given amount of time. Additionally, the manual method is more prone to subjective influences and is designed to produce homogenous-looking maps, not necessarily homogenous clusters.

The estimation areas produced from the automatic method are expected to remain stable for some time; the earliest possible date for a new set would be

[8] By long–standing convention, the Census Bureau uses an α of 0.10.

[9] Assumes equal underlying variances, pools sample variances, and uses Satterthwaite's approximation for degrees of freedom. F–tests of this assumption have p–values above 0.15.

sometime after long-form equivalent data are available for every county, in 2010, for use beginning with the 2011 sample year. Nevertheless, research on improvements to the automatic method continues. Preliminary investigations indicate that the algorithm is fairly robust to changing the order of the counties to be clustered, and to changing the levels of the distance adjustment factors. Possible tweaks to the algorithm that have not been investigated include changing the population percentages used in collapsing the race groups, the choice of transformation of the centroid latitude and longitude, and further modifying the distance metric. Investigating these possibilities would help determine that the algorithm is a robust procedure, and that the current choices are reasonable.

Resources permitting, it may be possible to move away from a custom algorithm and towards off the shelf clustering procedures combined with a jackknife–like system that will keep the constraints on cluster membership in place. Alternatively, since the choice of clustering system corresponds to a minimization over a space of set partitions, simulated annealing or other Monte Carlo methods may be worthwhile, particularly if the set of all partitions can be restricted to the partitions that correspond to valid clustering systems.

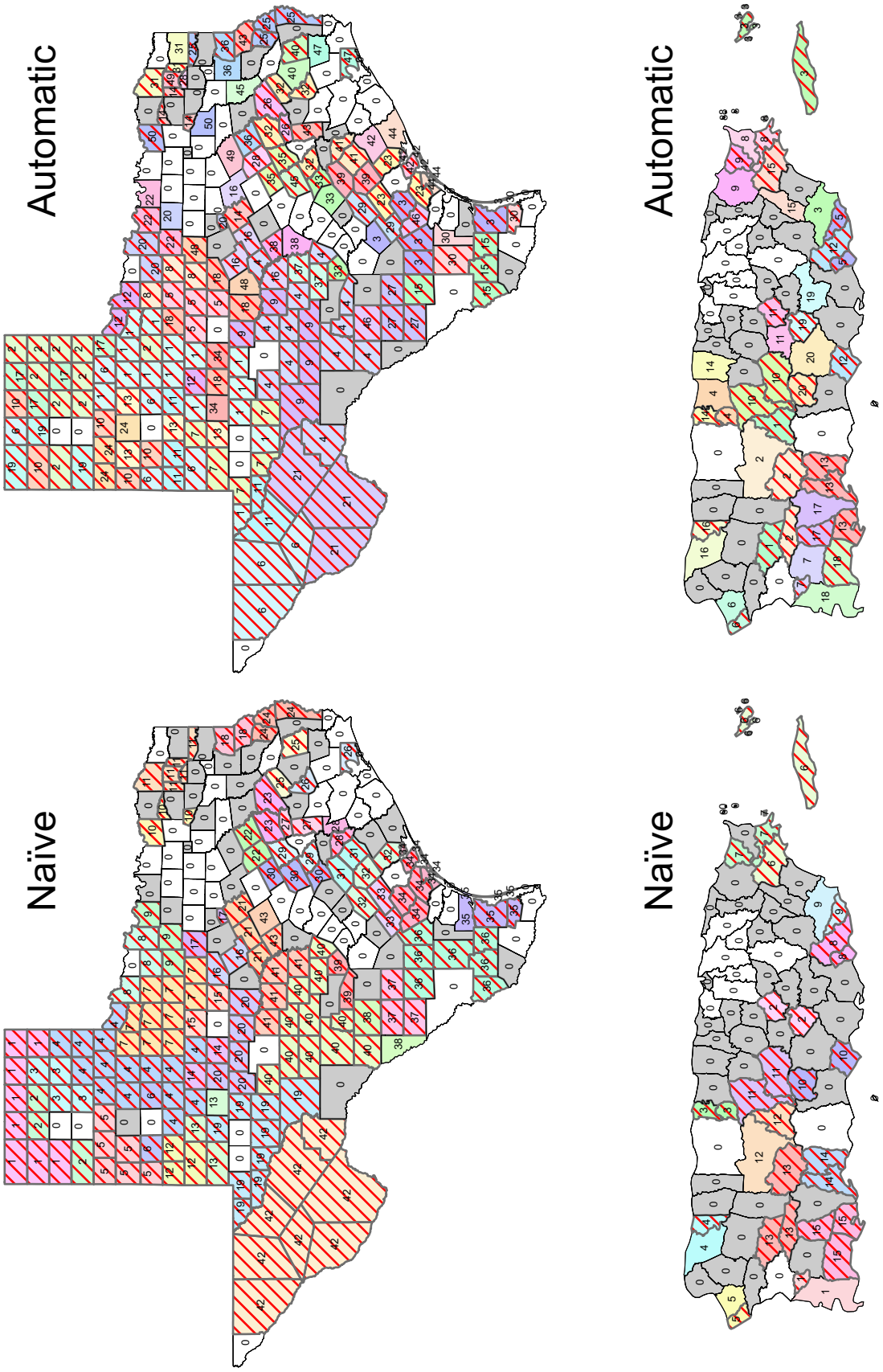
Acknowledgements

The authors wish to thank David Hubble and Rajendra Singh; their questions helped shape much of the research that went into this paper. Mark Asiala provided invaluable SAS® and LaTeX 2_ε advice at a number of places, and without his help things would not have turned out half as well as they did. Thanks also to Dale Garrett and Eric Schindler for their helpful comments on this paper.

References

Weisstein, Eric W. "Bell Number." From MathWorld–A Wolfram Web Resource. http://mathworld.wolfram.com/BellNumber.html Accessed 2006-06-23.

Memorandum for documentation from H. Shoemaker, "Documentation of PSU Definitions, Stratification, and Selection for the ACS National Sample 2000–2002 (ACS–S–16)."


Automatic
Automatic
Naïve
Naïve