Assigning PSUs to a Stratification PSU

Susan L. King[1], John Schilp, Erik Bergmann

Abstract

After every decennial census, many surveys including the Consumer Price Index (CPI) and the Consumer Expenditure Survey (CE) redefine their primary sampling units (PSUs), which are sets of contiguous counties. Since the CE survey is used to weight the CPI, the two surveys use a common set of PSUs. There are two types of PSUs: self representing and non-self representing PSUs. Self representing PSUs are selected with certainty, whereas non-self representing PSUs are grouped into a stratification PSU and one PSU is randomly selected to represent the stratification PSU. To minimize survey variance, the stratification PSUs should be homogeneous and have approximately equal populations. This is a constrained clustering problem and is solved using heuristic algorithms. This paper presents a new heuristic solution procedure that uses a "pseudo" assignment algorithm to assign PSUs to a stratification PSU. This heuristic procedure found a lower within cluster variability, Trace (W), than other procedures.

Key Words: clustering, integer programming, assignment algorithm, stratification PSU, Consumer Expenditure Survey, Consumer Price Index

1. Introduction

Every ten years the Consumer Price Index (CPI) and Consumer Expenditure Survey (CE) redefine their primary sampling units (PSUs) using the latest population estimates from the decennial census. Both surveys share the same sample design because the CPI uses CE's expenditure estimates for its survey weights. PSUs are small clusters of counties that are grouped together into entities called "core-based statistical areas" (CBSAs), which are defined by the U.S. Office of Management and Budget based on their degree of economic and social integration as measured by commuting patterns. Large CBSAs (populations over 2.35 million) are self-representing PSUs, and small CBSAs (populations under 2.35 million) are non-self representing PSUs. The non-self representing PSUs are grouped into stratification clusters, which in the literature are referred to as stratification PSUs. One PSU is randomly selected from each cluster to be its representative. Each PSU's probability of selection is based on size, the ratio of an individual PSU's population to its total cluster population. The number of stratification PSUs is determined in advance. The objective is to form homogeneous stratification PSUs and each stratification PSU should have approximately the same population to minimize the within stratum portion of the total survey variance. This is a constrained clustering problem and there is not an exact solution. Traditional clustering algorithms find homogenous stratification PSUs, but do not balance the population. This paper presents a heuristic algorithm for solving the PSU assignment problem using k-means clustering and zero-one integer linear programming. Due to the optimization, this approach finds more homogeneous clusters than other heuristic procedures such as the Friedman-Rubin hill climbing algorithm. An example is

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

given for clustering smaller metropolitan CBSAs (populations between 50,000 and 2.35 million) for the Northeast region.

2. Data

The data came from several sources. The CBSA's are from OMB. The variables used in the model are: median property value, median income, latitude, longitude, and the population of a CBSA. The clustering variables: median property value and median income are calculated from the American Community Survey (ACS) using data from 2005-2007. The median property value and income are averaged for all of the counties in a CBSA. The stratification PSUs are balanced by population, which is also from the ACS. The ACS county population is summed for each county in the CBSA. The latitude and longitude centroid for each CBSA is found from the U.S. Census Bureau's TIGER/Line® Shapefiles. For each region and size, average median property value, average median income, latitude, and longitude are standardized. All of the procedures were coded in SAS®. However, other software packages could be used.

3. Trace (W)

In clustering, Trace (W) is a measure of cluster homogeneity (Everitt *et al.)*. PSUs assigned to a stratification PSU should be homogeneous and there should be variability between the stratification clusters. The total variability (Total) is decomposed into within cluster variability (W) and between cluster variability (B), Total = W + B. Clustering either minimizes the within cluster variability (W) or maximizes the between cluster variability (B). In this project, the objective is to minimize the within cluster variability using the four model variables: median property value, median income, longitude and latitude. For a multivariate problem, W is defined as:

$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (v_{ij} - \overline{v}_i)(v_{ij} - \overline{v}_i)^T$$

where:

g is the number of clusters;

v_{ij} is the vector of standardized median property value, median income, longitude, and latitude of the j^{th} PSU in cluster i;

$\overline{v}_i$ is the mean of the standardized median property value, median income, longitude, and latitude for the PSUs in cluster i;

n_i is the number of PSUs in cluster i.

By taking the Trace of the W matrix, the sum of the diagonal elements, the within cluster variance is summarized by a single number. Minimizing Trace (W) is equivalent to minimizing the sum of squared Euclidean distances between the set of observations and the group mean. Smaller Trace (W)'s indicate homogeneity of the clusters.

4. Methods

The first step in the non self-representing stratification PSU assignment algorithm is to solve a relaxed clustering problem by ignoring the population constraint. Since the number of strata, clusters, is fixed for each region-size (e.g. Northeast and metropolitan CBSAs), the clustering algorithm must find a pre-specified number of clusters. K-means clustering is a partitive clustering technique. The standardized four model variables, median property value, median income, latitude, and longitude are used in the clustering. The k-means clustering algorithm may converge with a cluster having only one PSU or at the other extreme a cluster may contain the majority of the PSUs. Thus, in the relaxed clustering problem, the clusters are not balanced either in the number of PSUs or by population. However, the cluster centers are used in the next step.

The second step in the non self-representing stratification PSU assignment algorithm is to solve a "pseudo" assignment problem. The assignment problem is a classic problem in operations research. The assignment of PSUs to stratification PSUs is illustrated for the Northeast in Figure 1 of the Appendix. There are 41 PSUs and 6 stratification PSUs (k=6 for the Northeast). The PSUs are located at nodes and arcs connect the nodes. The arcs flow in one direction, from left to right. The decision variable, x_{ij}, equals 1 if PSU i is assigned to stratification PSU j. Otherwise, $x_{ij} = 0$.

The homogeneity requirement is accounted for in the linear objective function. Associated with each PSU is a standardized value for the four model variables: median property value, median income, latitude, and longitude. From the first step, the cluster centers have four standardized variables, median property value, median income, latitude, and longitude, and a cluster center is assigned to each one of the stratification PSU nodes. The Euclidean distance is calculated between each PSU and each stratification PSU. Let c_{ij} be the distance or cost of assigning PSU i to the stratification PSU j. The objective function is:

$$\textit{Minimize} \qquad \sum_{i=1}^{41}\sum_{j=1}^{6} c_{ij}x_{ij}$$

Every PSU be assigned to one and only one stratification PSU. This constraint is formulated as:

$$\sum_{j=1}^{6} x_{ij} = 1 \qquad \text{for every i}$$

The population of each stratification PSU, must be between an upper and lower bound on the population. These bounds are the total PSU population of a metropolitan, micropolitan, or rural PSUs in a region divided by the number of stratification PSUs plus or minus a tolerance, usually 10 percent. For the metropolitan PSUs in the Northeast, the bounds are 2,753,303 and 3,365,148. Let p_i equal the population of PSU i. The bounds on the population are:

$$\sum_{i=1}^{41} p_i x_{ij} \geq 2,753,305 \qquad \text{for every j}$$

$$\sum_{i=1}^{41} p_i x_{ij} \leq 3,365,148 \qquad \text{for every j}$$

Combining the objective function and constraints, the zero-one integer programming problem is expressed as:

$$\textit{Minimize} \quad \sum_{i=1}^{41}\sum_{j=1}^{6} c_{ij} x_{ij}$$

$$\textit{Subject to}: \quad \sum_{j=1}^{6} x_{ij} = 1 \qquad \text{for every i}$$

$$\sum_{i=1}^{41} p_i x_{ij} \geq 2,753,305 \qquad \text{for every j}$$

$$\sum_{i=1}^{41} p_i x_{ij} \leq 3,365,148 \qquad \text{for every j}$$

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

The third step is the stratification cluster update. New cluster centers are calculated by finding the cluster mean for median property value, median income, latitude, and longitude. The second step is repeated with the new cluster centers. The third step is repeated until the stopping criterion is met. The algorithm stops when either the Trace (W) is the same for two consecutive iterations or when the objective function has the same value for two consecutive iterations. The user selects one of these two stopping criterion. The only difference between the two stopping criterion is that the latter requires an extra iteration. They produce the same stratification PSU assignment.

The number of arcs in the network is the product of the number of PSUs and the number of stratification PSUs. As the number of arcs increase, both the time required solving and the complexity of the problem increases. Computer memory size may restrict larger problems from being solved. K-means clustering finds good clusters, but not necessarily the best and the quality of the solutions may be affected by the initial clusters and the size of the problem.

5. Results

The results for the Northeast are shown in Figure 2 and Table 1a - Table 1f in the Appendix. The stratification PSU colors are linked to the legend block in the tables. By

definition, a map shows the spatial distribution of the clusters for two of the four variables: latitude and longitude. The other two variables, median property value and median income, have the same weight as latitude and longitude and also influence the cluster assignment. From the tables, the total population of each cluster is between the lower and upper population bounds. Due to the Step 3 cluster updating, the stratification cluster center has the cluster average for median property value, median income, latitude, and longitude. For the Northeast, the objective function and Trace (W) were: 37.80 and 41.98, respectively.

Several alternative approaches were investigated to assign PSUs to a stratification PSU. Of the alternative procedures, the Friedman-Rubin hill climbing algorithm (Everitt *et al*., Khandaker and Reist, and Johnson *et al*.) had the lowest Trace (W)'s. This is the current procedure that CPI uses for assigning PSUs to a stratification PSU. The hill climbing algorithm is a greedy heuristic. The first step finds an initial partition of the PSUs into a pre-specified number of stratification groups. The PSUs are randomly assigned to stratification groups while maintaining the strata population bounds. In the second step, hill climbing, individual PSUs are moved one at a time from one stratum to another to reduce the Trace (W) or other criterion. The third step, the exchange procedure also attempts to minimize the Trace (W) by selecting pairs of PSUs from different strata and interchanging them. The final step is the size adjustment procedure which attempts to move and exchange PSUs to maintain the population bounds of the stratification cluster while simultaneously minimizing the objective criterion. Different initial partitions may lead to different final solutions. At every stage the algorithm makes one greedy choice after another without consideration or revision of decisions made in previous stages. Only at the final step is any attempt made to balance the stratification cluster populations.

The "pseudo" assignment algorithm finds a lower Trace (W) than the Friedman-Rubin hill climbing algorithm for all regions (Table 1). The Northeast has a smaller land area than the other regions, but is heavily populated with many adjacent PSUs. This contributes to the lower Trace (W)'s. The South includes a land area from Delaware to Texas, which contributes to a high Trace (W). The Midwest has several cities that are almost self representing and to meet the population bounds, smaller PSUs farther away are assigned to the stratification cluster. This contributes to the higher Trace (W).

Table 1: Metropolitan Trace (W) s for the "Pseudo "Assignment and the Friedman-Rubin Algorithm

Area	"Pseudo" Assignment Algorithm	Friedman-Rubin Algorithm
Northeast	41.98	52.21
Midwest	109.02	149.61
South	98.27	167.69
West	45.43	79.16

6. Conclusion and Implementation

The "pseudo" assignment algorithm provides superior stratification PSU clustering assignments than the Friedman-Rubin algorithm as measured by the Trace (W). The

algorithm uses commercially available software, is explainable, and fast. The only limitations are those from the software for the k-means clustering and the zero-one integer programming. It is a promising approach for solving the assignment of PSUs to stratification PSUs due to population changes reflected in the decennial census.

Historically, CPI and CE have stratified PSUs using geographic regions as defined by the U.S. Census Bureau and size, metropolitan, micropolitan, and rural. The U.S. Census Bureau (http://www.census.gov/geo/www/us_regdiv.pdf) also subdivides the regions into divisions. After this research was completed, a management decision was made to change the stratification to Census divisions and to collapse the metropolitan and micropolitan CBSAs into the same strata. The Trace (W)s are lower for stratification by divisions because of the smaller geographic area. The purpose of this paper was to introduce a heuristic algorithm for assigning PSUs to stratification PSU. The procedure in this paper was applied without difficulty to the new stratification.

7. Acknowledgements

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

8. Literature Cited

Everitt, B. S., Landau, S., Leese, M. (2001). Cluster Analysis, Fourth Edition. Oxford University Press, Inc., New York.

Johnson, W. H., Shoemaker, O. W., and Rhee, Y. W. (2002). Redesigning the Consumer Price Index Area Sample. Proceedings of the Section on Government Statistics, American Statistical Association, 1671-1676.

Khandaker, A.M., Reist, B.M. (2010). Evaluating alternative criteria for primary sampling units stratification. Proceedings of the Section on Survey Research Methods, American Statistical Association, 4664-4672.

Appendix: Figures and Tables

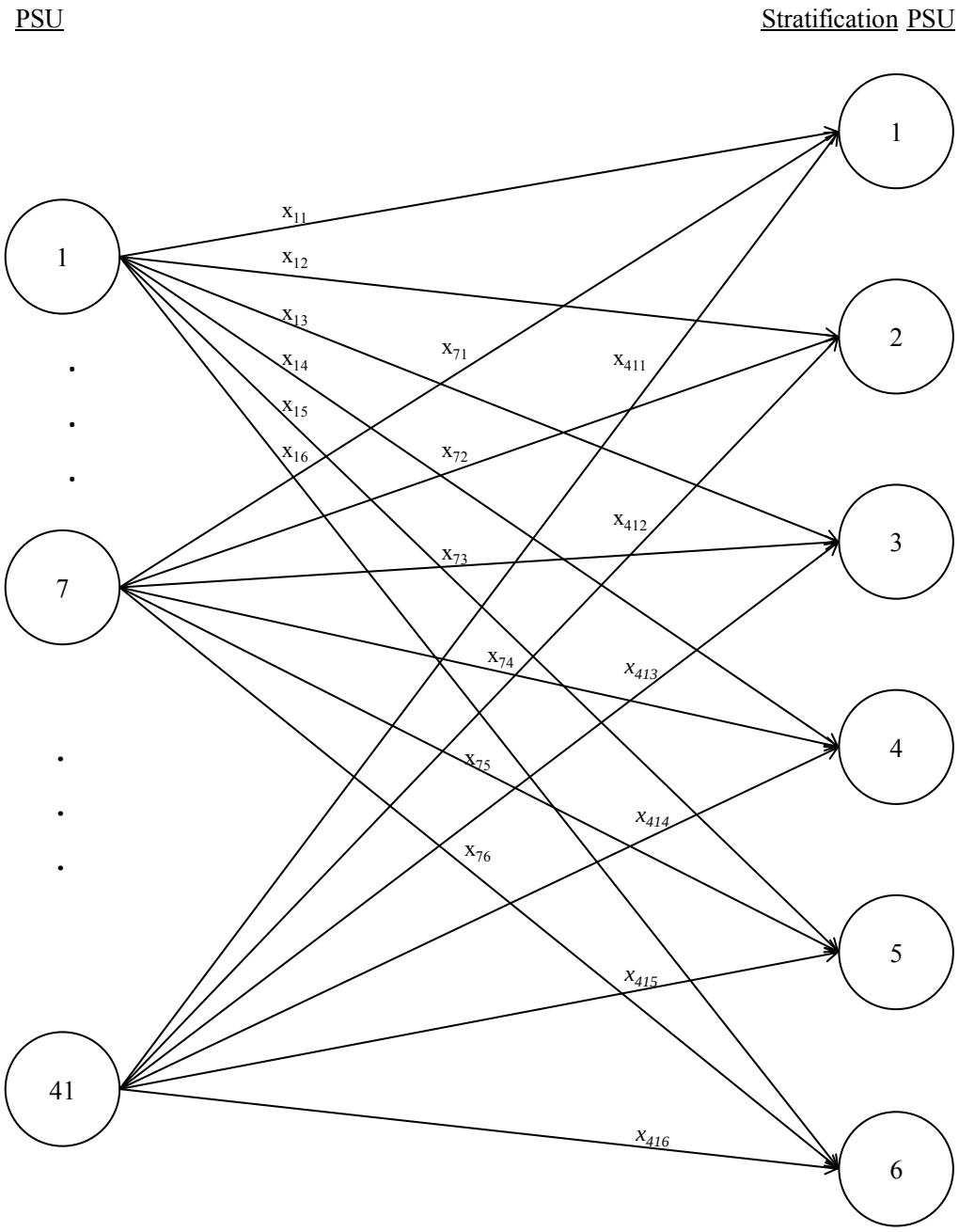

Figure 1: The Northeast has 41 PSUs and six stratification PSUs. There are 246 zero-one decision variables. Each PSU can be assigned to only one of the six stratification PSUs.

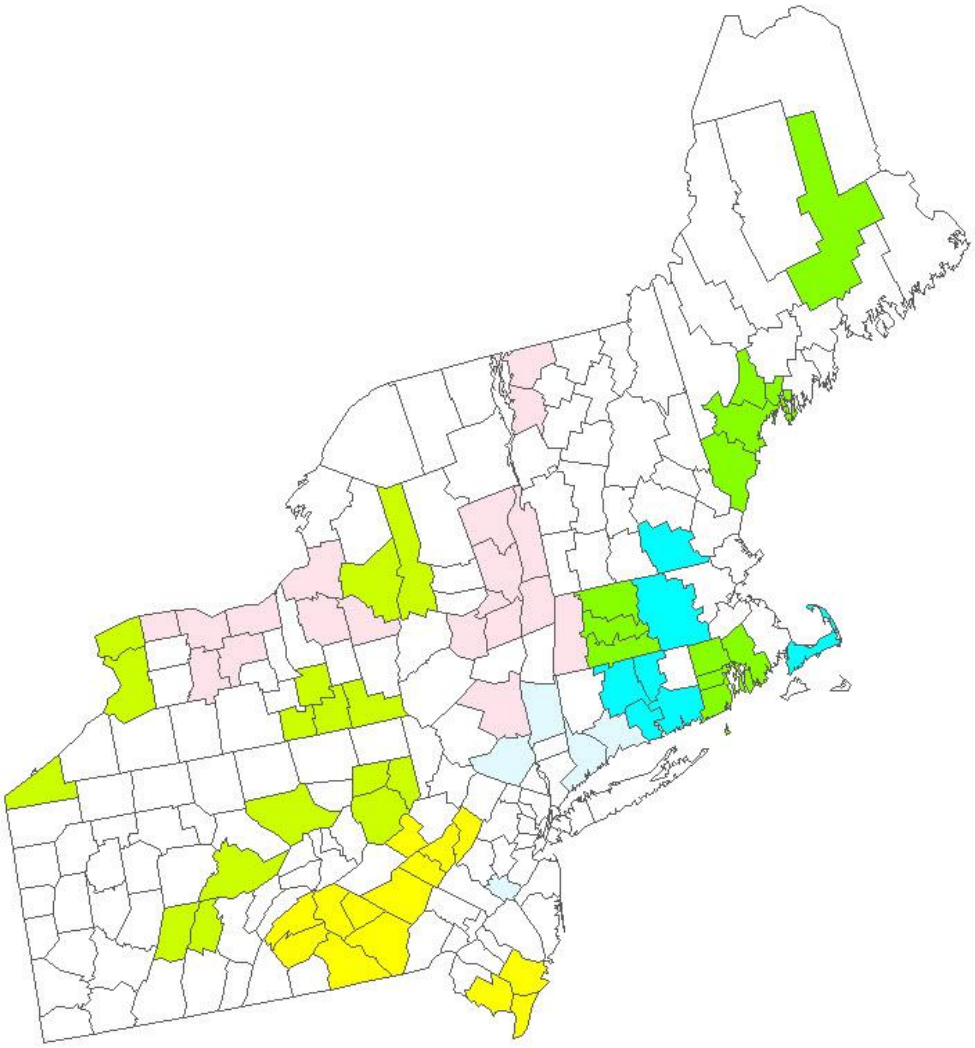

Figure 2: This map shows the geographic distribution of the six stratification PSUs in the Northeast. The PSU names, median property value, median income, latitude, longitude, and stratification cluster colors are given in Table 1a-f.

Table 1a: Cluster assignments for the first stratification PSU

Stratification PSU	PSU	Population	Median Property Value	Median Income	Latitude	Longitude
1	Final Stratification Cluster Center	0	294,840	62,648	42.03	-71.67
	Barnstable Town, MA	223,574	412,900	58,422	41.70	-70.30
	Norwich-New London, CT	267,029	252,400	61,842	41.45	-72.09
	Manchester-Nashua, NH	400,855	276,300	67,276	42.90	-71.58
	Worcester, MA	779,386	289,600	60,709	42.33	-71.84
	Hartford-West Hartford-East Hartford, CT	1,185,150	243,000	64,989	41.77	-72.54
Total		2,855,994				

Table 1b: Cluster assignments for the second stratification PSU

Stratification PSU	PSU	Population	Median Property Value	Median Income	Latitude	Longitude
2	Final Stratification Cluster Center	0	200,560	47,932	43.43	-70.64
	Lewiston-Auburn, ME	106,837	144,800	42,725	44.14	-70.22
	Bangor, ME	147,971	118,200	41,336	45.17	-68.72
	Portland-South Portland-Biddeford, ME	512,189	234,500	53,270	43.78	-70.33
	Springfield, MA	683,262	203,700	48,265	42.32	-72.57
	Providence-New Bedford-Fall River, RI-MA	1,605,211	301,600	54,064	41.71	-71.37
Total		3,055,470				

Table 1c: Cluster assignments for the third stratification PSU

Stratification PSU	PSU	Population	Median Property Value	Median Income	Latitude	Longitude
3	Final Stratification Cluster Center	0	105,791	41,767	41.77	-77.40
	Elmira, NY	88,199	78,300	39,989	42.13	-76.79
	Ithaca, NY	100,590	147,900	46,225	42.46	-76.48
	Williamsport, PA	117,311	108,700	40,430	41.27	-77.00
	Altoona, PA	125,711	91,100	40,196	40.48	-78.38
	State College, PA	143,557	157,600	42,976	40.89	-77.83
	Johnstown, PA	145,984	80,500	37,030	40.44	-78.79
	Binghamton, NY	246,800	91,300	42,930	42.14	-76.11
	Erie, PA	279,252	102,600	42,073	42.05	-80.06
	Utica-Rome, NY	295,059	89,100	42,105	43.18	-75.18
	Scranton--Wilkes-Barre, PA	548,942	112,100	40,737	41.43	-75.95
	Buffalo-Niagara Falls, NY	1,134,280	104,500	44,747	43.00	-78.82
Total		3,225,685				

Table 1d: Cluster assignments for the fourth stratification PSU

Stratification PSU	PSU	Population	Median Property Value	Median Income	Latitude	Longitude
4	Final Stratification Cluster Center	0	188,544	52,360	39.97	-75.83
	Ocean City, NJ	97,555	328,600	52,771	39.09	-74.80
	Lebanon, PA	126,426	140,800	49,805	40.35	-76.45
	Vineland-Millville-Bridgeton, NJ	154,086	156,500	48,464	39.40	-75.09
	Atlantic City-Hammonton, NJ	269,774	247,000	53,473	39.45	-74.62
	Reading, PA	398,155	153,100	52,241	40.38	-75.91
	York-Hanover, PA	414,023	156,300	53,641	39.93	-76.74
	Lancaster, PA	493,910	169,500	52,933	40.07	-76.28
	Harrisburg-Carlisle, PA	524,665	146,500	53,496	40.32	-77.13
	Allentown-Bethlehem-Easton, PA-NJ	794,961	198,600	54,420	40.75	-75.41
Total		3,273,555				

Table 1e: Cluster assignments for the fifth stratification PSU

Stratification PSU	PSU	Population	Median Property Value	Median Income	Latitude	Longitude
5	Final Stratification Cluster Center	0	346,025	67,960	41.10	-73.76
	Trenton-Ewing, NJ	364,567	304,600	68,582	40.25	-74.71
	Poughkeepsie-Newburgh-Middletown, NY	666,388	319,500	66,376	41.56	-74.04
	New Haven-Milford, CT	843,571	264,800	58,528	41.39	-72.94
	Bridgeport-Stamford-Norwalk, CT	894,724	495,200	78,353	41.21	-73.35
Total		2,769,250				

Table 1f: Cluster assignments for the sixth stratification PSU

Stratification PSU	PSU	Population	Median Property Value	Median Income	Latitude	Longitude
6	Final Stratification Cluster Center	0	167,929	51,105	43.05	-74.53
	Glens Falls, NY	128,279	137,600	46,168	43.37	-73.60
	Pittsfield, MA	130,346	184,900	48,836	42.40	-73.21
	Kingston, NY	181,755	237,400	54,871	41.86	-74.15
	Burlington-South Burlington, VT	206,437	225,200	56,284	44.82	-73.13
	Syracuse, NY	646,597	104,400	47,315	43.05	-76.18
	Albany-Schenectady-Troy, NY	850,506	168,500	54,755	42.70	-73.85
	Rochester, NY	1,031,480	117,500	49,508	43.15	-77.63
Total		3,175,400				